Das Demokratiebarometer: Führt Quantität zwangsläufig zu Validität?

Uwe Wagschal – Universität Freiburg
Sebastian Jäckle – Universität Freiburg
Rafael Bauschke – Universität Heidelberg

‚Prädemokratie', ‚Postdemokratie', ‚Autokratie'? Zum Stand vergleichender Herrschaftsforschung
Tagung der Sektion „Vergleichende Politikwissenschaft" an der Philipps-Universität Marburg,

vom 29. bis 31. März 2012

work in progress – bitte nicht zitieren!

Über Kommentare freuen wir uns hingegen sehr!

Korrespondenzanschrift:

Dr. Sebastian Jäckle
Seminar für Wiss. Politik
Werthmannstr. 12
79098 Freiburg i. Br.

phone: +49 - 761 - 203 9368
e-mail: sebastian.jaeckle@politik.uni-freiburg.de

Einleitung: Ein neuer Fokus der Demokratiemessung?

In den vergangenen zwei Jahrzenten entwickelte sich die empirische Demokratieforschung von einem politikwissenschaftlichen Randthema zu einem der wichtigsten Forschungsbereiche der vergleichenden Politikwissenschaft. Selbst wenn man sich auf die deutschsprachige Literatur beschränkt zeugen zahlreiche Publikationen von diesem Trend. Hier seien nur einige stellvertretend für die Breite der vorhandenen Arbeiten genannt: (Abromeit 2004; Berg-Schlosser 1999: 9; Fuchs 2004; Lauth 2004; 2008; Merkel et al. 2003; Merkel et al. 2006; Müller und Pickel 2007). Seit geraumer Zeit hat das Thema auch einen Stammplatz in den meisten Curricula politikwissenschaftlicher Studiengänge. Gerade in den letzten Jahren wurden eine ganze Reihe neuer Demokratiemaße entwickelt, so dass sich dem Forscher eine Vielzahl von Messkonzepten anbieten. Tabelle A1 im Anhang gibt einen Überblick in Bezug auf ihre Autoren, das jeweils zugrunde liegende Demokratieverständnis, ihre ländermäßige sowie zeitliche Abdeckung sowie die verwendete Operationalisierung. Mit der gestiegenen Relevanz einher geht auch eine Ausdifferenzierung der behandelten Forschungsfragen. Im Zentrum stehen dabei Fragen zu den Zusammenhängen und den kausalen Einflussmechanismen von Demokratie (und Autokratie) auf gesellschaftliche Phänomene wie etwa Wachstum, Wohlstand, Entwicklung und politische Konflikten (Büger 2007; Geis und Wagner 2006; Lauth et al. 2000; Müller 2002; Obinger 2001; Sunde 2006).

Auch lässt sich eine Verschiebung der Schwerpunktsetzung in der Auseinandersetzung mit der Demokratiemessung beobachten. Während das zentrale Forschungsanliegen in den vergangen Jahrzehnten insbesondere auf der Abgrenzung zwischen demokratischen und nichtdemokratischen Systemen lag, stehen in der aktuellen Debatte und Forschung zunehmend die Fragen der Messmethodik sowie der Binnendifferenzierung demokratischer Systeme im Vordergrund. Statt sich mit der Frage zu beschäftigen, wann ein politisches System als Demokratie bezeichnet werden kann, untersucht eine zunehmende Zahl von Beiträgen und Forschungsprojekten stattdessen die Frage, inwiefern Qualitätsunterschiede innerhalb der Demokratien existieren und wie diese einer Messung zugänglich gemacht werden können. Die Ursprünge dieser Qualitätsdebatte lassen sich primär in den stärker theoretisch orientierten Arbeiten zur „Quality of Democracy“ (Beetham et al. 2009; Diamond und Morlino 2005) verorten. Parallel hierzu lässt sich insbesondere in der deutschsprachigen Literatur eine deutliche Zunahme an Beiträgen beobachten, die die Debatte über die adäquate „Vermessung“ von Demokratien sowohl aus theoretisch-konzeptioneller als auch aus empirisch-messmethodischer Sicht deutlich vorangetrieben haben (Abromeit 2004; Fuchs 2004; Fuchs und Roller 2008; Lauth 2004; Müller und Pickel 2007; Stoiber 2007) oder zumindest auf den fahrenden Zug aufgesprungen zu sein scheinen (Campbell und Barth 2009).

Als einer der relevanten Auslöser für die intensive Beschäftigung mit der Thematik kann die extrem geringe Varianz angeführt werden, die vorhandene Demokratiemaße für die generell als entwickelte Demokratien bezeichneten Länder angeben. Vergleicht man die Demokratieindizes des

Demokratiebarometers, von Freedom House, Polity IV, Vanhanen, dem Demokratie-Index des Economist, dem Status-Index des SGI, das Demokratie-Ranking sowie dem Unified Democracy Score und, so lässt sich für das Sample der OECD Staaten eine unterschiedliche Varianz feststellen (vgl. Tabelle 1). Bei den ordinalen Messkonzepten ist die Spannweite (Differenz zwischen Minimum und Maximum) entweder Null oder sehr gering. Die Indizes mit Intervallskalenniveau weisen – da feiner gemessen wird – eine tendenziell höhere Varianz und Spannweite auf. Selbst ein Index wie die SGI der Bertelsmann Stiftung, der dezidiert für die OECD-Welt entwickelt wurde, weißt immer noch eine vergleichsweise geringe Varianz in der Demokratiekomponente auf. Allerdings zeigt sich auch, dass das Demokratiebarometer – dessen selbst erklärtes Ziel eine Erhöhung der Varianz unter den etablierten Demokratien ist – nicht die höchste Varianz aufweist (gemessen am Variationskoeffizienten CV). Zudem werden wir zeigen, dass diese Varianz vor allem durch das Aggregationsverfahren technisch erzeugt wurde und somit höchst problematisch zu bewerten ist.

Dennoch sind Zweifel an der Ähnlichkeit des demokratischen Niveaus in den etablierten Demokratien angebracht. Oft zitierte Beispiele bilden Italien unter Berlusconi oder die USA unter der Bush Administration (Merkel et al. 2006). Diese auf anekdotischer Evidenz basierenden Einschätzungen mangelt es jedoch an einer grundlegenden Systematik. Gleichwohl geraten vor dem Hintergrund der Diskrepanz zwischen Messeergebnissen und subjektiver Perzeption, die etablierten Messansätze zunehmend in die Kritik. Das Kernproblem wird hierbei in der Tatsache gesehen, dass bestehende Konzepte zwar eine größtenteils robuste Differenzierung von Systemtypen ermöglichen[1], jedoch nicht die Qualitätsunterschiede abbilden können, die sich bei einer genaueren, qualitativen Betrachtung zwangsläufig feststellen lassen (Lauth 2004; Lauth et al. 2000; Merkel et al. 2003). Während sich einige wenige Autoren aufgrund der geschilderten Problematik mittels vorhandener Messkonzepte die demokratischen Länder adäquat zu differenzieren für eine generelle Abkehr von der quantitativen Messung aussprechen (Abromeit 2004), versuchen andere die Schwierigkeiten mehr oder minder pragmatisch und konstruktiv anzugehen (Foweraker und Krznaric 2000; Fuchs 2004; Stoiber 2007; 2011).

Der jüngste in dieser Tradition stehende Beitrag ist das von einer Forschergruppe um Marc Bühlmann und Wolfgang Merkel entwickelte Demokratiebarometer. Laut den Autoren ist es das erklärte Ziel ihres Messansatzes: „die konzeptionellen und methodologischen Schwächen bisheriger Demokratiemaße zu überwinden, um so die Qualitätsunterschiede von etablierten Demokratien messen und analysieren zu können“ (Bühlmann et al. 2011a).

[1] Doch auch im Hinblick auf die Leistungsfähigkeit bei der Unterscheidung zwischen Autokratien und Demokratien sehen sich die Messansätze etwa aufgrund der Diskussion um hybride Systeme (Diamond 2002) oder defekte Demokratien (Merkel et al. 2003) einer zunehmenden Kritik ausgesetzt.

Tabelle 1: Demokratiewerte für entwickelte Demokratien in etablierten Indizes

Land	DB 2007 (0-100)	FH PR (1-7)	FH CL (1-7)	PolityIV 2010 (0-10)	Vanhanen 2000 (0-100)	SGI 2011 (1-10)	Democracy Ranking 2011 (0-100)	Democracy Index 2011 (1-10)	UDCS 2010 (-1,5-2,0)
Dänemark	87,2	1	1	10	41,19	9,05	83,5	9,52	2,01
Finnland	86,9	1	1	10	35,60	9,37	85,8	9,06	2,00
Schweden	86,3	1	1	10	37,69	9,38	87,2	9,5	2,01
Island	83,3	1	1	-	36,13	8,41	-	9,65	1,99
Norwegen	83,3	1	1	10	37,93	9,43	88,2	9,8	2,01
Kanada	83,0	1	1	10	24,16	8,52	79,8	9,08	1,40
Belgien	81,2	1	1	8	42,72	7,78	79,7	8,05	1,24
Deutschland	77,6	1	1	10	35,53	8,76	80,9	8,34	2,00
Schweiz	77,1	1	1	10	18,96	8,66	85,2	9,09	2,02
Neuseeland	76,3	1	1	10	34,91	9,22	82,9	9,26	1,57
Niederlande	75,2	1	1	10	38,42	8,50	82,6	8,99	2,01
Slowenien	70,4	1	1	10	29,04	-	75,2	7,76	1,40
Luxemburg	67,2	1	1	-	29,10	7,97	-	8,88	1,31
USA	66,8	1	1	10	19,08	8,60	79	8,11	1,57
Australien	64,9	1	1	10	35,34	8,46	79,5	9,22	1,40
Österreich	64,9	1	1	10	37,94	7,40	79,9	8,49	2,01
Spanien	64,5	1	1	10	31,89	7,24	77,6	8,02	1,57
Ungarn	62,8	1	1	10	25,42	6,39	68,4	7,04	1,40
Portugal	61,7	1	1	10	28,06	7,54	75,7	7,81	1,41
Irland	61,4	1	1	10	30,13	8,64	81	8,56	1,40
Tsch, Rep	56,6	1	1	8	39,26	7,42	71,1	8,19	1,11
Italien	51,3	1	2	10	42,75	6,26	71	7,74	1,28
Japan	46,5	1	2	10	24,38	6,93	74,1	8,08	1,17
Polen	45,3	1	1	10	22,3	7,31	70,5	7,12	1,29
UK	43,6	1	1	10	30,15	7,66	79,9	8,16	1,40
Frankreich	41,5	1	1	9	29,26	7,32	76,2	7,77	1,12
Arith. Mittel	67,95	1	1,08	9,79	32,21	8,09	78,95	8,51	1,58
SD	14,22	0	0,27	0,59	6,97	0,92	5,38	0,77	0,34
Spannweite	45,7	0	1	2	23,79	3,17	19,80	2,76	0,91
Var. Koef. CV	20,9	0	25,2	6,93	22,0	11,36	6,93	10,1	34,8

