Fleuß, Dannica; Helbig, Karoline

Measuring Nation States' Deliberativeness: Systematic Challenges, Methodological Pitfalls, and Strategies for Upscaling the Measurement of Deliberation

Political Studies

Political Studies Association | POLITICAL STUDIES

Article

Political Studies
2021, Vol. 69(2) 307–325

DOI: 10.1177/0032321719890817
journals.sagepub.com/home/psx
SAGE

Measuring Nation States' Deliberativeness: Systematic Challenges, Methodological Pitfalls, and Strategies for Upscaling the Measurement of Deliberation

Dannica Fleuß[1] 
and Karoline Helbig[2]

Abstract

A theoretically reflected and empirically valid measurement of nation states' democratic quality must include an assessment of polities' deliberativeness. This article examines the assessment of deliberativeness suggested by two sophisticated contemporary measurements of democratic quality, that is, the *Democracy Barometer* and the *Varieties of Democracy*-project. We feature two sets of challenges, each measurement of deliberativeness must meet: First, it must address the methodological challenges arising in the course of conceptualizing, operationalizing, and aggregating complex concepts (see Munck and Verkuilen, 2002). Second, attempts to measure nation states' deliberativeness are confronted with specific conceptual and systematic challenges which we derive from recent deliberative democracy scholarship. We argue that both *Democracy Barometer* and *Varieties of Democracy*-project provide highly sophisticated assessments of democratic quality, but ultimately fail to capture nation states "deliberativeness" in a theoretically reflected and methodologically sound manner. We examine the methodological, pragmatic, and systematic reasons for these shortcomings. The crucial task for measurements of nation states' deliberativeness consists in providing a conceptual approach and methodological framework for "upscaling" existing meso-level measurements (such as the DQI). The concluding section presents conceptual and methodological strategies that can enable researchers to meet these challenges and to provide a theoretically grounded and empirically valid measurement of nation states' deliberativeness.

Keywords

deliberative democracy, deliberation, measurement of democracy, measurement of democratic quality, deliberative systems, upscaling deliberation

Accepted: 5 November 2019

[1]Department of Political Science, Helmut-Schmidt-University, Hamburg, Germany
[2]WZB Berlin Social Science Center, Berlin, Germany

Corresponding author:
Dannica Fleuß, Department of Political Science, Helmut-Schmidt-University, Holstenhofweg 85, Hamburg 22043, Germany.
Email: dannica.fleuss@hsu-hh.de

Introduction

Deliberative democratic theories formulate a major contemporary paradigm of democratic legitimacy (Dryzek, 2015; Elstub, 2015). According to deliberative democrats, democratic self-legislation manifests not only in institutionalized electoral procedures, but is realized first and foremost in society-wide, free and unrestrained "taking and giving reasons" about politically relevant issues. From this perspective, civil society communications and the corresponding *bottom up*-input that is fed into empowered, decision-making institutions are crucial for political systems' democratic quality (Dryzek and Niemeyer, 2010: 11). This *normative* claim may be all but undisputed among theoretical and empirical scholars. Nevertheless, all major theoretical paradigms presupposed in established indices of democratic quality agree on the *functional value* of deliberative practices for inclusive and egalitarian democratic decision-making.[1] The merits of deliberation for democratic quality are increasingly acknowledged by practitioners worldwide who consider deliberation as a means for tackling the most pressing issues of today's representative democracies: The Belgian G1000 project employed deliberative mini-publics to overcome "the limits of representative democracy" (G1000, 2012; see Caluwaerts and Reuchamps, 2015). In Ireland, "We the Citizens" was established with the purpose to "enhance [. . .] democracy" in the context of citizens' decreasing trust in established political institutions (We the Citizens, 2011). "America in One Room" is supposed to tackle inner-societal polarization and to provide a democratic alternative to populism and technocracy with a view to the 2020 presidential election (Fishkin, 2018: 3, 70; see America in One Room, 2019).

For a long time, measurements of democracy did not reflect this widespread acknowledgment of the significance of deliberative procedures in their empirical assessments: Neither the *Index of Freedom in the World* nor *Polity IV* or Vanhanen's *Polyarchy Index* integrate deliberation at a conceptual level (not to mention: measure it in a systematic manner) (Freedom House, 2019; Marshall and Jaggers, 2018; Vanhanen, 2000). In recent years, two measurement instruments have been developed that explicitly aim at providing a more differentiated measurement of democratic quality[2] that is grounded in democratic theory:[3] The *Democracy Barometer* (DB) represents its own procedure as "basically theory-driven" (Bühlmann et al., 2012: 116; see Jäckle et al., 2012) and the V-Dem integrates extensive reflections of different democratic theory-paradigms in conceptualizing "democratic quality" (c.f. Coppedge et al., 2016c: 9–21, 2018b: 4–14).

This article's point of departure is the premise that measures of democracy should include an assessment of nation states' overall deliberative quality (for short: "deliberativeness") in order to achieve valid and empirically meaningful measurements (see Fleuß et al., 2018). From a democratic theory perspective, measurements must not only consider the processes of public opinion and will-formation that are ultimately conveyed into decision-making, but also provide a measurement of their *quality*. This task does, however, prove demanding from a conceptual as well as from a methodological point of view: First, there are several competing theories of deliberative democracy (Elstub, 2010). Accordingly, there are multiple definitions of deliberation and deliberative quality that would suggest divergent normative criteria for evaluating democracies' deliberativeness which, in turn, would lead to different operationalizations and strategies of measurement. Second, real-world deliberative procedures occur in various spaces or *sites* of a political system (Dryzek and Niemeyer, 2010; Mansbridge et al., 2012). Especially informal deliberations in the civil society and so-called "everyday political talk" (Mansbridge, 1999)

do, however, pose enormous challenges for comparative large-scale measurements, for example, in terms of data availability. A valid and encompassing measure of democracies' deliberative quality currently constitutes a gap in research, but "embarking on an attempt to do so will yield considerable benefit to both democratic assessment and deliberative democracy" (Niemeyer et al., 2015: 2).

Against this background, the article examines the assessments of democracies' deliberativeness provided in the DB's measurement approach and the V-Dem to answer two overarching research questions:[4] (1) Do the measurements of democratic quality offered by DB and V-Dem integrate an assessment of nation states' deliberativeness in a conceptually and methodologically sound manner? (2) To what extent are both indices able to cope with the systematic challenges that any assessment of democracies' deliberativeness must deal with—and what can be learned from their respective accomplishments and shortcomings for developing a comprehensive measurement of democratic nation states' deliberative quality?

In order to answer question (1), section "Reviewing DB's and V-Dem's Measurement Methodology" of this article will aim at a careful representation of the measurements of deliberative quality provided in DB and V-Dem. These assessments will be structured in accordance with Munck's and Verkuilen's (2002; also see Munck, 2009; Pickel et al., 2015) well-known criteria for evaluating each democracy indices' conceptual and methodological validity. Section "Three Conceptual Challenges of 'Upscaling' Measurements of Deliberative Quality and How to Address Them" takes a step back and provides an assessment of DB's and V-Dem's measurements' content validity, that is, it addresses both indices' potential to cope with systematic challenges of measurements of democracies' deliberativeness: Section "Systematic Challenges of Measuring Democratic Deliberation" outlines these challenges based on a discussion of recent theoretical and empirical research from the field of deliberative democracy. Subsequently to applying these considerations to DB's and V-Dem's measurement approaches (in section "Do DB and V-Dem Meet the Systematic Challenges of Measuring Deliberation?"), we argue that crucial flaws in current assessments of nation states' deliberativeness result from inadequate conceptual approaches for "upscaling" existing measurements of deliberative quality (such as the DQI) from the meso- to the macro-level. In the concluding section, we outline strategies to address this challenge (section "Upscaling measurements of deliberative quality: Lessons to learn for measuring deliberativeness").