Anmerkungen und Quellen:
Es wurden nur die Daten verwendet, für die im Vergleich hinreichend Datenpunkte vorlagen (n = 26)
Demokratiebarometer: http://www.democracybarometer.org/
Freedom House (FH; PR = Political Rights, CL = Civil Liberties): http://www.freedomhouse.org/report/freedom-world/freedom-world-2011;
Polity IV: http://www.systemicpeace.org/polity/polity06.htm;
Vanhanen: http://www.prio.no/CSCW/Datasets/Governance/Vanhanens-index-of-democracy/;
SGI Sustainable Governance Indicators (Demokratiewert): http://www.sgi-network.org/index.php?page=category&category=SA;
Democracy Ranking: http://www.democracyranking.org/en/;
Economist Democracy-Index: http://www.eiu.com/public/topical_report.aspx?campaignid=DemocracyIndex2011;
UCDS = Unified Democracy Scores (ein über einen Bayesianischen latenten Variablen Ansatz aus 10 bestehenden Demokratiemaßen synthetisierter kombinierter Demokratiewert (vgl. Pemstein et al. 2010)): http://www.unified-democracy-scores.org/

Diese vielversprechende Zielsetzung ist zu begrüßen. Gleichzeitig muss einem solchen Vorhaben im Hinblick auf die zahlreichen theoretischen und praktischen Schwierigkeiten die bis dato Messversuche unterminierten, mit gesunder Skepsis begegnet werden. Der vorliegende Beitrag möchte vor diesem Hintergrund analysieren, welche Strategie das Demokratiebarometer zur Überwindung des Problems der mangelnden Varianz vorschlägt, sowie dessen theoretische Grundlagen und die eigentliche Messung einer kritischen Begutachtung unterziehen. Die Form der externen Evaluation – also letztlich

des Vergleichs der Ergebnisse unterschiedlicher Demokratie-Messansätze unter der Grundannahme, dass hohe Korrelationen hier gleichsam als Indiz dafür gewertet werden können, dass eine gemeinsame latente Variable gemessen wird (= nomologische Validität) – hat in der deutschen Demokratieforschung insbesondere aus vergleichender Perspektive eine gewisse Tradition (Berg-Schlosser 1999; Lauth 2004; Lauth et al. 2000; Pickel und Pickel 2006).[2] Statt die Leistungsfähigkeit des Demokratiebarometers jedoch ausschließlich im Vergleich zu den Ergebnissen herkömmlicher Messansätze zu evaluieren, schlagen wir eine Überprüfung des Konzepts anhand der für die Demokratiemessung entwickelten und vielbeachteten Kriterien von Gerardo Munck und Jay Verkuilen vor (2002). Anhand dieser Kriterien lässt sich ein Raster aus Konzeptionalisierung, Operationalisierung und Messung sowie Aggregation entwickeln, das es ermöglicht, die Qualität des Demokratiebarometers im Hinblick auf die Indexbildung entsprechend systematisch einzuschätzen.

Bevor wir uns dem Aufbau und der Zielsetzung des Demokratiebarometers en détail zuwenden, soll zunächst knapp skizziert werden, welche Aspekte und Kriterien für die Evaluation von Messansätzen laut Munck und Verkuilen (2002) (im folgenden M&V) zentral sind. Im Anschluss an diese für unseren Artikel grundlegenden theoretischen Vorüberlegungen, wird das Demokratiebarometer auf den drei Ebenen Konzeption, Messung und Aggregation diskutiert und bewertet. Im letzten Abschnitt werden die gewonnenen Erkenntnisse synthetisiert sowie potentielle Verbesserung im Hinblick auf die zugrunde liegende Messung diskutiert.

1. Theoretische Vorüberlegungen: das Konzept von Munck & Verkuilen

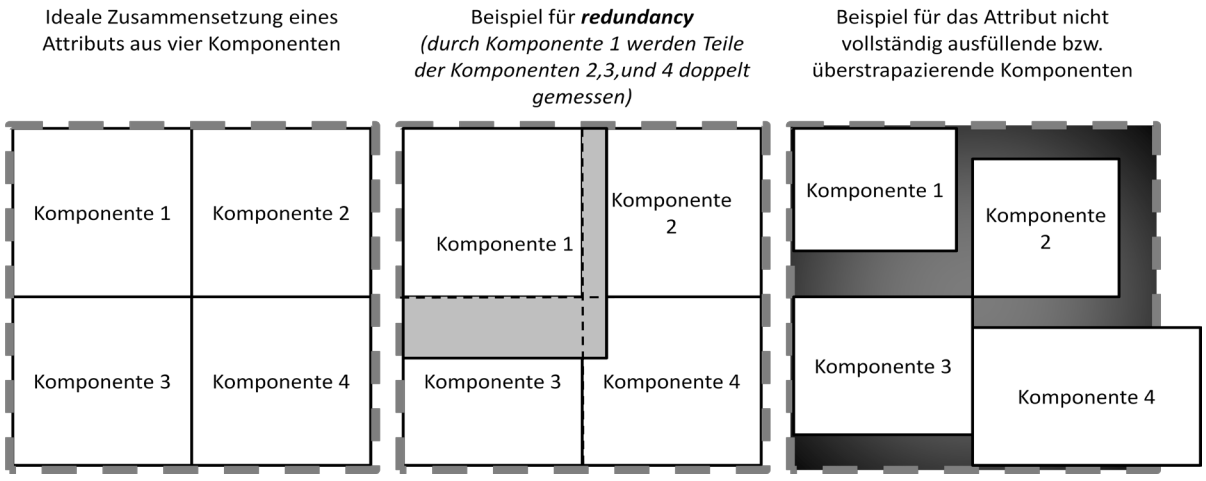
Konzeptionell basieren wir unsere Kritik am Demokratiebarometer auf dem von M&V (2002) erarbeiteten Vorschlag zur Bewertung von Messansätzen. Hierfür sprechen insbesondere zwei Gründe. Erstens, hat sich diese Herangehensweise in der Debatte um die Güte von Messvorschlägen mittlerweile als zentrales Analyseraster etabliert (Jäckle und Bauschke 2009; 2010; Müller und Pickel 2007). Zweitens, beziehen sich die Autoren des Demokratiebarometers selbst – wenn auch wie zu zeigen sein wird selektiv und damit in der Konsequenz unzureichend – auf die von M&V als relevant erachteten „Konstruktionsempfehlungen“ (vgl. Bühlmann et al. 2011a: 2). M&V folgend lassen sich drei wesentliche Aspekte bzw. Ebenen der Messung unterscheiden: Die *Konzeptualisierung*, die *Operationalisierung* mit der eigentlichen *Messung* im Anschluss, sowie die *Aggregation*.

Auf der Ebene der *Konzeptionalisierung* muss zunächst das zu messende Konstrukt präzisiert werden. M&V (vgl. 10-14) identifizieren in diesem Zusammenhang mehrere Anforderungen. Zunächst gilt es ein Konzept zu entwickeln, das weder zu maximalistisch noch zu minimalistisch ausgestaltet ist. Das Problem einer maximalistischen Definition ist, dass sie das grundlegende Konzept überfrachtet, so

[2] Der Demokratiebarometer ist am stärksten mit dem Democracy-Ranking (r =0,72) sowie dem Democracy-Index des Economist (r = 0,70) korreliert, negativ ist die Korrelation mit dem Civil Liberty Index von Freedom House (r = -0,39).

dass sich in der Empirie keine entsprechenden Fälle mehr identifizieren lassen. Im Gegensatz dazu wird bei einer minimalistischen eine zu sparsame und damit unvollständige Definition gewählt, die keinerlei ausreichende Differenzierung der zu untersuchenden Fälle mehr ermöglicht. In einem zweiten Schritt gilt es die verschiedenen Bestandteile des Konzepts ihrem Abstraktionsgrad entsprechend logisch anzuordnen. Hier führen M&V das intuitiv eingängige Bild des Konzeptbaums (*concept tree*) ein. Das abstrakte Konzept der Demokratie – versinnbildlicht als Stamm eines Baumes – lässt sich demnach in einzelne Äste aufteilen, die wiederum in Zweige ausdifferenziert werden können. Die Äste repräsentieren hierbei die das Gesamtkonzept konstituierenden Dimensionen bzw. Komponenten, während die Zweige als deren Subkomponenten aufgefasst werden können. Zum Teil können mehrere Ebenen einer solchen Verzweigung notwendig sein, um ein Konzept adäquat in all seinen Feinheiten erfassen zu können. An den Enden dieses Konzeptbaums finden sich schließlich die Blätter (*concept leaves*). Diese sind es, denen im Rahmen der Operationalisierung einzelne messbare Indikatoren zugeordnet werden. Die eigentliche Messung findet entsprechend auf der Ebene der Blätter statt. Die Aufgabe des Indexentwicklers ist es zum einen, die richtige horizontale und vertikale Differenzierung sicherzustellen, zum anderen die beiden Gefahren der Redundanz (*redundancy*) sowie Aufblähung (*conflation*) zu verhindern. Unter *redundancy* verstehen M&V, dass die auf einer Abstraktionsebene befindlichen Komponenten sich gegenseitig ausschließende Aspekte des genau über ihnen angesiedelten Attributs ausmachen sollten, es also komplett erfassen, sich dabei jedoch nicht überlappen sollten.