Reviewing DB's and V-Dem's Measurement Methodology

Methodological Standards of Measurement Instruments

The measurement of democratic quality has been an important area of empirical research during the last 50 years and gained momentum after the cold war (Pickel and Pickel, 2007). Munck and Verkuilen (2002) criticized the then existing indices for their having "paid sparse attention to the quality of the data on democracy" and mostly ignoring "problems of conceptualization and measurement" (Munck and Verkuilen, 2002: 5–6). Their framework was not only meant to enable the comparative evaluation of indices, but also to provide the basis for a more systematic development of new measurement approaches.[5] Leaning on this framework (see Munck and Verkuilen, 2002: 8; see Munck, 2009; Pickel et al., 2015), our assessment of DB's and V-Dem's approaches to deliberation will differentiate between conceptualization, measurement and aggregation—and the corresponding standards of assessment.

At the level of *conceptualization*, a concept of deliberative quality must be specified. In order to do so, specific attributes have to be identified, while *avoiding both a maximalist or minimalist definition*, that is, the concept should not include theoretically irrelevant attributes or exclude theoretically relevant ones (Munck and Verkuilen, 2002: 9). The first methodological challenge (*MC*1) highlighted by Munck and Verkuilen is to comply with those claims. The challenge *MC*2 refers to the adequateness of the "concept trees": The previously identified attributes have to be organized in a concept tree. The attributes form the "leaves" of that tree, and are subsumed under categories of hierarchically organized levels of abstraction (see Munck and Verkuilen, 2002: 12). In the course of organizing the attributes, they have to be clearly isolated, avoiding redundancy (including a criterion repeatedly) and conflation (one attribute containing components that should actually be assigned to at least two different attributes) (Munck and Verkuilen, 2002: 13–14).

On the level of *measurement*, attention has to be paid to the selection of indicators, the selection of measurement levels and the recording of the measurement process. With regard to the selection of indicators, validity can be ensured by using multiple indicators and establishing the cross-system equivalence of these indicators (see Pickel et al., 2015: 506–507). Those multiple indicators can be used to minimize measurement error and should be cross-checked through multiple sources (see Munck and Verkuilen, 2002: 15–16). The validity of the *selection of indicators* comprises challenge *MC*3. The validity of the *measurement levels* (*MC*4) has to be secured by avoiding:

> The excesses of introducing distinctions that are either too fine-grained, which would result in statements about measurement that are simply not plausible in light of the available information and the extent to which measurement error can be minimized, or too coarse-grained, which would result in cases that we are quite certain are different being placed together (Munck and Verkuilen, 2002: 17).

The selection thus requires theoretical justification as well empirical testing (c.f. Munck and Verkuilen, 2002: 17).

The selection of both indicators and measurement levels has to allow for a reliable measurement of the concept. On the level of *aggregation*, choices have to be made regarding the level of aggregation and an adequate aggregation rule. The *level of aggregation* has to "balance the goal of parsimony with the concern with underlying dimensionality and differentiation" (*MC*5), while the *development of an aggregation rule* should "ensure the correspondence between the theory of the relationship between attributes and the selected rule of aggregation" (*MC*6; Munck and Verkuilen, 2002: 8; also see Møller and Skaaning, 2012; Pickel et al., 2015: 508–509). Throughout the whole process, the conceptualization has to be made transparent, the measuring process has to be recorded by means of coding rules and coding processes, the disaggregate and aggregate data must be made accessible, and the aggregation rule has to be published to ensure transparency and replicability (also see Pickel et al., 2015: 500). In the following representations of DB's and V-Dem's assessments of deliberative quality, we will apply criteria *MC*1–*MC*6 to evaluate their measurement of deliberativeness in democratic polities.

The Democracy Barometer

The first thing that catches one's eye with regard to DB's assessment is that the term "deliberation" appears neither in the measurements of the instrument (Merkel et al., 2016a) nor in its methodological descriptions (Merkel et al., 2016b). Since Bühlmann

et al. (2012) do mention deliberation in their introduction of DB as part of a democratic model, the instrument is based on *and* claim to propose a "basically theory-driven" measurement of democratic quality, this is quite an astonishing fact: In delimitation against a minimalistic model on the one hand and an output-oriented model of democracy on the other hand, they define the participatory model which considers participation as the core of democracy and stipulates that collective decisions should be derived from a *deliberative process* (Bühlmann et al., 2012: 117). However, the selection of the specific elements of this conceptual model that are transferred into a measurement approach (and, in consequence, the omission of others) is not explained. In the following, we will try to determine the reasons for this omission of the term deliberation and illustrate in what other ways the inherent principles of deliberation appear in DB. To provide the basis for this, we consult the concept of "embedded democracy" as it constitutes DB's theoretical foundation (Bühlmann et al., 2008: 117). Since the term "deliberation" itself does not appear in the instrument, we will not be able to discuss the details of the authors' understanding of deliberation and applying Munck's and Verkuilen's criteria as summarized in *MC*1–*MC*6 will only be partly possible. Therefore, this section will focus on reconstructing the conceptualization process and feature the reasons for the authors' decision to omit the concept of "deliberation."

DB's conceptual structure is derived from the "embedded democracy"-concept and is characterized by three principles (Munck and Verkuilen, 2002: "attributes" as they will be called henceforth), that is "freedom," "equality" and "control" (Bühlmann et al., 2008: 117, 118). The basic idea of embedded democracy is the definition of democracy as "a union of interdependent and independent partial regimes" (Merkel and Croissant, 2003: 59; translation t.a.). If one of the regimes is impaired, this affects the functioning of all other regimes (Merkel and Croissant, 2003: 59). Those partial regimes are elections, the right to political participation, civil rights, horizontal accountability and the effective power of government, the second of which is concerned with political communication in the public sphere. Merkel and Croissant view the public sphere as "an independent sphere of political action, in which organizational as well as communicative power can be built under the surveillance of the public, and which can support the competition for political power by collective opinion and will-formation." (Merkel and Croissant, 2003: 62; translation t.a.). Although DB's authors would hardly describe themselves as "deliberative democrats," this quote represents classic ideas of deliberative democracy. Furthermore, freedom of opinion and speech, the right to demonstrations and petitions (Merkel and Croissant, 2003: 62, 71), as well as grassroots movements, NGOs, lobby groups, direct democratic forms of participation like referenda, civil society discussion forums, participation in the planning of large infrastructure projects and a pluralistic media system are perceived as indications of an enabled public sphere in an embedded democracy. The indicators used by DB to *measure* democratic quality do, however, only partly cover these democratic institutions and practices: While all the above-mentioned constitutional rights, the media and even the existence of NGOs and lobby groups have been considered, bottom-up participatory practices like grassroot movements, deliberative discussion forums or the participation in democratic innovations such as the planning of infrastructure projects are not considered in the measurement. Though these ideas derived from deliberative democratic theory constitute only a small part of the overall theoretical framework "embedded democracy," their omission in the Barometer obviously cannot be explained by the authors' oblivion to the importance of political deliberation for the quality of democracy. Even if deliberation ranks very lowly in their conception of democracy,

this can only be part of the reason for its exclusion. Based on a closer assessment of DB's dealing with democratic deliberation, we identified three systematic reasons for this mismatch between the general (theoretical) acknowledgment of deliberation's significance for democratic quality and its omission in the measurement instrument.

The concept tree is not an exact representation of the concept "embedded democracy." The indicators associated with deliberation appear within different branches of the concept tree (*MC*2). Although they are prominently enumerated under the heading "participation" in the embedded democracy-model (Merkel and Croissant, 2003: 62, 71), they are scattered among all three attributes and various components[6] of DB.[7] Thus, the concept tree of DB does not—and is not intended to—mirror the model of embedded democracy, not only with regards to deliberation, but in large parts of the structure.

Rules in Law are Privileged Over Rules in Use. It is easy to see that most indicators associated with deliberation in the embedded democracy-model are concerned with its constitutional and institutional framework, that is, the *rules in law* (Bühlmann et al., 2012: 131). The indicators concerning the freedom of opinion, speech, and the press (Constspeech, Constrpress), the freedom to associate (Constfras, Constass), as well as the freedom of information (RestricFOI, EffFOI) and the legal environment of press freedom (Legmedia) can all be seen as parts of the *constitutional and legal framework* of democratic deliberation. This institutional framework for deliberation is only partly mapped in DB. DB does, however, offer a very thorough examination of the democratic quality for one significant sphere of public political communication, that is, the media that are addressed with regards to their pluralism (Newsimp, Newspaper), neutrality (Balpress, Neutrnp), and independence (Polmedia). In addition, a selection of indicators measures the density of organizations representing economic and public interests that are potential institutionalized contexts of political deliberation, such as trade unions and other professional organizations (Union, Memproorg) or humanitarian and environmental organizations (Memhuman, Memenenviron). Although these institutions represent *rules in use*, they constitute the context of deliberation and thus, from this perspective, reify as *rules in law*, since the mere existence of such organizations provides a context for, but does *not guarantee* deliberation.

The Measurement Prefers to Rely on "Hard Data." The overrepresentation of *rules in law* over *rules in use* indicates that the reason for omitting deliberation from DB might not merely result from decisions associated with *conceptualizing* "democratic quality," but even more from the *measurement* strategy applied in the instrument. Measuring deliberation only by using "hard data," as aspired by DB (Merkel et al., 2016a: 8), is hardly a feasible strategy. Therefore, the more than 300 indicators of secondary data out of which the indicators of the Barometer have been selected (Merkel et al., 2016a: 7) might not have contained many items suitable for measuring deliberation in the first place. The indicators allegedly were derived by a stepwise deduction of principles, components, subcomponents, and indicators (Merkel et al., 2016b). However, the pool of the 300 potential indicators seems to have existed beforehand,[8] so there must be an inconsistency between the deduction of the subcomponents and the selection of indicators. Therefore, even if the subcomponent "political deliberation" would have been inserted into the function "participation," there probably would have been too few indicators to measure it—or not even any at all. For a smooth appearance of the step-by-step-deduction from concept

to indicators it appears, thus, rather convenient that deliberation has not been conceptualized in the first place.

Since DB does not explicitly measure the *concept of deliberation* and does not build an index of deliberation, addressing the methodological criteria *MC*4–*MC*6 is obsolete. The scattered indicators associated with deliberation could, however, in principle be aggregated to an index of deliberative quality: It might be possible to simply "collect" the associated data, aggregate them into a deliberation score, and integrate this score into the aggregation of the general country-year score. However, such a procedure would compromise the stability of the whole instrument, either by inherently doubling the values of the respective indicators or by moving them from one component to another, thus leaving instable[9] components. In the latter case, indicators relevant for measuring deliberative quality would have to be distributed more regularly across all attributes of DB and would have to be weighted according to their importance in the respective context. Furthermore, the only partial transfer of deliberation-associated ideas into the instrument, the overrepresentation of *rules in law* over *rules in use*, and of quasi-metric scales over "hard numbers" would call for a very elaborate (and highly artificial) weighing and aggregation in order to produce reliability while still reflecting the theoretical interrelations of the indicators.

The Varieties of Democracy-Project

In its conceptualization of democracy, V-Dem explicitly acknowledges that democracy is an essentially contested concept which is understood, interpreted, and specified in many different (and often incompatible) ways (Coppedge et al., 2016c: 22). These "varieties of democracy" render it impossible to represent democracy as a "single point score" (Coppedge et al., 2016c: 22). V-Dem's conceptual part refers to seven major traditions of democratic thinking. The measurement provides by now the corresponding indices for five of them, one being deliberative democracy (Coppedge et al., 2019: 39–41). Each of these high-level democracy indices is comprised of (a) the polyarchy index (reflecting the electoral principle) and (b) a component index measuring the specific understanding of democratic quality. V-Dem's basic understanding of deliberative democratic quality is represented in the "deliberative principle of democracy" which claims that "[. . .]there should [. . .] be respectful dialogue at all levels—from preference formation to final decision—among informed and competent participants who are open to persuasion" (Coppedge et al., 2018a: 41). This section examines the conceptualization, measurement, and aggregation procedure proposed to translate the "deliberative principle" into a measurement of nation states' deliberativeness. The aspects to be measured in this dimension thus concern a reasonable, respectful dialogue about political issues that displays a certain epistemic quality and aims at reaching a decision that can be justified in terms of the common good. The extent to which the deliberative principle is realized in a country is measured by the deliberative component index. Its operationalization essentially draws from established DQI-criteria (Steenbergen et al., 2003): V-Dem applies seven indicators labeled reasoned justification (v2dlreason), common-good orientation (v2dlcommon), the respect for counter arguments (v2dlcountr), and the range of consultation in elite deliberations (v2dlconslt), as well as the extent of public deliberation (v2dlengage), and the target range of national spending (v2dlencmps) and polities in general (v2dlunivl), the latter two being excluded from the index (c.f. Coppedge et al., 2016a: 192–197).

Although V-Dem's explicit acknowledgment of the deliberative democracy-tradition in the course of index-building is highly welcomed, the current approach displays a range of crucial conceptual shortcomings. The most striking feature about V-Dem's measurement is that it lacks a crucial element considered relevant from a theoretical and conceptual point of view: The statement "political decisions [. . .] should be informed by a process characterized by respectful and reason-based dialogue *at all levels*" (Coppedge et al., 2018b: 5; accentuation by t. a.) leads to the conclusion that deliberative quality must also be *measured* at all levels. There is no specification as to what exactly these levels are, although one could assume that a Habermasian differentiation of political center, periphery (public sphere, media and associations) and civil society is implicated (see Habermas, 1996). However, the only items referring to deliberation other than elite deliberation are the indicators "range of consultation" (v2dlconslt) and "engaged society" (v2dlengage). The range of consultation is to be assessed with the following question: "*When important policy changes are being considered, how wide is the range of consultation at elite levels?*" (Coppedge et al., 2018a: 145). The six ordinal categories coders can choose from include circles of increasing range, only the two highest of which encompass members of elites other than the political one (civil society, labor, business). The consultation of (individual) *citizens* is not even an option. Thus, democratic innovations such as mini-publics or citizen juries and informal "everyday political talk" cannot even be registered. The engaged society-indicator addresses the width and independence of *public* deliberations "manifested in discussion, debate, and other public forums such as popular media" (Coppedge et al., 2018a: 146). Again, there are six ordinal categories that mirror the claim of deliberation "at all levels" by taking both elites and non-elites into consideration. The quality of non-elite deliberations is, however, not addressed in this item. The indicators "reasoned justification" (v2dlreason), "common good" (v2dlcommon), and "respect counterarguments" (v2dlcountr) contrastingly measure the *quality* of deliberations, but are merely applied to deliberations among political elites (Coppedge et al., 2018a: 144–145).

In consequence, V-Dem measures the *existence* of deliberation on different societal levels, but does not evaluate the *quality* of deliberation at all these levels: In accordance with the understanding of democracies' deliberativeness represented in the deliberative principle it would, however, be essential that the quality of reason giving-procedures is evaluated for *all levels* of a polity. The reason behind this selectivity seems to be purely pragmatic as an evaluation of deliberation procedures in non-institutionalized contexts is rather hard to accomplish. Building on these analyses concerning *MC*1, further observations can be made for *MC*2, that is, the composition of the concept tree. The indicators do not mirror the conceptualization in a balanced way: Three indicators are concerned with the quality of elite deliberation, and two with the range of elite as well as non-elite deliberation. Thus, there are not only aspects left out (as presented earlier), but there also seems to be an *overrepresentation* of indicators measuring elite deliberations' quality.[10] This imbalance would have to be counteracted in the aggregation process.