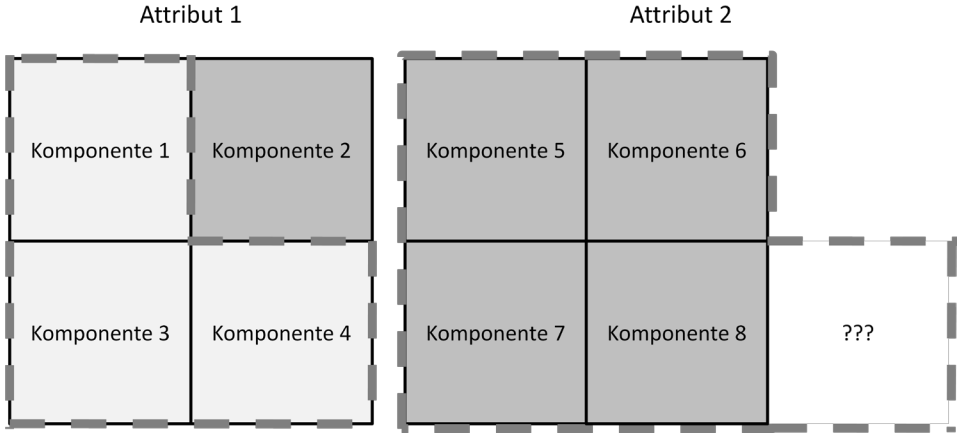
Abbildung 1: Beispiel für *redundancy* und für unvollständige sowie aus dem Attribut herausragende Komponenten



In Abbildung 1 wird dies nochmals dargestellt: Das große grau gestrichelte Quadrat repräsentiert das übergelagerte Attribut, welches aus vier Komponenten zusammengesetzt sein soll. In der linken Grafik gelingt das passgenaue Zusammensetzen der vier Komponenten zum Attribut perfekt. Die mittlere Grafik zeigt ein Beispiel für *redundancy*. Die rechte illustriert zwei ähnlich gelagerte, so von M&V jedoch nicht explizit beschriebene Problematiken, nämlich dass die verwendeten Komponenten nicht

groß genug sind um ein Attribut vollständig auszufüllen oder über die konzeptionellen Ränder des Attributs hinausragen und somit Aspekte beinhalten, die nicht mehr als konstitutiv für das Attribut zu betrachten sind. Solche Komponenten können ein Attribut „überstrapazieren“. Dazu kommt es auch bei der zweiten von M&V genannten Gefahr, der der *conflation* (vgl. Abbildung 2). Dabei fließt eine Komponente in ein Attribut mit ein, die eigentlich nicht zu diesem dazugehört, wodurch das Attribut künstlich aufgeblasen wird. Eventuell fehlt exakt diese Komponente dann an anderer Stelle, was für ein solches Attribut bedeuten würde, dass es unzureichend ausgefüllt wird. In Abbildung 2 ist exakt dies der Fall. Komponente 2 wurde unpassenderweise dem Attribut 1 untergeordnet, obgleich sie in ihrer Gänze besser unter Attribut 2 subsumiert würde. Allgemein kann man diese Gefahren in dem Maße umgehen, „in dem die Attribute und Komponenten erstens jeweils eine eigenständige Bedeutung haben und zweitens zueinander trennscharf sind“ (Fuchs und Roller 2008: 89).

Abbildung 2: Beispiel für *conflation*



Auf der Ebene der *Operationalisierung* und *Messung* lassen sich, neben der grundsätzlichen Anforderung einer theoretisch begründeten und nachvollziehbaren Indikatorenwahl, drei zentrale Anforderungen bzw. Gütekriterien an einen Messvorschlag formulieren (Munck und Verkuilen 2002: 15-22). Erstens sollten die Blätter des Konzeptbaums, möglichst durch mehrere Indikatoren abgebildet werden. Hierdurch können einerseits systematische Verzerrungen, die durch eine zu starke Fokussierung auf einzelnen Indikatoren verursacht würden, ausgeschlossen und andererseits der durch zufällige Messfehler entstehende bias reduziert werden. Ziel ist eine Minimierung der Messfehler durch geeignete Indikatoren, die auch anhand dritter Quellen gegengeprüft werden können. Gerade bei internationalen Vergleichen sollten transnational reliable Indikatoren verwendet werden. Dies setzt jedoch voraus, dass die gewählten Indikatoren unabhängig vom jeweiligen kulturellen, sozialen oder politischen Kontext, d.h. insbesondere in unterschiedlichen Ländern dasselbe messen. In Anlehnung an Przeworsky und Teune (1970) sprechen die Autoren von „cross-system equivalence“, die bei der Auswahl von Indikatoren zu gewährleisten sei (Munck und Verkuilen 2002: 16). Generell sollten die Indikatoren also eine hohe Zuverlässigkeit (Reliabilität) aufweisen. Zweitens, sollte die Auswahl des Messniveaus theoretisch begründet sein, der prinzipiellen externen Überprüfung offenstehen und

sicherstellen, dass weder zur grob noch zu differenziert gemessen wird. M&V schlagen eine Maximierung der Homogenität innerhalb der Messklassen mit einem Minimum an notwendigen Unterscheidungen vor. Drittens g die Replizierbarkeit der Messung sicherzustellen, also die prinzipelle Wiederholung der Messung muss gewährleistet sein.

Der Prozess der *Aggregation* führt die Messanlage wieder zurück zum Stamm des Konzeptbaums. Hierbei sieht man sich jedoch laut M&V mit einem generellen Trade-off konfrontiert (vgl. Munck und Verkuilen 2002: 22-27). Mit jeder Verdichtung der Rohdaten bzw. der einzelnen Indikatoren geht ein Informationsverlust einher den es theoretisch zu begründen gilt. So mag eine Aggregation in einen einzelnen Wert zwar im Sinne einer besseren Verständlichkeit und v.a. Vergleichbarkeit einleuchten, jedoch gleichzeitig ebenso die potentielle Mehrdimensionalität des zugrunde liegenden Konzepts ignorieren und die Differenzierungsmöglichkeiten der untersuchten Fälle reduzieren. Eine weitere Herausforderung stellt die Auswahl einer geeigneten Aggregationsregel dar, die wiederum theoriebasiert erfolgen sollte. In anderen Worten müssen sich die theoretisch postulierten Beziehungen zwischen den einzelnen Ebenen des Konzeptbaumes (Blätter – Zweige – Äste – Stamm) im mathematischen Ausdruck der angewandten Aggregationsregel widerspiegeln.[3] Eine weitere entscheidende Frage bei der Aggregation ist die Standardisierung der Variablen. Wenn einzelne Variablen nicht nur auf unterschiedlichen Skalen messen, sondern auch unterschiedliche Größenordnungen aufweisen, müssen sie zunächst vergleichbar gemacht werden, um eine sinnvolle Aggregation durchführen zu können. Dies kann beispielsweise über ein Aufspannen des empirisch vorgefundenen Zahlenraums auf eine feste Skala (z.B. 0 bis 100) oder über die z-Standardisierung geschehen, bei der alle Variablen so transformiert werden, dass sie einen Mittelwert von 0 und eine Standardabweichung von 1 erhalten.[4] Auch diese Weichenstellung beeinflusst das Ergebnis (s.u.). Schließlich ist die Frage der Gewichtung bei der Aggregation von Relevanz. Selbst wenn alle Variablen und Komponenten gleichgewichtig eingehen, ist dies eine inhaltliche Entscheidung. Dies gilt gerade auch für das Demokratiebarometer, da etwa die Zahl der Subkomponenten und Indikatoren unterschiedlich groß ist und so eine implizite – und nicht diskutierte – Gewichtung vorgenommen wird (vgl. Abbildung 3).

[3] Unseres Erachtens wird dieser Punkt häufig nur wenig bei der Indexbildung reflektiert. So liegt der Entscheidung für eine additive Aggregation (z.B. Mittelwertbildung) oder multiplikative Aggregation (z.B. geometrisches Mittel) eine bedeutsame demokratietheoretische Entscheidung zugrunde. Bei einer Verwendung des arithmetischen Mittels ist die (oft unausgesprochene) Weichenstellung, dass sich die einzelnen Komponenten des Konzepts gegeneinander ausgleichen können, d.h. Defizite etwa bei der Freiheit können durch höhere Werte bei (sozialer) Gerechtigkeit kompensiert werden. Beim geometrischen Mittel bedingen sich dagegen die einzelnen Komponenten stärker. Ist der Wert einer Teilkomponente (z.B. Freiheit) „null“ dann ist auch der Gesamtindex null. Je nach theoretischer Fundierung ist demzufolge entweder einer multiplikativen Verknüpfung über das geometrische Mittel oder einer additiven Verknüpfungslogik, wie sie dem arithmetischen Mittel zugrunde liegt, sachlogisch der Vorzug zu geben.

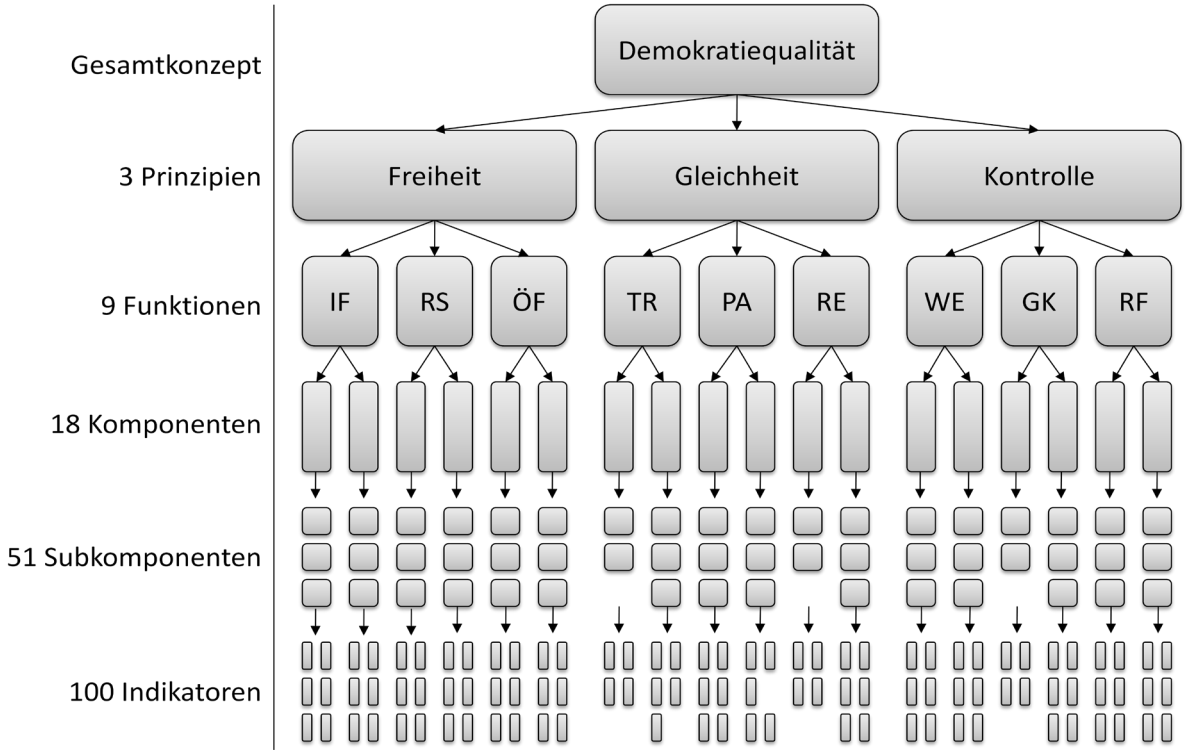
[4] Sie wird durch folgende Formel berechnet: $z = \frac{x_i - \bar{x}}{\sigma_x}$, mit Sigma = Standardabweichung der Variablen x.

Die von M&V entwickelten Kriterien bieten uns nicht nur erste Anhaltspunkte für die Begutachtung der Messvorschläge sondern ebenso eine Strukturierungsmöglichkeit der nun folgenden Analyse. Bevor jedoch die eigentliche Evaluation beginnen kann, soll jedoch zunächst das Demokratiebarometer knapp skizziert werden. Im Anschluss wird dessen zugrunde liegende Konzeptionalisierung kritisch diskutiert um in den beiden darauf folgenden Abschnitten die Operationalisierung und Messung selbst sowie die verwendete Aggregation zu evaluieren.