In an assessment which is analogous to the DQI's measurement of deliberative quality at the meso level (see Steenbergen et al., 2003), V-Dem measures deliberative quality in political systems with expert evaluations for each of the five indicators named earlier. These indicators resemble each other concerning coding type and scale: Country experts—"typically a scholar or professional with deep knowledge of a country" (Coppedge et al., 2018a: 29)—evaluate the data concerning each variable and range them on an ordinal scale. This scale is linearized by a measurement model using posterior predictions, more

precisely: with a Bayesian item response theory measurement model. How exactly experts come to their evaluations is, however, not transparent. They are asked to evaluate naturally heterogeneous features of a polity by choosing between (at best) six predetermined categories. Being aware of that difficulty, the codebook requests the coders to "base your answer on the style that is most typical of" the respective aspect (Coppedge et al., 2018a: 144–145). This instruction is quite vague in itself. The resulting possibilities for inconsistency are claimed to be mitigated by bridge and lateral coding.[11]

The indicators selected for measuring deliberation remain at a very abstract level. This raises further problems with regard to the measurement's validity (*MC*3 and *MC*4): The cross-system equivalence cannot be reached by aggregating multiple indicators targeting similar aspects across the whole system and cross-checking them through multiple sources, but only by transferring this task to the country experts themselves. This leaves the experts with very much room for interpretation and influence over the actual value of the deliberative component index assigned to a country. The problems caused by the indicators' high level of abstraction can be easily detected based on an analysis of the indicators' categories (*MC*4): The scales of all five indicators used in the deliberative component index are "[o]rdinal, converted to interval by the measurement model" (Coppedge et al., 2018a: 144–145). However, taking the example of the indicator *engaged society* (c.f. Coppedge et al., 2018a: 146), it becomes clear that several dimensions are merged in the categories, thereby not only omitting the possibility of combinations of features beyond those given by V-Dem, but also compromising the "ordinal style" of the indicator:

> 0: Public deliberation is never, or almost never allowed.
>
> 1: Some limited public deliberations are allowed but the public below the elite levels is almost always either unaware of major policy debates or unable to take part in them.
>
> 2: Public deliberation is not repressed but nevertheless infrequent and non-elite actors are typically controlled and/or constrained by the elites.
>
> 3: Public deliberation is actively encouraged and some autonomous non-elite groups participate, but it is confined to a small slice of specialized groups that tends to be the same across issue-areas.
>
> 4: Public deliberation is actively encouraged and a relatively broad segment of nonelite groups often participate and vary with different issue-areas.
>
> 5: Large numbers of non-elite groups as well as ordinary people tend to discuss major policies among themselves, in the media, in associations or neighborhoods, or in the streets. Grass-roots deliberation is common and unconstrained.

Based on a closer assessment of this operationalization, we can observe that at least two dimensions should be differentiated with regard to the six coding options, which are, moreover, treated as cumulative (and thus ordinal) in themselves: First, the range of deliberation from "merely elite" to the inclusion of non-elite deliberation is omitting the possibility of a deliberating public commenting decisions of non-deliberative elites. Second, *rules in law* and *rules in use* are treated as cumulative attributes of one variable, thus negating the possibility of public political deliberation in spite of an official prohibition or the elite's attempt at suppressing it (like, for example, in the Arabian Spring). Because of the omission of democratically relevant possibilities, the ordinal treatment of

both dimensions appears to be invalid. Moreover, the two dimensions do not necessarily have to complement each other, as demonstrated by the example for the second dimension. Therefore, treating the combination of both dimensions as ordinal seems even more problematic.

Aggregation rules must reflect the logical and conceptual relationship between the "aggregated" attributes or components of the index (Møller and Skaaning, 2012; Munck and Verkuilen, 2002; Pickel et al., 2015). As mentioned earlier, the *deliberative democracy index* is based on that *polyarchy index* as well as on the *deliberative component index*. The latter comprises the five indicators associated with deliberation presented earlier. The aggregation level is adequate (*MC*5), since the deliberative component index is considered as a means of assessing the role of deliberation within the country as a whole. The aggregation rules (*MC*6) of the *deliberative component index* are essentially additive formulas, with each indicator being weighted by the strength of its own factor loading.[12] The factor loadings are derived from a Bayesian Factor analysis (BFA),[13] the indicator thus "allow[s] the structure of the underlying data to promulgate through the hierarchy [. . .] while being faithful to the theoretically informed aggregation formula" (Coppedge et al., 2016b: 11). Unfortunately, it remains unclear how exactly the authors intend this aggregation formula to reflect the relationship between the attributes of the concept of "deliberation." It can only be stipulated that the authors of V-Dem view reasoned justification, common-good orientation, the respect for counter arguments, and the range of consultation in elite deliberations as well as the extent of public deliberation as sufficient and (taking into account their respective importance for deliberation as established by the factor analysis) more or less substitutable conditions for realizing the deliberative principle. The weighting with the factor loadings somewhat defuses the notion of the overrepresentation of elite deliberation suspected regarding *MC*2, since the weight of the indicators depends on their *actual* measurable influence on deliberation. Since certain aspects and levels of deliberation are not even measured and thus cannot assert their own weights in the formula, some portion of that critique must, however, be sustained. Finally, the aggregation of the polyarchy index and the deliberative component index to the *deliberative democracy index* (deliberative DI) is achieved by a mix of addition and multiplication similar to the aggregation of the polyarchy index (P being the polyarchy index and HPC—in our case—the deliberative component index; Coppedge et al., 2018b: 9):

$$\mathrm{DI} = 0.25 \times \mathrm{P}^{1.6} + 0.25 \times \mathrm{HPC} + 0.5 \times \mathrm{P}^{1.6} \times \mathrm{HPC}$$

This aggregation rule is justified by the authors as there can be no deliberative democracy without polyarchy *and* deliberation (such that the multiplicative element represents this logic of necessary conditions), as well as by reference to family resemblance-logics allowing for the substitutability of both indices (Coppedge et al., 2015: 7). The exponential logic of the formula induced by the exponentiation of P is explained as follows: "The more a country approximates polyarchy, the more its combined DI score should reflect the unique component" (Coppedge et al., 2018b: 9). This argument does, however, not sufficiently establish why exactly an exponential function using the exponent 1.6 (since Version 6, March 2016) must be applied. It could, for example, also be argued that higher the polyarchy approximation, the higher its weight for the DI score, since its slope increases, especially in comparison with the linear graph of the deliberative component index. In that line of argument, a polyarchy graph resembling a radical function or arcustangents would be more plausible, since the decreasing slope for higher values of the

polyarchy index would give a heavier weight to a linear graph of the deliberative component index (with a constant slope of 1) in comparison. In conclusion, it remains questionable in how far V-Dem's aggregation rules for the deliberative component- and the deliberative democracy-index meet the standards to provide a theoretically founded aggregation procedure.

This section revealed that the measurements of democratic quality proposed by DB and V-Dem fail to capture nation states' deliberativeness for different reasons: Strictly speaking, judged by its own standards, DB does not fail to provide a sound measurement of deliberativeness. Rather, DB does not offer a comprehensive conceptualization and operationalization of "deliberation" or "deliberativeness" to start with. V-Dem, by contrast, presents sophisticated theoretical reflections rooted in the deliberative democracy-tradition and offers a theory-driven conceptualization of nation states' deliberativeness as well as a separate deliberative democracy-index. A detailed assessment of this index based on Munck's and Verkuilen's (2002) criteria nonetheless reveals that the operationalization as well as the aggregation procedure in many cases cannot match V-Dem's own ambitions to provide a thoroughly theory-guided, conceptually and methodologically sound measurement of this deliberativeness: V-Dem's measurement ultimately falls short of translating the presupposed understanding of "deliberativeness" into a measurement that sufficiently reflects the conceptual logic and the underlying theoretical premises.

Three Conceptual Challenges of "Upscaling" Measurements of Deliberative Quality and How to Address Them

Systematic Challenges of Measuring Democratic Deliberation

DB and V-Dem provide highly sophisticated, "theory-driven" measurements of democratic quality. Nevertheless, our examination of their approaches concluded that they display deficits when it comes to integrating a measurement of democracies' deliberativeness. In this section, we will argue that a good deal of these shortcomings can be traced back to systematic challenges involved in any attempt to measuring deliberation at the nation state-level and indicate strategies for addressing them. While the DQI (Steenbergen et al., 2003) provides the gold standard for such measurements at the meso level, the conceptualization, operationalization, and aggregation procedure applied for measuring deliberative quality at the macro level must come to terms with different challenges. Based on recent developments in deliberative theory and research, this section will outline the specific challenges of conceptualizing and measuring nation states' deliberativeness (section "Systematic Challenges of Measuring Democratic Deliberation"), examine to what extent the approaches offered by DB and V-Dem are able to meet them (section "Do DB and V-Dem Meet the Systematic Challenges of Measuring Deliberation?"), and offer perspectives for developing a valid assessment of democracies' deliberativeness (section "Upscaling measurements of deliberative quality: Lessons to learn for measuring deliberativeness").

Although, there are certainly "no absolute criteria" for evaluating the conceptualization of deliberative quality, its adequacy "can nevertheless be evaluated in terms of methodological standards as to how well the term [. . .] is aligned with the phenomenon it defines" (Pickel et al., 2015: 504–505). In assessing a polity's deliberativeness, scholars need to examine a broad variety of deliberative practices that occur throughout the political system:

In a democracy, *reason giving* on political issues and decisions takes place in various institutionalized and non-institutionalized for a such as parliaments, constitutional courts, the civil society, "designed" democratic innovations, mass media or social media platforms. The *first* conceptual *challenge* of measuring deliberativeness results from the heterogeneity of deliberative sites in real world-polities: To provide a comprehensive measurement of nation states' deliberative quality, scholars must presuppose a clear conceptual account of the sites within a polity that should be considered. This is even more significant since the character of deliberations is likely to differ between those sites (see Fleuß et al., 2018; Esau et al., in press; Pedrini, 2014: 18). This finding can—at least partly—be traced back to the fact that deliberative procedures in different sites of democratic systems fulfill different functions. A valid measurement of nation states' deliberativeness therefore must also be able to deal with the heterogeneity of communicative styles in different sites which are subsumed under the heading "deliberation" (*SC*1).

Against this background, the *second challenge* consists in providing a suitable concept of "deliberation" and "deliberative quality" which is applicable to heterogeneous deliberative practices occurring throughout a political system. Scholars tend to disagree with regard to the *rigidity* of criteria that should be applied for distinguishing deliberative from non-deliberative communicative practices (e.g. Goodin, 2018; Owen and Smith, 2015). Original normative approaches used a narrow concept of deliberation: "deliberation" was conceptualized as "the exchange of rational (non-emotional), neutral, impartial arguments" (Fleuß et al., 2018: 15, see Cohen, 1989; Habermas, 1996). In the course of deliberative theories' development, "we see a definite expansion of sorts of things that could be considered arguments and reasons" (Chambers, 2003: 322; see Fleuß et al., 2018: 15).[14] The conceptualization chosen within this spectrum of wider or more narrow concepts of deliberation has crucial impact on the measurement of deliberative quality as different concepts implicate divergent evaluation standards: In extreme cases, real world-deliberations can fully satisfy the standards derived from a broad concept of deliberation although they would be evaluated as "performing poorly" if a classical Habermasian concept was applied (*SC*2).

The *third challenge* results from the fact that the complex interplay and interactions of these heterogeneous forms of deliberation must be reflected to provide a valid measurement of deliberativeness at the nation state-level. From a theoretical point of view, this challenge is addressed most explicitly in 4th generation deliberative thinking, that is, the systemic approach to deliberation as developed by Mansbridge et al. (2012) and Dryzek (2000; Dryzek and Niemeyer, 2010). One suitable strategy for reflecting interactive relationships in the overall score for a polity's deliberative quality consist in developing an aggregation formula that mathematically mirrors individual deliberations' interactions (see Fleuß et al., 2018: 17–19). From these systematic reflections on the requirements of conceptualizing and measuring nation states' deliberative quality, we derive the following list of systematic challenges (SC) that we will refer to in further evaluating the measurement approaches proposed by DB and V-Dem:

*SC*1: *Does the index differentiate between different types of democratic deliberation that may occur in different sites of a political system* (*having different normative goals, fulfilling different functional roles, etc.*)?

*SC*2: *Does the conceptualization decide upon either a broad or a narrow concept of deliberation* (*or does it, alternatively, distinguish different concepts of deliberation and measure/aggregate them separately*)?

*SC*3: *Does the index consider the interactive relationships between different types of deliberative practices and their interplay's impact on the deliberative performance of the democratic system as a whole*?

Do DB and V-Dem Meet the Systematic Challenges of Measuring Deliberation?

Because DB only scarcely and non-systematically considers deliberative quality, our evaluation with regard to its meeting the requirements outlined in section "Systematic Challenges of Measuring Democratic Deliberation" can be brief. Although we were able to demonstrate that DB integrates bottom-up deliberative participatory means in different branches of its concept tree (see section "The Democracy Barometer"), it does not offer a systematic assessment of different deliberative sites in a polity (*SC*1). Due to the lack of a systematic conceptualization of deliberation and deliberative quality, a detailed discussion of the concept of deliberation applied (SC2) or asking whether deliberative practices' interactive relationships are considered (SC3) is dispensable. V-Dem's deliberative component index, on the other hand, puts a good deal of effort in providing a theoretically reflected concept of deliberation. V-Dem claims to aim at a measurement of deliberative quality "at all levels of society." As outlined in section "The Varieties of Democracy-Project", this approach lacks specification with regards to the exact sites of deliberation that are to be taken into account: The codebook requires experts to evaluate the quality of deliberations occurring "at all levels," but rather vaguely specifies that this includes "political elites" and the "civil society" (Coppedge et al., 2018b: 5). In consequence, the authors of V-Dem are certainly aware of the fact that deliberation occurs in different institutional and societal contexts, but they do not offer a systematic account of these deliberative sites (*SC*1). In addition, the authors do not systematically distinguish different styles, contexts and functions of deliberative procedures, for example, in commonly binding decision-making or public opinion and will-formation. In consequence, neither DB's nor V-Dem's theoretical groundworks and conceptual frameworks are able to account for different types of democratic deliberation that occur in different contexts and fulfill different functions within the polity at large. As these contexts and forms of deliberation and their potential impact on the overall deliberative quality are not reflected on a conceptual level, the measurement does not differentiate between them either. It might be expected that country experts, in coding a polity's deliberative quality, are automatically integrating this "context-sensitivity" in applying the coding-criteria, for example, to parliamentary procedures and civil society deliberations. As the codebook and coding instructions do not only imply a focus on elite deliberation, but also lack specific instructions for different deliberative sites (see Coppedge et al., 2018a: 40–51, 2018b: 18), this nevertheless seems to be left to pure chance.