2. Das Demokratiebarometer – Zielsetzung und grundlegende Herangehensweise

Wie bereits weiter oben erläutert, ist das Hauptanliegen des Demokratiebarometers eine Binnendifferenzierung etablierter Demokratien zu ermöglichen. Als weiteres und hiermit verknüpftes Anliegen betrachten die Autoren die Identifizierung verschiedener (funktionaler) Realisierungsmöglichkeiten von Demokratie im Sinne einer „varities of democracies“ (Bühlmann et al. 2011a: 9). Theoretisch basiert der Messvorschlag auf einer Demokratiekonzeption, die sich aus den Dimensionen – bzw. im Jargon des Demokratiebarometers – den Prinzipien Freiheit, Gleichheit und Kontrolle konstituiert.[5] Diese drei Prinzipien lassen sich wiederum in jeweils 3 Funktionen aufgliedern (vgl. Abbildung 3).

Abbildung 3: Der Konzeptbaum des Demokratiebarometers

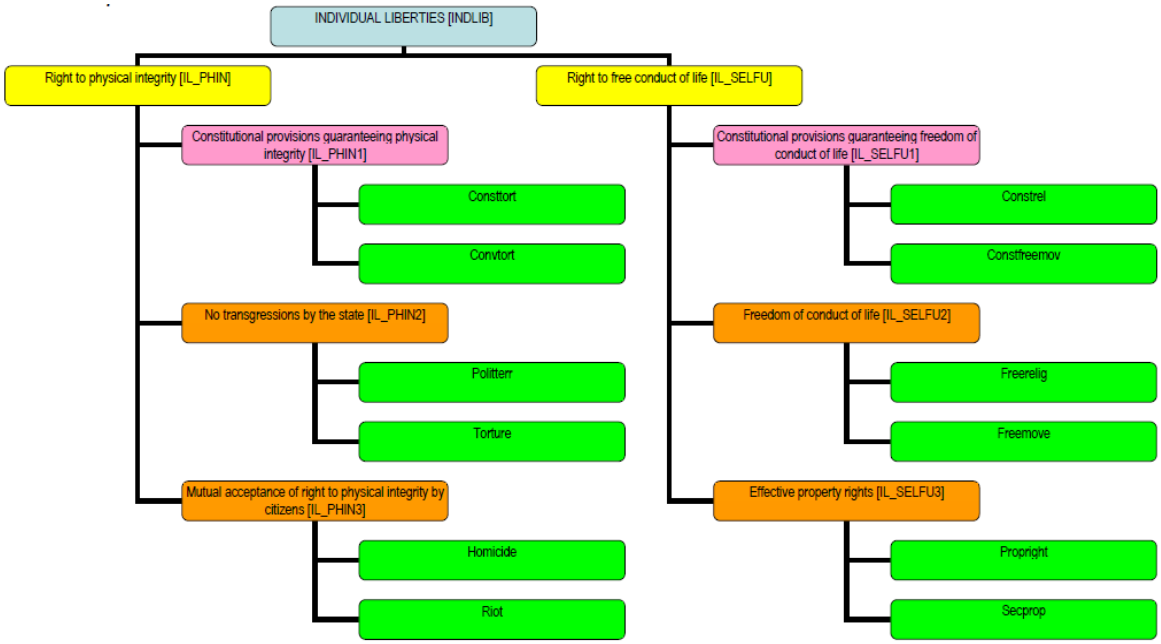


Eigene Darstellung nach (Bühlmann et al. 2011c: 18-42). IF: Individuelle Freiheiten; RS: Rechtsstaatlichkeit; ÖF: Öffentlichkeit; TR: Transparenz; PA: Partizipation; RE: Repräsentation; WE: Wettbewerb; GK: Gewaltenkontrolle; RF: Regierungsfähigkeit;

[5] Das Demokratiebarometer bezieht sich damit implizit auf das von Hans-Joachim Lauth (2004) systematisierte Demokratieverständnis.

Die Funktionen werden in einem weiteren Schritt in jeweils 2 Komponenten gegliedert, die über mehrere (zumeist drei) Subkomponenten ausdifferenziert werden, deren Operationalisierung zumeist über zwei Indikatoren erfolgt. Die Autoren folgen damit der von M&V eingeführten Metapher des Konzeptbaums. In der folgenden Abbildung 4 ist im Sinne eines besseren Verständnisses, die Ausdifferenzierung für die Funktion „individuelle Freiheit“ exemplarisch dargestellt.[6] Die Messung erfolgt letzten Endes über insgesamt 100 Indikatoren, die nach einer Reskalierung (sh. 3.1) in die Konstruktion des Demokratiebarometers einfließen. Bis dato liegen Daten für den Zeitraum 1995 bis 2007 für 30 etablierte Demokratien vor.

Abbildung 4: individuelle Freiheiten



Quelle:(Bühlmann et al. 2011c)

2.1. **Konzeptionelle Kritik**

Die Diskussion der konzeptionellen Grundlagen muss zwangsläufig an der verwendeten Demokratiedefinition ansetzen. Die Autoren schlagen hier eine „Demokratiekonzeption mittlerer Reichweite“(Bühlmann et al. 2011a: 2) vor. Man könnte auch von einem dynamischen und durch Interaktionen geprägten Demokratiemodell sprechen, da die Autoren Demokratie basierend auf der „Prämisse, dass demokratische Systeme eine Balance zwischen den interdependenten Werten ‘Freiheit’ und ‘Gleichheit’ herzustellen versuchen und [sie] sich dazu einer dritten demokratieinhärenten Dimension bedienen: Kontrolle“ (Bühlmann et al. 2011a: 4) nicht als starres Konzept denken.

[6] Die Konzeptbäume für alle Funktionen bzw. Komponenten finden sich in (Bühlmann et al. 2011c).

3.1.1 Fundamentale Kritik an der grundlegenden Konzeption nach Victoria Kaina

Diese drei Prinzipien stehen dabei laut Bühlmann et al. in einem Spannungsverhältnis. Einerseits werden sie als normativ gleichwertig erachtet und andererseits – und das ist das entscheidende Novum zu alternativen Messansätzen – stehen sie gleichzeitig in einem interdependenten Verhältnis zueinander. Dies bedeutet jedoch auch, dass die Maximierung einer Dimension zu Lasten einer oder der anderen beiden Dimensionen geht. Entsprechend kann man auch nicht den Fall als optimal für die Demokratiequalität bezeichnen, wenn alle drei Dimensionen vollständig bzw. maximal ausgeprägt vorliegen. Eine solche Konstellation ist in dem vorliegenden Demokratiekonzept per definitionem ausgeschlossen. Die vom Demokratiebarometer vermutete, aber nicht explizit formulierte, grundlegende Vorstellung von Qualität ist insofern, wie es Victoria Kaina in ihrem Beitrag zum Demokratiebarometer pointiert herausgearbeitet hat, durchaus problematisch.[7] Das Demokratiebarometer postuliert, dass sich „der Maßstab für die Demokratiequalität eines demokratischen politischen Systems an einem Optimum orientier[t], das sich für das Spannungsverhältnis jener drei Demokratieprinzipen sensibel zeigt“ (Kaina 2008: 521). Damit liegt aber eben gerade kein gemeinsamer Maßstab vor, der an alle drei Prinzipien immer in derselben Weise angelegt werden kann und die Bildung einer Rangfolge erlauben würde, sondern lediglich eine Orientierung für die fallspezifische Bestimmung eines optimalen (Spannungs-)Verhältnisses zwischen den betrachteten Demokratien. Neben der praktischen Schwierigkeit ein allgemeingültiges Optimum zu bestimmen, das einen Vergleich zumindest potentiell ermöglichen würde, verweist Kaina auf das Problem der mangelnden Kontextualität, da „sich das jeweilige Optimum in der Verwirklichung der drei Demokratieprinzipien in einem gegebenen demokratischen politischen System nicht kontextfrei bestimmen lässt“ (Kaina 2008: 522). In Ermangelung eines theoretischen Maßstabes, läuft eine empirische Bestimmung des Optimums laut Kaina jedoch Gefahr arbiträr zu werden.[8] Im Ergebnis bedeutet dies, dass das Demokratiebarometer dem eigenen Anspruch eine Rangfolge der Staaten entsprechend ihrer Demokratiequalität zu erstellen, nur teilweise gerecht werden kann. Stattdessen erscheint der Messvorschlag als ein Instrument zur Klassifikation verschiedener Typen der Demokratie oder wie Bühlmann et al. formulieren zur Identifikation der „Varieties of Democracies“.

[7] Die Autorin greift mit der Frage nach der Bedeutung von Qualität eine hochrelevante, jedoch weitestgehend ignorierte Frage im Rahmen der aktuellen Demokratieforschungsdebatte auf (Ausnahmen bilden (Abromeit 2004; Diamond und Morlino 2005; Fuchs 2004; Fuchs und Roller 2008).

[8] Ein drittes Problem sieht Kaina in der Verwendung eines „optimalen Erfüllungsgrades“ als Maßstab höchster Demokratiequalität. „Das Problem an diesem Argument ist, dass sich in der Sprachlogik der Worte „Optimum“ oder „optimal“ graduelle Abstufungen verbieten. Es geht hier also um Unterschiede der Art (und nicht des Grades), weil ein Optimum entweder erreicht ist oder nicht, etwas entweder optimal ist oder nicht.[…] Es können auf diesem Wege demzufolge nur dichotome Aussagen über die Demokratiegüte demokratischer politischer Systeme getroffen werden“ (Kaina 2008: 522). Rein semantisch betrachtet hat Kaina hier durchaus recht, allerdings erscheint uns ihre Kritik doch überzogen und letztlich am Punkt vorbei. Die Frage ist nicht ob es eine optimalere Demokratiequalität als die optimale gibt, sondern wie man zu welchem Grad dem Optimum nahekommen kann.

Je nachdem, für wie kritisch man die angesprochenen und zumindest von Kaina als durchaus fundamental erachteten theoretisch-konzeptionellen Defizite des Demokratiebarometers hält, könnte man bereits an dieser Stelle jede weitere Evaluation abbrechen und den Daumen über dieses Messinstrument senken. Im Folgenden werden wir jedoch nicht den Weg einer solchen Fundamentalopposition einschlagen, sondern wir werden zeigen, dass abseits dieser konzeptionellen Ungereimtheiten auch noch weitere Kritikpunkte am Demokratiebarometer anzubringen sind, die zum Teil allerdings durchaus ihre Grundlegung in eben dieser problematischen Konzeption von demokratischer Qualität erfahren haben.

3.1.2 Zu maximalistisch? – die konzeptionelle Breite des Demokratiebarometers

Projiziert man das gewählte (funktionalistische) Demokratiemodell auf das von M&V erwähnte Kontinuum maximalistischer und minimalistischer Definitionen, lässt sich das Demokratiebarometer als eher maximalistisches Konzept auffassen, was ebenfalls in der Literatur bereits kritisiert wurde (Lauth 2010). Diese Entscheidung wird von den Autoren folgendermaßen begründet: Das Problem existierender Messansätze sei, „dass diese auf einer (zu) minimalistischen Demokratiekonzeption basieren." Und weiter: „ Um das Phänomen 'Demokratie' jedoch in seiner ganzen Komplexität erfassen zu können, muss eine Demokratiequalitätsmessung auf einem breiten Demokratiekonzept fußen." (Bühlmann et al. 2011a: 2). Letzterer Aussage ist grundsätzlich zunächst zuzustimmen. Tatsächlich scheint eine Spezifizierung bzw. Erweiterung des Demokratieverständnisses die einzig sinnvolle Option darzustellen um eine adäquate Differenzierung unter den generell demokratischen Staaten abzubilden. Dies muss jedoch – wie auch M&V hervorheben – in einer Weise geschehen, die verhindert, dass Elemente aufgenommen werden, die für die Qualität einer Demokratie nicht konstitutiv sind (siehe hierzu auch: Fuchs 2004; Lauth 2004). Überlegungen beispielsweise ökonomisches Wohlergehen als Charakteristikum ausgeprägter Demokratien anzusehen wären aus zweierlei Gründen fatal. Erstens würde hierdurch die Gefahr bestehen, dass die eigentlichen Kernelemente, auf die es bei einer Messung demokratischer Qualität ankommen sollte, verwässert würden. Zweitens böte sich hierdurch nicht mehr die Möglichkeit, den Zusammenhang zwischen Demokratie und Ökonomie selbst in den Fokus der Analyse zu nehmen. Dies ist nur möglich, wenn Demokratie nicht vollkommen maximalistisch gedacht wird.[9] Den Autoren des Demokratiebarometers kann jedoch bescheinigt werden, dass sie die von ihnen verwendete, vergleichsweise breite Demokratiedefinition über die drei Prinzipien Freiheit, Gleichheit und Kontrolle überzeugend begründen und damit zumindest auf der obersten Ebene des Konzeptbaumes kein überdehntes

[9] Dieselbe Problematik zeigt sich auch im umgekehrten Fall. Bei den Sustainable Governance Indicators (SGI) lässt sich beispielsweise argumentieren, dass der Einbezug der Demokratiekomponente in den Gesamtindex im Sinne eines konstituierenden Elements von Reformfähigkeit dessen eigentliches Kernkonzept zu maximalistisch abbildet (Jäckle und Bauschke 2010: 80).

Demokratiekonzept aufstellen.[10] Das bedeutet jedoch nicht, dass *conceptual overstretching* für das gesamte Demokratiebarometer kein Problem darstellen würde. Vielmehr wird im Folgenden gezeigt werden, dass durch die weitere Ausdifferenzierung über die insgesamt fünf Ebenen, die unterhalb der Demokratiequalität anzusiedeln sind, das ursprüngliche Konzept Schritt für Schritt so ausgeweitet wird, dass letztlich ein deutlich breiteres Konzept auf der untersten Ebene gemessen wird als eigentlich intendiert.

3.1.3 Die Indikatorenwahl – Unzureichende Herleitung und schrittweise Ausweitung des Messkonzepts

Eine weitere Anforderung in der Phase der Konzeptionalisierung ist die Schrittweise Ausdifferenzierung des Demokratiemodells im Sinne des Konzeptbaums.[11] Der Schritt der weiteren Spezifizierung der drei Prinzipien in jeweils drei Funktionen wird von den Autoren zwar relativ überzeugend geleistet aber bereits Lauth kritisiert die nicht folgerichtige Zuordnung der Funktionen zu den Prinzipien, da sie „weder ausreichend begründet noch überzeugend sind." (2010: 518). So kann die Funktion „Transparenz" ohne weiteres auch dem Prinzip „Kontrolle" subsumiert werden und nicht nur dem der „Gleichheit". Gleiches gilt, wenn man sich die Einsortierung des „Wettbewerbs" unter das Prinzip „Kontrolle" ansieht, der doch mindestens genauso plausibel auch unter „Gleichheit" verortet werden könnte. Der Erfüllungsgrad der insgesamt neun Funktionen bildet den zentralen Maßstab der Demokratiequalität (vgl. Bühlmann et al. 2011a: 9). Unterhalb der Funktionenebene fehlt es dem Demokratiebarometer oftmals jedoch noch deutlicher an einer klaren, nachvollziehbaren und vor allem theoretisch begründeten Ausdifferenzierung. Die Verbindung der neun Funktionen mit den darunter liegenden und sie damit konstituierenden Komponenten bzw. Subkomponenten erfolgt wesentlich oberflächlicher als die ersten beiden Differenzierungsschritte vom „Demokratiestamm" über die drei Prinzipien zu den neun Funktionen. Doch gerade auch auf den unteren Ebenen ist eine differenziertere Darstellung angezeigt um einerseits eine gewisse Nachvollziehbarkeit zu gewährleisten und andererseits vor allem auch Anhaltspunkte für die Validität der Messung zu bieten. Schließlich ist es nochmals ein Unterschied ob Indikatoren, d.h. die Blätter des Baumes beschrieben, oder zumindest in tabellarischer Form gelistet werden, wie dies beim Demokratiebarometer in der Regel geschieht oder ob die jeweiligen Zusammenhänge zwischen Indikatoren, Subkomponenten, Komponenten bis zu dem eigentlich interessierenden Gesamtkonzept wirklich expliziert werden und somit die Verästelungen des Konzeptbaumes nachvollziehbar gemacht werden. Ohne hier der

[10] An dieser Stelle sei nochmals auf die Vorarbeiten von Lauth (2004; 2000) verwiesen, der eine noch umfangreichere Begründung des „dreidimensionalen" Demokratieverständnisses über Freiheit, Gleichheit und Kontrolle liefert als die Autoren des Demokratiebarometers.

[11] Interessanterweise scheinen die Autoren diesen auch bei M&V explizit erwähnten Schritt nicht als zentrale Aufgabe anzusehen:„Using Munck and Verkuilen (2002) as a starting point, there are three critical tasks to accomplish: (1) Appropriateindicators have to be collected. (2) The scaling of the indicators as well as the (sub)components, functions, andprinciples needs to be determined carefully, especially when it comes to the identification of the lowest and highestpossible values. (3) The aggregation of values from each level of the concept tree to the next higher levelshould be in line with the theoretical assumptions of the concept." (vgl. Bühlmann et al. 2011c: 3).

Diskussion der eigentlichen Messung vorgreifen zu wollen, lässt sich hinterfragen ob der eigene Anspruch einer theoriegeleiteten, schrittweisen Entwicklung der Indikatoren ausgehend von einer grundlegenden Demokratiedefinition über mehrere Zwischenebenen beim Demokratiebarometer tatsächlich ausreichend eingelöst worden ist.

Kontrastiert man das Vorgehen des Demokratiebarometers mit den von M&V formulierten Regeln, lassen sich auf der Konzeptualisierungsebene folgende konkreten Defizite konstatieren: Zunächst lässt sich aufgrund der unzureichenden theoretischen Herleitung die Geeignetheit mancher Indikatoren hinterfragen. So mag noch einleuchten, dass die öffentliche Unterstützung einer Regierung einen positiven Effekt auf die Regierungsfähigkeit hat,[12] doch inwiefern sich diese Unterstützung über die persönliche Einstellung zum Schwarzfahren erheben lässt (vgl. Bühlmann et al. 2011b: 28), selbst wenn dies nur einen Teilindikator darstellt, ist zumindest fraglich. Durch die mangelnde Diskussion der Indikatoren in Bezug auf ihre vertikale Verknüpfung und ihre letztliche Geeignetheit, muss der Leser die vom Demokratiebarometer angenommenen Zusammenhänge zwischen Indikatoren und den übergeordneten Ebenen entweder unreflektiert hinnehmen, oder versuchen sie sich selbst zu erschließen – die Ausführungen in den Publikationen zum Demokratiebarometer sind hierbei nur bedingt hilfreich.[13]

Neben der mangelnden Stringenz in der Herleitung und der großen Anzahl an Indikatoren – was im Übrigen auch eine umfassende Diskussion und Evaluation im Hinblick auf die Problematik der *conflation* und *redundancy* erschwert, ist es der von den Autoren beschriebene Prozess der Indikatorenauswahl, der Zweifel an einer ausschließlich theoriegeleiteten Entwicklung aufkommen lässt. Zunächst wurden 300 potentielle Indikatoren kompiliert. In einem weiteren Schritt wurden über eine negative Auswahl dann 200 Indikatoren ausgeschlossen, so dass 100 für die weitere Analyse verblieben. Diese Reduzierung der Indikatorenzahl erfolgte anhand folgender Vorgaben: Erstens sollten Indikatoren die auf Expertenbefragungen beruhen, aufgrund möglicher Reliabilitätsprobleme, größtenteils ausgeschlossen werden. Zweitens, wurde versucht durch die Verwendung verschiedener Quellen und multipler Indikatoren, systematische Verzerrungen möglichst auszuschließen sowie zufällige Fehler zu reduzieren. Drittens, wurden jeweils Indikatoren ausgewählt, die die de jure und die de facto Ebene abbilden können. Viertens wurden nur jene Indikatoren ausgewählt die für einen längeren Zeitraum und für alle im Sample enthaltenen Fälle verfügbar waren. Sowohl der Auswahlprozess der ursprünglichen Indikatoren als auch ihre Reduktion erfolgte dabei laut den Autoren „basically theory driven“ (Bühlmann et al. 2011b: 4). Trotz dieses Hinweises drängt sich

[12] Sehr viel grundlegender kann man sich natürlich die Frage stellen, ob nicht auch nicht-demokratische Staaten, wie z.B. Singapur oder die VAE eine sehr gute Regierungsfähigkeit aufweisen können und inwiefern ein solcher Indikator, der sowohl in demokratischen als auch in autoritär geführten Systemen hoch ausgeprägt sein kann als Maß für Demokratiequalität taugt.