Throughout its assessment, V-Dem presupposes a rather narrow, classical ("Habermasian") concept of deliberation which is outlined in the "deliberative principle" (Coppedge et al., 2018a: 41). Potential biases or distortions in evaluations of non-institutionalized, less formal civil society deliberations that may result from applying these standards without further differentiations to all deliberative sites are not further considered in the codebook or instructions of country experts (*SC*2). Meeting challenge *SC*3, that is, to reflect the interactive relationships between individual instances of deliberation, is impeded by the measurement approach V-Dem provides: Country experts do not evaluate individual instances of deliberation separately, but offers expert evaluations

of the polity's deliberative qualities which remain at a high level of abstraction. Due to the indicators' high level of abstraction, a *systematic* assessment of interactions between deliberations in different sites is impossible. Therefore, V-Dem is at least at danger to miss one crucial feature of a well-functioning democracy: Civil society deliberations or deliberative mini-publics contribute to democratic quality *only if* the results of opinion- and will-formation processes are actually transmitted into the empowered, decision-making branch of the polity (Dryzek and Niemeyer, 2010: 11–12).

Upscaling Measurements of Deliberative Quality: Lessons to Learn for Measuring Deliberativeness

Based on our examinations of DB's and V-Dem's assessments of deliberation and the reflections offered in sections "Systematic Challenges of Measuring Democratic Deliberation" and "Do DB and V-Dem Meet the Systematic Challenges of Measuring Deliberation?" of this article, we are able to identify one overarching requirement for measurements of nation states' deliberative quality: Any measurement approach must consider the specific characteristics of deliberative quality at the nation state level. By considering a nation state's deliberativeness merely analogous to a small-scale setting's deliberativeness, essential elements and features may be overlooked in its conceptualization and accordingly will not be appropriately reflected in the measurement and aggregation procedure. V-Dem's application of meso-level standards to the overall deliberative quality of a polity is an illustrative example for this. In order to provide a valid assessment of nation states' deliberative quality, scholars accordingly need to respect the diversity of sites (and styles) of deliberation within a polity and must reflect this acknowledgment in their measurement approach. We argued that it is crucial for measurements of nation states' deliberativeness to reflect at least three core features, that is, to respect the *diversity of deliberative sites* and *deliberative practices* within a polity in the conceptualization and measurement and to integrate an assessment of their *interactiveness* in the overall measurement approach.

Upon careful consideration, this conceptualization of deliberative quality and the underlying theoretical reflections predetermine that certain *methodological* strategies are (not) suitable: In our assessment of DB, we pointed out that exclusively relying on "hard data" and *rules in law* can hardly be considered appropriate for measuring deliberative quality. Expert evaluations of polities (based on DQI-analogous criteria) might, however, appear as a feasible and resource-saving strategy. Based on a close examination of V-Dem's deliberative component index, it nevertheless becomes clear that these expert evaluations must be conducted at the appropriate level of abstraction to avoid an invalid measurement due to conceptual mismatches as well as reliability problems. A similar mismatch between the concept of "(nation states') deliberativeness" and the corresponding index results from applying inappropriate aggregation rules: As we pointed out in *SC*3, a valid measure of *large scale* deliberative quality cannot just accumulate the (DQI-) scores of individual deliberative procedures that occur, for example, in a parliament, the media, and civil society organizations. It must also reflect their interactive relationships. Introducing such differentiations to "upscale" existing measurements of deliberation will largely increase the complexity of the measurement approach. It will require an assessment that combines different methodological strategies.

Recent deliberative scholarship provides a promising point of departure for *SC*1 and *SC*2, that is, a context-sensitive measurement of deliberative procedures' quality. Bächtiger and Parkinson (2019: 43) argue that "the configuration of deliberation and its relationship

to other communication modes depends on goals and contexts, [while] the deliberative core remains stable" and suggest to adapt established DQI-measurements in line with this conceptualization (Bächtiger and Parkinson 2019: 70–78; see Gerber et al., 2018). With regard to strategies for addressing *SC*3, case studies could prove computerized content-analyses of discourses (such as structural topic modeling) and network analyses to be suitable methodological approaches for assessing the interactive relationships between different deliberative practices within the democratic system (Bächtiger and Parkinson 2019; Beste, 2016; Gerber, 2015: 142–150). An appropriate aggregation rule for combining the scores reached in these different assessments must comply with three basic conceptual premises: (a) In weighting individual deliberative procedures' quality, it must reflect their relative importance for a systems' overall deliberativeness. The normative perspective of deliberative theory suggests that deliberations in the public space must be attributed a higher weight than deliberations in the empowered space.[15] Its operations[16] need to reflect (b) that the transmission of claims and topics of deliberations from the public to the empowered space is a *necessary condition* and (c) that individual deliberations in different sites of a polity are, at least to a certain degree, *substitutable* and able to *compensate* for a lack of deliberative quality in other sites (see Fleuß et al., 2018: 18; Møller and Skaaning, 2012).[17]

Summary

Deliberative democratic theories formulate a major paradigm of democratic legitimacy. This article argues that measurements of democratic quality must include an assessment of nation states' deliberativeness. Contemporary measurements of democratic quality such as DB and V-Dem aim at providing "basically theory-driven" measurements of democratic quality. Although DB and V-Dem differ with regard to their democratic theoretical frameworks and premises, they consider nation states' deliberativeness as an important feature of democratic quality. In their measurements both indices do, however, fall short of complying with the standards for measuring deliberativeness as laid out in this article (*MC*1–*MC*6 and *SC*1–*SC*3). We examine the methodological, pragmatic, and systematic reasons for these shortcomings. The crucial task for measurements of nation states' deliberativeness consists in providing a conceptual approach and methodological framework for "upscaling" existing meso-level measurements (such as the DQI). Although recent research provides useful pointers for addressing the systematic challenges coming along with this task, Bächtiger and Parkinson (2019: 151) are right to "advocate some modesty when measuring deliberation and deliberativeness." As the final section of this article indicates, "only a plurality of methodological tools will unravel the deliberative dimensions of democratic systems." Such measurements must, however, not only be complemented by "thorough, theoretically informed, well-grounded interpretive work by researchers who have taken the time to understand meanings from the insider's point of view" (Bächtiger and Parkinson, 2019: 151). There may also be no one-fits-all solution for upscaling existing measurements and integrating them in existing indices. Rather, the methodological strategies suitable for this undertaking will necessarily depend on democracy indices' normative premises, their overall structure and preceding methodological preferences and determinations.

Acknowledgements

The authors are indebted to André Bächtiger, Simon Niemeyer, Toralf Stark, Gary S. Schaal and Philipp Weinmann as well as two anonymous reviewers for commenting on earlier versions of this article. They would

like to thank Friederike Uhl for her support in finalizing this article. Parts of this article are based on a working paper that was presented at a NCCR Democracy-Workshop at Zürich University in 2017 by Gary S. Schaal, Karoline Helbig, and Dannica Fleuß.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Dannica Fleuß https://orcid.org/0000-0003-4214-4806

Notes

1. Most measurements of democracy presuppose a liberal conceptualization of democracy and democratic quality. Liberal democratic theorists highlight the functional value of processes of public deliberation for democratic quality (see Dahl, 1971; Locke, 2016; Rawls, 2011).
2. For a differentiation of measurements of *democracy* and *democratic quality* and the respective standards of assessment, see Pickel et al. (2015).
3. This article focuses on strategies for measuring deliberativeness in *established* democracies. Therefore, the measurement proposed by the Bertelsmann Transformation Index is not taken into account (see Bertelsmann Foundation, 2018). In addition to being theory-driven (which distinguishes them from indices such as the SGI (Pickel et al., 2015: 519; see Schraad-Tischler and Seelkopf, 2018)), DB and V-Dem are also remarkably transparent with regard to their methodology and data bases (which distinguishes them from, for example, the EIU, see Coppedge et al., 2011: 251). Both characteristics make DB and V-Dem prime candidates for assessing the merits and shortcomings of contemporary measurements of nation states deliberative quality.
4. This analogous assessment of DB and V-Dem is not meant to obscure the fact that DB and the V-Dem project at large pursue different goals. We will elaborate on these comparability problems below (sections "The Democracy Barometer" and "The Varieties of Democracy-Project").
5. Further development of the framework (Munck, 2009). For a critical discussion of the framework see Fuchs and Roller (2009) and Pickel et al. (2015).
6. In the Democracy Barometer: "function"; for comparability reasons, the original terminology is transferred into the terminology of Munck and Verkuilen (2002). It should be pointed out here that the authors of DB are well-aware of the fact that not all of the indicators used to measure democratic quality can be logically derived from the (sub-)components at higher levels of abstraction and that, due to the vast number of indicators, it is "nearly impossible to avoid redundancies and conflations and to provide a thorough theoretical foundation up to the final leaves of the concept tree" (see Pickel et al., 2015: 510).
7. For example, in the attribute "freedom," the component "public sphere" and various subcomponents appear freedom to associate (Constass), trade union density and membership in professional organizations (Union, Memproorg), membership in public interest organizations (Memhuman, Memenviron), freedom of opinion, speech, and the press (Constspeech, Constrpress), pluralism of the media (Newsimp, Newspaper), and the neutrality of the media (Balpress, Neutrnp). In the attribute "equality," the freedom of information (RestricFOI, EffFOI), legal and political environment of press freedom (Legmedia, Polmedia) are positioned under the component "transparency."
8. At least the wording "(o)verall, about 300 indicators were collected from existing data sets as well as produced or calculated by the project team on the basis of various types of documents and information. From this collection 105 indicators were selected to build the Democracy Barometer" (Merkel et al.,2016a: 7; see also Bühlmann et al., 2012: 131; Merkel 2016b: 3) hints to that.
9. The indicators for the components and subcomponents are selected "to ensure content validity and to prevent concept overstretching" (Merkel et al., 2016b: 4). Inter alia, every aspect is measured by two different indicators that "do so in a different fashion or originate from different sources," and every component should include rules in law as well as rules in use (Merkel et al., 2016b: 4). Extracting an indicator would counteract these guidelines.

10. Our examination of V-Dem's indicators for elite and non-elite deliberations reveals that they are located at different analytical levels (albeit at the same level of the concept tree).
11. " [. . .] we encourage Country Experts to conduct bridge coding (coding of more than one country through time) and lateral coding (coding limited to a single year–2012). The purpose of this additional coding is to assure cross-country equivalence by forcing coders to make explicit comparisons across countries. This helps the measurement model estimate, and correct for, systematic biases across coders and across countries [. . .]" (Coppedge et al., 2016b: 26).
12. The authors thank Jan Teorell of the V-Dem Project for his notes on that matter.
13. Probably the BFA is executed for all five indicators at the level of country-year as analogously described for the Women Civil Liberties Index (c.f. Sundström et al., 2015: 13). Concerning this method, some critical remarks seem in order: The factor loadings used for weighting the indicators are not constant, since with every new variable being integrated in the draws for the BFA the factor loading can potentially change (if only by fractions). So either the data sets have to be updated not only with new values but with newly computed indices for all prior years—thus making the data sets over the years incompatible and old publications unreliable. Or the data of every year have different aggregation formulas because of the different weights and are thus not directly comparable. Furthermore, the values within one index are not independent of each other, since the whole data set of the indicators involved is reflected in their weights in the aggregation formula. In the long-run, for example general world-wide changes of democracy might be less visible, since the trend of the new data is outweighed by the prior data.
14. For a critical examination of these extensions of the concept of deliberation see Owen and Smith (2015) and Goodin (2018).
15. The exact weights of public and empowered space-deliberations can then be determined either based on theoretical considerations or by applying an empirical strategy such as a factor analysis.
16. We suggest to calculate the overall score for a nation states' "deliberativeness" by combining multiplicative operations (b) with additive operations (c).
17. From the perspective of contemporary systemic deliberative theory, "[t]hough there may be little or no perfect democratic deliberation in any site, the collective work done across the system may still produce a suitably deliberative democratic whole" (Boswell et al., 2016: 263, see Fleuß et al., 2018: 18).

References

America in One Room (2019) Executive Summary and Overall Results. Available at: https://cdd.stanford.edu/2019/america-in-one-room-results/ (accessed 21 October 2019).

Bächtiger A and Parkinson J (2019) *Mapping and Measuring Deliberation: Towards a New Deliberative Quality*. Oxford: Oxford University Press.

Bertelsmann Foundation (2018) *Transformation Index BTI 2018: Governance in International Comparison*. Gütersloh: Verlag Bertelsmann Stiftung.

Beste S (2016) 'Legislative Frame Representation': Towards an Empirical Account of the Deliberative Systems Approach. *The Journal of Legislative Studies* 22 (3): 295–328.

Boswell J, Hendriks CM and Ercan SA (2016) Message Received? Examining Transmission in Deliberative Systems. *Critical Policy Studies* 10 (3): 263–283.

Bühlmann M, Merkel W, Müller L, et al. (2008) Wie lässt sich Demokratie am besten messen? Zum Forumsbeitrag von Thomas Müller Und Susanne Pickel. *Politische Vierteljahresschrift* 49 (1): 114–122.

Bühlmann M, Merkel W, Müller L, et al. (2012) Demokratiebarometer: Ein neues Instrument zur Messung von Demokratiequalität. *Zeitschrift für Vergleichende Politikwissenschaft* 6 (S1): 115–159.

Caluwaerts D and Reuchamps M (2015) Strengthening Democracy through Bottom-Up Deliberation: An Assessment of the Internal Legitimacy of the G1000 Project. *Acta Politica* 50 (2): 151–170.

Chambers S (2003) Deliberative Democratic Theory. *Annual Review of Political Science* 6 (1): 307–326.

Cohen J (1989) Deliberation and Democratic Legitimacy. In: Pettit P and Hamlin A (eds) *The Good Polity: Normative Analysis of the State*. Oxford: Polity Press, pp.17–34.

Coppedge M, Gerring J, Altman D, et al. (2011) Conceptualizing and Measuring Democracy: A New Approach. *Perspectives on Politics* 9 (2): 247–267.

Coppedge M, Gerring J, Lindberg SI, et al. (2016a) V-Dem Codebook. v6. Varieties of Democracy (V-Dem) Project. Available at: https://www.v-dem.net/ (accessed 18 April 2017).

Coppedge M, Gerring J, Lindberg SI, et al. (2016b) V-Dem Methodology. v6. Varieties of Democracy (V-Dem) Project. Available at: https://www.v-dem.net/ (accessed 18 April 2017).

Coppedge M, Gerring J, Lindberg SI, et al. (2018a) V-Dem Codebook. v8. Varieties of Democracy (V-Dem) Project. Available at: https://www.v-dem.net/ (accessed 26 November 2018).