[13] So ist beispielsweise auch der angenommene positive Zusammenhang zwischen der Dauer der Legislaturperiode und der Regierungsfähigkeit (vgl. Bühlmann et al. 2011b: 28) durchaus diskussionswürdig, nicht nur wenn man wiederum den Vergleich zu autokratischen Systemen anstrengt, in denen die „Legislaturperioden“ entweder unbegrenzt sind, oder keinerlei Auswirkungen auf die Politik zeitigen.

aufgrund der verfügbaren Informationen der Verdacht auf, dass es sich bei der Konzeption der Messung und der Identifikation von Indikatoren um teilweise voneinander losgelöste Prozesse handelt. Statt einer stringenten theoriegeleiteten Entwicklung über die verschiedenen Ebenen des Konzeptbaums, die in einer positiven Auswahl der letztlich gemessenen Indikatoren münden sollte, wird aus einer Vielzahl verfügbarer Indikatoren ein Indikatorenset zusammengestellt, dass dann mit den separat entwickelten Funktionen „gematcht" wird. Der Konzeptbaum wächst damit einerseits vom Stamm der Demokratiedefinition ausgehend bis zu den neun Funktionen – dort bricht diese theoriegeleitete Vorgehensweise jedoch größtenteils ab – und andererseits wächst er ausgehend vom Universum potentieller Indikatoren (den Blättern) durch ein möglichst plausibles Zusammenfassen der Einzelindikatoren zu Subkomponenten und Komponenten gewissermaßen rückwärts in die andere Richtung. Die Gefahr dabei ist selbstverständlich, dass Äste und Zweige des Baums nicht so recht zusammenpassen, dass es also zu einem gewissen „mismatch" zwischen den Ebenen kommt.[14]

Grundsätzlich ist auch der Wunsch zu begrüßen, die einzelnen Komponenten und Subkomponenten jeweils durch mehrere Indikatoren abzubilden und so systematische wie zufällige Messfehler zu reduzieren, gleichzeitig muss man sich aber aufgrund der sich hierdurch ergebenden sehr großen Anzahl von insgesamt 100 Indikatoren die Frage stellen, ob wirklich eine ausreichende Überprüfung auf *redundancy* und *conflation* stattgefunden hat. Eine intensivere Diskussion der einzelnen Komponenten, Subkomponenten und Indikatoren wäre genau aus diesem Grund wünschenswert, denn hier tun sich beim Demokratiebarometer doch einige Lücken auf. Die Zusammensetzung der neun Funktionen aus Komponenten und Subkomponenten wird zwar noch beschrieben (Bühlmann et al 2011a: 9-16), über eine Nennung der einzelnen Subkomponenten reicht diese „Diskussion" jedoch nur in den seltensten Fällen hinaus. Eine wirkliche Besprechung der untersten Ebene, also der verwendeten Indikatoren fehlt vollends.

Betrachtet man diese jedoch genauer, so tauchen durchaus potentielle Probleme auf: allgemein scheinen die vielen Ebenen zu einer unbewussten Ausweitung des grundlegenden Demokratiekonzepts zu führen, so dass in die Messung vermehrt Aspekte eingehen, die mit dem eigentlichen Kern von Demokratie und ihrer Qualität wenn überhaupt nurmehr peripher zusammenhängen. Auch wenn die Autoren mit dem Verweis auf ihr „grundsätzlich theoriegeleitetes Vorgehen" behaupten das von Sartori problematisierte „concept overstretching" (Bühlmann et al. 2011a: 9-16) zu umgehen, scheint exakt dies der Fall zu sein. So erscheint die Verwendung der Mordrate und der Anzahl gewalttätiger Proteste zwar für die Gesamtkomponente „Recht auf körperliche Unversehrtheit" durchaus notwendig, aber ob diese Komponente in ihrer Gänze wirklich einen notwendigen Teilaspekt der Demokratiequalität darstellt, kann wohl bezweifelt werden. An diesem Beispiel kann man zwei

[14] In eine ähnliche Richtung weist auch der mehrfach von den Autoren des Demokratiebarometers gegebene Hinweis (vgl. z.B. Bühlmann et al. 2011a: 20), dass sich die Messanlage nach Belieben durch Hinzufügen weiterer Indikatoren erweitern ließe. Ein in sich schlüssiges, kohärentes Bild gibt das Demokratiebarometer damit nicht ab.

grundlegende Fragen verdeutlichen, die sich im Rahmen des Demokratiebarometers an unterschiedlichen Stellen immer wieder stellen: Erstens, kann es sein, dass dieselbe Ausprägung eines Indikators sowohl als Indiz für eine hohe Demokratiequalität als auch als Indiz für ein autoritäres Regime erachtet werden kann? Denn sowohl in autoritären Regimen, die Proteste jedweder Art von Grund auf nicht zulassen als auch in Demokratien, in denen diese Proteste in der Regel friedlich ablaufen, würde der Indikator „Anzahl gewaltsamer Proteste" dieselbe Ausprägung annehmen. Ist er damit wirklich geeignet als Teilmaß für Demokratiequalität, kann er also um mit Przeworsky und Teune zu sprechen als Maß gesehen werden „that is reliable across systems and valid within systems" (Przeworsky und Teune 1970: 114)? Zweitens, selbst wenn man sich dazu entschließt Indikatoren, die eigentlich nicht zwischen Autokratien und Demokratien differenzieren als Maß für demokratische Qualität zu verwenden, stellt sich die Frage, ob man mit der Mordrate nicht etwas misst, was vielmehr ein Outcome guter Regierungsführung ist als ein konstituierendes Element demokratischer Qualität; und Bühlmann et al. geben ja selbst an, dass sie dezidiert keine Outputs und Outcomes messen wollen: „Outputs [werden] als genuin politische Entscheidungen betrachtet, die durch Demokratie als Mittel hervorgebracht werden sollten, nicht aber Bestandteil von Demokratie sind."

Ein weiteres Beispiel für ein Überdehnen des eigentlichen Konzepts findet sich bei der Komponente Öffentlichkeit. Eine funktionierende Öffentlichkeit sei über die beiden Komponenten der Vereinigungs- und Meinungsfreiheit realisierbar. Dem ist soweit durchaus zuzustimmen, wenn nun jedoch als Indikator für die real wahrgenommene Vereinigungsfreiheit die Anzahl gewerkschaftlich Organisierter sowie die Anzahl an Mitgliedern von Menschenrechts- und Tierschutzorganisationen herangezogen wird, so handelt es sich hierbei um Indikatoren, bei denen der Konnex zur Demokratiequalität letztlich fehlt. Der gewerkschaftliche Organisationsgrad hängt aufs stärkste mit der historischen Entwicklung des Wirtschafts- und Sozialsystems zusammen, die eben nur zum Teil durch die Politik und damit Demokratie geprägt war. Aus diesem Grund kann dieser Indikator in unseren Augen auch nicht als Maß für die Qualität einer Demokratie dienen. Die in einzelnen Ländern fast schon obligatorische Gewerkschaftsmitgliedschaft (in Island listet die auch vom Demokratiebarometer als Quelle verwendete ILO zum Teil mehr als 90%, ähnlich hohe Werte finden wir auch in den übrigen skandinavischen Ländern) deutet vielmehr auf einen gewissen sozialen Druck hin zumindest offiziell Gewerkschaftsmitglied zu sein, als darauf, dass diese Personen sich wirklich, gewissermaßen basisdemokratisch, in der Gewerkschaftsumgebung ausleben. Die relativ betrachtet wenigen Gewerkschaftsmitglieder in anderen Staaten sind dagegen eventuell deutlich politisch-aktiver und tragen damit zur Qualität der Demokratie sogar eher mehr bei als die Masse der skandinavischen Gewerkschaftsmitglieder in ihren Ländern. Zudem finden sich auch ähnlich hohe Werte in postkommunistischen Staaten direkt nach dem Fall des Eisernen Vorhangs. In der tschechischen Republik waren 1990 noch 80,5 % der Arbeitnehmer gewerkschaftlich organisiert. Dieser Wert nahm im Laufe der demokratischen Konsolidierung kontinuierlich ab auf 20,5% im Jahr 2007. Das Demokratiebarometer interpretiert dies als einen klaren Verlust demokratischer Qualität. Wie sinnvoll

der Indikator Gewerkschaftsdichte damit ist möge jeder selbst beurteilen. Gleichzeitig werden die für die demokratische Willensbildung wohl relevantesten Vereinigungen ausgeblendet. Wenn schon ein hoher Prozentsatz an in Tierschutzorganisationen engagierten Bürgern die Demokratiequalität erhöht, dann muss dieselbe Argumentation erst recht für die Mitgliederzahlen von Parteien gelten. Parteimitgliedschaft wird jedoch vom Demokratiebarometer nicht erfasst. Unserer Meinung nach böte zudem ein Rückgriff auf Umfragedaten, die politische Partizipation als demokratisches Gut selbst in den Fokus nehmen (z.B. Eurobarometer EB 60.1) eventuell eine Alternative bzw. Bereicherung der reinen mitgliederzahlenbasierten Indikatoren, die augenblicklich Verwendung finden.

Betrachtet man die Konzeption des Demokratiebarometers abschließend, ergibt sich ein gemischtes Bild: Problematisch erscheint die mangelnde Detailliertheit in Bezug auf die Entwicklung des Konzeptbaums und die daraus resultierenden Unsicherheiten bei der Zuordnung von Indikatoren. Positiv ist hingegen zu bewerten, dass sich die Autoren im Hinblick auf die Auswahl von Indikatoren der de jure/de facto Problematik bewusst sind. Grundsätzlich wird der Versuch unternommen, Komponenten sowohl über institutionelle Variablen als auch Variablen, die die Verfassungswirklichkeit erfassen, abzubilden. Dies stellt eine der zentralen Weiterentwicklungen gegenüber bisherigen Messansätzen dar, die sich primär auf institutionelle Aspekte der Demokratie kaprizieren. Allerdings läuft man hier leicht Gefahr Aspekte in die Messung mit aufzunehmen, die kein Element der Demokratiequalität selbst sind, sondern die ein Outcome guter demokratischer Qualität sind.

2.2. Messung: Wiegt die Flut an reliablen, quantitativen Indikatoren deren mangelnde Validität auf?

Bei der Messung beschränken sich die Autoren des Demokratiebarometers auf „‘objektive‘ Daten aus statistischen Sekundärquellen oder repräsentativen Bevölkerungsumfragen“ (Bühlmann et al. 2011a: 17; Hervorhebung im Original), Daten aus Expertenbefragungen oder Expertenbeurteilungen[15] sollen aufgrund fraglicher Reliabilität und einer intransparenten Kategorienzuordnung hingegen gezielt keine Verwendung finden. Der diesem Vorgehen zugrundeliegende Wunsch nach nachvollziehbaren und dabei für die Messung wirklich geeigneten Indikatoren ist vollends verständlich, es bleibt allerdings die Frage, ob ein alleiniges Zurückgreifen auf *hard data* wirklich ausreicht um die Komplexitäten erfassen zu können, die das Konzept demokratischer Qualität auszeichnen. Es stellt sich in unseren

[15] Expertenbefragungen sind zumeist offener, wohingegen die Experten bei Expertenbeurteilungen oftmals ihre Einschätzung standardisiert beispielsweise auf einer 10-Punkte-Skala abgeben (vgl. Jahn 2006: 198). Die beiden Ansätze können jedoch auch verbunden werden, wie z.B. bei den Sustainable Governance Indicators, bei denen die Länderexperten zunächst um ein Rating eines bestimmten Sachverhalts gebeten werden und diese Beurteilungen dann in einem zweiten Schritt ausführlicher in ihren eigenen Worten begründen sollen. Die Ausführungen der Länderexperten werden schließlich nicht nur zur besseren Nachvollziehbarkeit der Ratings verwendet, sondern im Sinne einer *thick description* zu Länderreports synthetisiert, so dass die Expertenbefragung einen doppelten Nutzen bietet: erstens trägt sie dazu bei ein theoretisch und konzeptionell schwierig zu fassendes Phänomen (Reformfähigkeit) passgenauer abzubilden und durch einen Index sinnvoll vergleichbar zu machen, zweitens können interessierte Leser durch die Länderreports besser nachvollziehen, wie es zu einer bestimmten Bewertung eines Landes gekommen ist (vgl. Brusis 2009a: 83 ff).

Augen also die Frage, ob die verwendeten harten Fakten wirklich so gut geeignet sind. Denn die reliabelsten Indikatoren helfen nicht viel, wenn sie am eigentlich interessierenden Konzept „vorbeimessen". Der vorherige Abschnitt hat bereits einige Indikatoren aufgezeigt, bei denen eine entsprechende Gefahr zumindest nicht auszuschließen ist. Insofern liegt beim Demokratiebarometer eventuell eine Überbetonung des Reliabilitätskriteriums vor, bei gleichzeitiger Vernachlässigung des eigentlich noch wichtigeren Validitätskriteriums. Aus forschungspragmatischer Sicht verwundert dies nicht, sind doch die so beworbenen reliablen *hard data*-Indikatoren auch gerade diejenigen Maße, die sich gut und günstig über die Zeit wie über verschiedene Länder erheben lassen. Gerade bei der Messung latenter Variablen, die sich bereits auf konzeptioneller Ebene als vergleichsweise komplexe Phänomene präsentieren, wie dies bei Demokratie und deren Qualität zweifellos der Fall ist, und bei denen klassische Proxy-Maße nur eine unzureichende Passgenauigkeit aufweisen, kann jedoch auf das Wissen von Experten eventuell nicht verzichtet werden.[16] Entsprechend dieser Argumentation kombinieren beispielsweise die Sustainable Governance Indicators Expertenbefragungen und *hard data*. Ein Zitat aus der Methodik der SGI fasst den trade-off zwischen statistischen Daten und Expertenwissen und gleichzeitig die hieraus erwachsenden Kombinationsmöglichkeiten passend zusammen:

„Statistical data are generally more reliable than expert opinions, particularly when they are collected by official institutions and by using methods that conform to cross national standards. At the same time, however, such data often do not adequately cover the full meaning of a concept. We therefore believe that complex concepts can be measured best through the use of expert assessments that take the country specific context into account and provide "thick" descriptions capturing the nuances of phenomena"(Brusis 2009a)

Auch wenn an der Art und Weise Kritik geübt werden kann, wie die SGI die Expertenbefragungen im Prozess der Indexbildung verarbeiten (Jäckle und Bauschke 2009; 2010), so ist der grundlegende Gedanke komplexe Sachverhalte über Expertenwissen anzunähern und diese Informationen mit den statistischen Daten zu einem Gesamtindex zu integrieren nicht von der Hand zu weisen. Zudem ist die pauschale Kritik an den Expertenbefragungen in Bezug auf deren Reliabilität ungerechtfertigt. Wird eine ausreichende Anzahl an Experten befragt – was freilich oftmals Probleme im Bezug auf die Kosten sowie die Identifizierung potentieller Experten aufwirft – lässt sich die interne Konsistenz ihrer Antworten problemlos mit statistischen Mitteln überprüfen und damit eine ähnlich reliable Datengrundlage schaffen wie sie die statistischen Daten bieten.[17]

[16] Bei der Messung von Parteiideologie, einem der Demokratiemessung im Bezug auf die Komplexität durchaus vergleichbaren Konstrukt, haben sich Expertenbefragungen als sinnvolles Instrument herausgestellt, gerade weil sie es erlauben einige Schlüsselparameter, die sich über härtere Datenquellen nur mangelhaft bestimmen lassen gezielt abzufragen (vgl. Castles und Mair 1984; Laver und Hunt 1992: 34).

[17] Zudem, wenn bereits Expertenbefragungen aufgrund einer mangelnden Reliabilität vermieden werden sollen, stellt sich die Frage, weshalb auf Daten des IMD zurückgegriffen wird, die aus einer Umfrage unter den Absolventen dieser Schweizer Business School gewonnen werden (vgl. http://www.imd.org/research/centers/wcc/research_methodology.cfm).

3.3 Aggregation: Problematische Reskalierung und unnötig komplexe Aggregationsregel

Nach der Messung der einzelnen Indikatoren ist es nötig die Komplexität der gesammelten Informationen durch Aggregation zu reduzieren. Hierzu muss nach M&V zunächst entschieden werden ob es sinnvoll ist einen einzigen Gesamtindex zu bilden, oder ob hierdurch nicht zu viele Informationen über unterschiedliche Aspekte des Gesamtkonstrukts verloren gehen würden. Für das Untersuchungsobjekt Demokratiequalität ist es sicherlich sinnvoll einen Gesamtindex zu entwickeln, wie das Demokratiebarometer dies auch tut. Gleichzeitig wäre es jedoch auch gerade im Hinblick auf die verwendete Konzeption von Demokratie als dynamisches Gleichgewicht der drei Dimensionen Freiheit, Gleichheit und Kontrolle, das eben unterschiedliche Schwerpunktsetzungen als durchaus gleichdemokratisch versteht, eventuell sinnvoll die Aggregation auf dieser Ebene abzubrechen um die unterschiedlichen Demokratieprofile auf diese Weise besser vergleichend gegenüberstellen zu können. Durch Spinnengrafiken, die von den Autoren des Demokratiebarometers extensiv eingesetzt werden um die neun Demokratiefunktionen für einzelne Länder abzubilden, wird rein visuell diesem Anspruch jedoch durchaus Rechnung getragen. Insofern ist auch das höchstmögliche Aggregationsniveau nicht als Problem aufzufassen.

Problematisch dagegen sind zwei weitere Aspekte der Aggregation. Zunächst müssen sämtliche Indikatoren so standardisiert werden, dass sie an sich vergleichbar sind, und damit auch über eine geeignete Aggregationsregel kombinierbar werden. Sowohl die Standardisierung als auch die verwendete Aggregationsregel sind beim Demokratiebarometer jedoch suboptimal.

3.3.1 Künstlich erzeugte Varianz durch Standardisierung

Die Standardisierung erfolgt entsprechend empirischer maximal und Minimalwerte, was als „best-practice-Verfahren“ bezeichnet wird. Hierzu wird ein *blueprint-sample* aus den 30 etablierten Demokratien zwischen 1995 und 2005 betrachtet. Der höchste Wert, den jeder Indikator über alle 330 Länderjahre annimmt, wird auf den Wert 100 gesetzt, der niedrigste auf den Wert Null. Die ursprünglichen Indikatorwerte dazwischen werden entsprechend linearer Transformation auf der hierdurch entstehenden Skala zwischen 0 und 100 verteilt. Problematisch an diesem Verfahren sind aus unserer Sicht zwei Aspekte: erstens werden entgegen der ursprünglichen Konzeption, die davon ausgeht, dass Demokratie aus unterschiedlichen Kombinationen von Gleichheit, Freiheit und Kontrolle bestehen kann, doch wieder jeweils ein möglichst hoher Wert bei jedem Indikator als besonders gut für die Demokratiequalität betrachtet. Hierdurch wird implizit nichts anderes gemacht als die Problematik der theoretischen Bestimmung der Maxima und Minima (die von den Autoren des Demokratiebarometer sehr schön anhand des Beispiels Wahlrecht/Wahlbeteiligung dargestellt wird (Bühlmann et al. 2011a: 18-19) auf ein theoretisch unreflektiertes „der Wert sollte möglichst groß sein“ auszulagern. Am Wahlbeteiligungsbeispiel hieße das, dass eine Demokratie in der Wahlpflicht

herrscht, den Standard für gute Demokratie setzt. Damit ist es doch die partizipatorische Demokratietheorie die über andere Demokratieverständnisse dominiert.

Zweitens spreizt das gewählte Skalierungsverfahren die Daten künstlich auf. Wenn beispielsweise einzelne Indikatoren nur eine minimale Varianz in den Rohdaten aufweisen werden diese Unterschiede so aufgeblasen, dass sie dieselbe Spannweite aufweisen, wie Indikatoren, in denen sich die Länder wirklich deutlich unterscheiden. Diese aufgespreizte Varianz fließt dann in gleicher Weise in die Indexberechnung ein, wie die reale Varianz anderer Indikatoren. Dies ist einerseits ein konzeptionelles Problem da eigentlich irrelevante Faktoren plötzlich wichtig werden und dabei Länder, die nur marginal schlechter abschneiden über Gebühr abgestraft werden und andererseits ein praktisches Problem, da durch die Transformation die Transitivität des Index verletzt wird. Dies bedeutet, dass aus einer vollständig transitiven Ordnung der ursprünglichen Rohindikatorwerte keinerlei Rangfolge bei den transformierten Werten mehr herausgelesen werden kann. Tabelle 2 verdeutlicht diese Situation:

Tabelle 2: Beispiel zur Verdeutlichung der Skalierungsproblematik

	In Blueprint Sample		**Rohdaten**			**Transformierte Werte**		
	Min	Max	Land A	Land B	Land C	Land A	Land B	Land C
Indikator X	10	30	10	20	30	0	50	100
Indikator Y	20	80	80	50	20	100	50	0
Mittelwert	-	-	45	35	25	50	50	50

Lässt sich aus den ursprünglichen Indikatorwerten noch eine eindeutige Rangfolge mittels Durchschnittsbildung ablesen[18] (A>B>C), ist dies nach erfolgter Transformation der Werte nicht mehr möglich (A=B=C). Damit macht die Transformation der Werte die Situation tendenziell unklarer, was im Gegensatz zu der Hauptaufgabe eines Index allgemein steht, nämlich die Komplexität der Realität klarer geordnet darzustellen. Im ungünstigsten Fall könnte es sogar sein, dass sich auch die Rangfolge der Länder im Demokratiebarometer aufgrund der Skalierung verändert. Dadurch, dass für die Transformation stets auf dasselbe Blueprint-sample zurückgegriffen wird, umgeht das Demokratiebarometer ein Problem, das bei den Sustainable Governance Indicators deutlich zu Tage tritt. Die SGI verwenden dieselbe „best-practice"-Skalierung, allerdings werden dort für jede Welle die Minima und Maxima neu bestimmt, wodurch ein Vergleich der Indexwerte über die Zeit, d.h. über mehrere Erhebungswellen hinweg, nicht möglich ist. Beim Demokratiebarometer ist dieser, solange die Transformation auf Basis derselben Vergleichsjahre durchgeführt wird möglich. Allerdings mit der Konsequenz, dass Werte über 100 und unter 0 sich ergeben können – problematisch ist dies jedoch nicht, so man klar wie die Autoren des Demokratiebarometers darauf hinweist, dass der Wert 100 kein absolutes Optimum darstellt. Ähnlich wie bei den SGI sind die aggregierten Werte des

[18] Die Aggregation von den Indikatorwerten zu den Subkomponenten erfolgt auch beim Demokratiebarometer mittels arithmetischem Mittel.