Coppedge M, Gerring J, Lindberg SI, et al. (2018b) V-Dem Methodology. v8. Varieties of Democracy (V-Dem) Project. Available at: https://www.v-dem.net/ (accessed 26 November 2018).
Coppedge M, Gerring J, Lindberg SI, et al. (2019) V-Dem Codebook. v9. Varieties of Democracy (V-Dem) Project. Available at: https://www.v-dem.net/ (accessed 10 July 2019).
Coppedge M, Lindberg SI, Skaaning S, et al. (2015) Measuring High Level Democratic Principles Using the V-Dem Data, Working Paper Series 6, The Varieties of Democracy Institute. Available at: https://www.v-dem.net/ (accessed 19 April 2017).
Coppedge M, Lindberg SI, Skaaning S-E, et al. (2016c) V-Dem Comparisons and Contrasts with Other Measurement Projects. *V-Dem Varieties of Democracy*. Available at: https://www.v-dem.net/ (accessed 19 April 2017).
Dahl RA (1971) *Polyarchie; Participation and Opposition*. New Haven, CT; London: Yale University Press.
Dryzek JS (2000) *Deliberative Democracy and Beyond: Liberals, Critics, Contestations*. Oxford: Oxford University Press.
Dryzek JS (2015) Deliberative Engagement: The Forum in the System. *Journal of Environmental Studies and Sciences* 5 (4): 750–754.
Dryzek JS and Niemeyer S (2010) *Foundations and Frontiers of Deliberative Governance*. Oxford: Oxford University Press.
Elstub S (2010) The Third Generation of Deliberative Democracy. *Political Studies Review* 8 (3): 291–307.
Elstub S (2015) A Genealogy of Deliberative Democracy. *Democratic Theory* 2 (1): 1–19.
Esau K, Fleuß D and Nienhaus S-M (in press) Different Arenas, Different Deliberative Quality? Using a Systemic Framework for Evaluating Online Deliberation on Immigration Policy in Germany. *Policy & Internet*, Special Issue "Political Online Participation and its Effects".
Fishkin JS (2018) *Democracy When the People Are Thinking: Revitalizing Our Politics through Public Deliberation*. Oxford: Oxford University Press.
Fleuß D, Helbig K and Schaal GS (2018) Four Parameters for Measuring Democratic Deliberation: Theoretical and Methodological Challenges and How to Respond. *Politics and Governance* 6 (1): 11–21.
Freedom House (2019) Methodology 2019. Available at: https://freedomhouse.org/report/methodology-freedom-world-2019 (accessed 10 June 2019).
Fuchs D and Roller E (2009) Die Konzeptualisierung der Qualität von Demokratie. Eine kritische Diskussion aktueller Ansätze. In: Brodocz A, Llanque M and Schaal GS (eds) *Bedrohungen der Demokratie*. Wiesbaden: VS Verlag für Sozialwissenschaften, pp.77–96.
G 1000 (2012) Manifesto. Available at: http://www.g1000.org/en/manifesto.php# (accessed 21 October 2019).
Gerber M (2015) Equal Partners in Dialogue? Participation Equality in a Transnational Deliberative Poll (Europolis). *Political Studies* 63: 110–130.
Gerber M, Bächtiger A, Shikano S, et al. (2018) Deliberative Abilities and Influence in a Transnational Deliberative Poll (EuroPolis). *British Journal of Political Science* 48 (4): 1093–1118.
Goodin RE (2018) If Deliberation Is Everything, Maybe It's Nothing. In: Bächtiger A, Dryzek JS, Mansbridge J, et al. (eds) *Oxford Deliberative Democracy Handbook*. Oxford: Oxford University Press, pp.883–899.
Habermas J (1996) *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Cambridge: Polity Press.
Jäckle S, Uwe W and Bauschke R (2012) Das Demokratiebarometer: "basically Theory Driven"? *Zeitschrift für Vergleichende Politikwissenschaft* 6 (1): 99–125.
Locke J (2016) *Second Treatise of Government and A Letter Concerning Toleration*. Oxford: Oxford University Press.
Møller J and Skaaning S-E (2012) Concept-Measure Inconsistency in Contemporary Studies of Democracy. *Zeitschrift für Vergleichende Politikwissenschaft* 6: 233–251.
Mansbridge J (1999) Everyday Talk in the Deliberative System. In: Stephen M (ed.) *Deliberative Politics: Essays on Democracy and Disagreement*. Oxford: University Press, pp.211–238.
Mansbridge J, Bohman J, Chambers S, et al. (2012) A Systemic Approach to Deliberative Democracy. In: Parkinson J and Mansbridge J (eds) *Deliberative Systems, Deliberative Democracy at the Large Scale*. New York: Cambridge University Press, pp.1–26.
Marshall M and Jaggers K (2018) *Polity IV Project: Dataset Users' Manual* (Polity IV Data Computer File, Version P4v2004, Center for International Development and Conflict Management, University of Maryland, College Park, Maryland). Center for Global Policy, George Mason University. Available at: www.systemicpeace.org/polityproject.html (accessed 10 June 2019).
Merkel W and Croissant A (2003) Liberale und defekte Demokratien. In: Karl S (ed.) *Herausforderungen der Repräsentativen Demokratie*. Baden-Baden: Nomos, pp.55–88.

Merkel W, Bochsler D, Bousbah K, et al. (2016a) *Democracy Barometer: Codebook*, version 5. Aarau: Zentrum für Demokratie.

Merkel W, Bochsler D, Bousbah K, et al. (2016b) *Democracy Barometer: Methodology*, version 5. Aarau: Zentrum für Demokratie.

Munck GL (2009) *Measuring Democracy: A Bridge between Scholarship and Politics*. Baltimore, MD: Johns Hopkins University Press.

Munck GL and Verkuilen J (2002) Conceptualizing and Measuring Democracy: Evaluating Alternative Indices. *Comparative Political Studies* 35 (1): 5–34.

Niemeyer S, Curato N and Bächtiger A (2015) Assessing the Deliberative Capacity of Democratic Polities and the Factors That Contribute to It. In: *Democracy: A Citizen Perspective Conference*, Turku, 27–28 May. Available at: http://www.abo.fi/fakultet/media/33801/niemeyerbachtigercurato_assessingthecapacity-ofdeliberativesystems.pdf (accessed 19 April 2018).

Owen D and Smith G (2015) Survey Article: Deliberation, Democracy, and the Systemic Turn. *Journal of Political Philosophy* 23 (2): 213–234.

Pedrini S (2014) Deliberative Capacity in the Political and Civic Sphere. *Swiss Political Science Review* 20 (2): 263–286.

Pickel S and Pickel G (2007) *Politische Kultur- und Demokratieforschung*. Wiesbaden: Springer-Verlag.

Pickel S, Stark T and 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.

Rawls J (2011) *Political Liberalism: Expanded Edition*. New York: Columbia University Press.

Schraad-Tischler D and Seelkopf L (2018) *Concept and Methodology: Sustainable Governance Indicators*. Gütersloh: Verlag Bertelsmann Stiftung.

Steenbergen MR, Bächtiger A, Spörndli M, et al. (2003) Measuring Political Deliberation: A Discourse Quality Index. *Comparative European Politics* 1: 21–48.

Sundström A, Paxton P, Wang Y, et al. (2015) Women's Political Empowerment: A New Global Index, 1900 2012. Varieties of Democracy (V-Dem) Project, Working Paper Series, 19. Available at: https://www.v-dem.net/ (accessed 19 April 2018).

Vanhanen T (2000) The Polyarchy Dataset: Vanhanen's Index of Democracy. Available at: https://www.prio.org/Data/Governance/Vanhanens-index-of-democracy/ (accessed 28 November 2019).

We the Citizens (2011) Final Report. Available at: http://www.wethecitizens.ie/wp-content/uploads/2015/05/We-the-Citizens-2011-FINAL.pdf (accessed October 21 2019).

Author Biographies

Dannica Fleuß is a postdoctoral research fellow and lecturer in Political Science at Helmut-Schmidt-University (Hamburg) and a research associate at the Centre for Deliberative Democracy and Global Governance (University of Canberra). Her research and teaching experience include political theory, measurements of democratic quality, and empirical studies on deliberation. After finishing her PhD in 2016 at Heidelberg University on the political philosophy of proceduralism, she started her postdoctoral project at Helmut-Schmidt-University that develops a theory-driven empirical assessment of deliberation in Western democracies.

Karoline Helbig is a doctoral researcher for Political Theory at Berlin Social Science Centre and the Weizenbaum Institute for the Networked Society. She studied sociology and mathematics and is currently working on her PhD project regarding deliberative democratic theory and procedures in the context of digitalized societies. Her further research interests include measuring democracy, methods of social research, democratic theory, digitalization, and the sociology of knowledge.