Demokratiebarometers aber sinnvollerweise nurmehr ordinal zu interpretieren (vgl. Brusis 2009b: 544-545), die verwendeten Aggregationsregeln (arithmetisches Mittel und Arkustangens-Funktion) benötigen jedoch metrisches Skalenniveau und auch die letztlich berichteten Indexwerte werden nicht als Ranking präsentiert, sondern als eine Intervallskala implizierende Komma-Werte.

Für das 2007 Demokratiebarometer haben wir als Alternative alle Rohdaten der 100 Indikatoren z-transformiert und dann als arithmetisches Mittel über die drei Prinzipien aggregiert. Die Befunde sind erstaunlich und zeigen die Sensitivität des Verfahrens. Zwar sind beide Ergebnisse mit r = 0,94 hoch korreliert, doch von den 30 Ländern im Datensatz bleiben lediglich bei 5 die Rangplätze unverändert. Im Durchschnitt verändern die Länder um 2,5 Rangplätze ihre Position. Am gravierendsten ist der Fall der Niederlande, die sich um 7 Plätze bei einer z-Standardisierung verbessern würden.

3.3.3 Arkustangensfunktion als Aggregationsregel

Die Aggregation erfolgt auf den untersten beiden Ebenen von den Indikatoren über die Subkomponenten zu den Komponenten mittels des arithmetischen Mittels. Für die folgenden drei Aggregationsschritte bis hin zum Gesamtindex wird die Arkustangens-Funktion verwendet. Zunächst wird also davon ausgegangen, dass die einzelnen Indikatoren (Subkomponenten) für die entsprechende Subkomponente (Komponente) alle dasselbe Gewicht haben und dass es sich um jeweils hinreichende Bedingungen handelt. Im späteren Verlauf werden dann die Komponenten, Funktionen und Prinzipien jeweils als notwendige Bedingungen für die jeweils höhere Ebene betrachtet, weshalb ein multiplikatives Aggregationsverfahren kombiniert mit der Arkustangens-Funktion gewählt wurde.[19] Warum gerade ab der dritten Ebene die Aggregation auf der Logik notwendiger Bedingungen beruht, darunter jedoch die einzelnen Indikatoren/Subkomponenten nicht notwendig sind, wird theoretisch nicht begründet. Betrachtet man dieses Phänomen jedoch aus empirischer Warte wird diese Entscheidung verständlich. Würde man bereits die einzelnen Indikatoren als notwendige Bedingungen auffassen, so würde dadurch, dass die Skalierung für jeden Indikator immer mindestens ein Land mit einer Null versieht, weder während des Zeitraums des Blueprint-Samples noch für das Jahr 2007 auch nur ein Land keine Null für den Gesamtdemokratiewert erhalten. Wenn man nur den ersten Aggregationsschritt mittels arithmetischem Mittel durchführt, dann wären es immer noch 24 von 30 Ländern die einen Demokratiewert von Null aufweisen würden. Wenn man hingegen wie das Demokratiebarometer erst ab dem dritten Aggregationsschritt die Logik notwendiger Bedingungen ansetzt, fällt über den gesamten Beobachtungszeitraum kein einziges Land auf den Wert Null. Dies würde nur geschehen, wenn ein Land in allen in der Regel sechs Indikatoren, die eine Komponente auszeichnen einen genauso schlechten Wert hätte wie das schlechteste Land während des Blueprint-

[19] Um den Wertebereich in etwa auf 0 bis 100 festzusetzen werden an der eigentlichen Arkustangens-Funktion noch gewisse arithmetische Umformungen vorgenommen:
(arctan(Komponente1 * Komponente2)***1,2/4000)*80**). Den Grund für diese Umformung muss sich der Leser jedoch selber herleiten, in den Publikationen des Demokratiebarometers finden sich hierzu keine Informationen.

Sample Vergleichszeitraums. Durch die Arkustangensfunktion soll eine begrenzte Substituierung ungleicher Werte ermöglicht werden oder anders ausgedrückt, es soll eine möglichst gleichmäßige Verteilung beispielsweise der drei Prinzipien belohnt werden. Letztlich vergrößert dieses Aggregationsverfahren mittels Progression die Messunterschiede und es wird wiederum eine künstliche Varianz geschaffen. Die Annahme dahinter ist, dass die Erhöhung eines an sich schon sehr hohen Wertes nurmehr einen marginalen Grenznutzen für die Demokratiequalität hat. Eine Annahme, die so aus dem grundlegenden Konzept von Demokratie als dynamisches Gleichgewicht zwischen Freiheit, Gleichheit und Kontrolle nur bedingt abgeleitet werden kann. Tabelle 3 verdeutlicht den Effekt der vom Demokratiebarometer verwendeten Arkustangens-Aggregationsregel im Vergleich zu den weitaus üblicheren Varianten des arithmetischen und geometrischen Mittels.

Tabelle 3: Unterschiedliche Aggregationsregeln im Vergleich

	Wert 1	Wert 2	arithmetisches Mittel	geometrisches Mittel	Arkustangens-Funktion
Land A	1	99	50,00	2,38	9,95
Land B	25	75	50,00	40,99	43,30
Land C	49	51	50,00	51,46	49,99
Land D	50	50	50,00	51,48	50,00

Was deutlich wird ist, dass die Arkustangens-Formel noch stärker als die reine Multiplikation, welche dem geometrischen Mittel zugrunde liegt, ungleiche Verteilungen der Basiswerte bestraft und gleichzeitig ähnliche Verteilungen so stark belohnt, dass der aggregierte Wert letztlich sogar größer sein kann als die Werte, aus denen er gebildet wurde.

Ferner ist auch die Gewichtung ein Einfallstor für Verzerrungen. Die Autoren des Demokratiebarometers möchten allen 100 Indikatoren in der Messung ein gleiches Gewicht zuweisen. Über die unterschiedlichen Besetzungszahlen der Subindikatoren und Indikatoren werden diese jedoch in letzter Konsequenz (vgl. Abbildung 1) nicht gleichgewichtet. Zudem ist inhaltlich die angestrebte Gleichgewichtung nicht plausibel: So wird in der gegenwärtigen Konstruktion den Einstellungen zum Schwarzfahren das gleiche Gewicht zugewiesen wie die Möglichkeit direktdemokratisch abzustimmen.

Insgesamt ist die Aggregationsmethode durchaus ausführlich beschrieben und damit nachvollziehbar, allerdings weist sie, wie gezeigt wurde, gewisse Differenzen zur grundlegenden Konzeption auf. Zudem, und dies ist ein Aspekt, der nicht unterschätzt werden sollte, dürfte die Aggregation mittels Arkustangensfunktion von einem Großteil der Leser mathematisch nur bedingt nachzuvollziehen sein. Eine simplere Aggregationsregel, beispielsweise über das geometrische Mittel, würde zwar keine so starke Abstrafung von Ländern mit unterschiedlich ausgeprägten Demokratiekomponenten ermöglichen, sie wäre jedoch weitaus besser nachzuvollziehen.

3. Die Qualität der Qualitätsmessung – zur Validierung des Demokratiebarometers

Die Autoren des Demokratiebarometers sind sich durchaus der Notwendigkeit bewusst ihre eigene Messung anhand gewisser Gütekriterien zu überprüfen. Sie beschränken sich hierbei jedoch auf die drei Validitätskriterien der Konstrukt-, Kriteriums- sowie der endogenen Validität. Das was man klassischerweise zunächst bei einer Indexkonstruktion unter Validität verstehen würde, nämlich die Inhaltsvalidität, die abprüft, ob das theoretisch entwickelte Konzept auch passgenau über messbare Indikatoren abgedeckt wird, wird hingegen vernachlässigt. Dabei bestehen in Anbetracht des gleichzeitig vom Stamm und den Blättern aufeinanderzuwachsenden Konzeptbaumes gerade hier in unseren Augen Zweifel, was die Passgenauigkeit der gewählten Indikatoren anbelangt.

3.1. Konstruktvalidierung über Clusteranalyse?

Als Möglichkeit einer Konstruktvalidierung schlagen die Autoren eine Clusteranalyse auf Basis der neun Funktionen vor.[20] Sie nehmen an, dass sich unterschiedliche Cluster bilden, die jeweils abweichende Demokratieprofile repräsentieren. Grundsätzlich ist dies eine sinnvolle Option, zumindest, wenn man davon ausgeht, dass die zugrundeliegende Theorie ebenfalls korrekt ist. Aus methodischer Perspektive ist freilich einzuwenden, dass einerseits alternative Clusterungsalgorithmen zu vollkommen anderen Gruppierungen führen und dieses Verfahren sehr sensitiv sein kann (Wagschal 1999). Der Versuch die Clusterlösung für das Jahr 2005 auf Basis des im Internet verfügbaren Demokratiebarometer-Datensatzes (standardisierte Version) zu replizieren führte nicht zu der von Bühlmann und seinen Kollegen präsentierten Clusterlösung(2011a: 27). Abbildung 5 zeigt die sich ergebende Clusterung in Form eines Dendrogramms. Wichtiger noch als die Frage was für Cluster sich bilden, ist jedoch für die Macher des Demokratiebarometers die Tatsache, dass sich überhaupt unterschiedliche Gruppen identifizieren lassen. Ob diese Gruppen aber entsprechend der neun Funktionen sich so stark unterscheiden, dass es auch inhaltlich sinnvoll ist von distinkten Demokratieprofilen zu sprechen, kann eine Clusteranalyse alleine nicht klären. Hierzu ist es sinnvoll die Mittelwerte der Gruppen zu vergleichen. Bühlmann et al. machen dies ebenfalls mittels einer einfachen deskriptiven Tabelle. Hier soll nun mittels eines Mittelwertvergleichs untersucht werden, ob sich die Cluster entsprechend der Mittelwerte der neun Funktionen über die in einer Gruppe enthaltenen Länder wirklich signifikant unterscheiden.[21]

Abbildung 5: Clusteranalyse für die neun Demokratiefunktionen (2005)

[20] Eine Alternative böten Korrelationsanalysen.

[21] Da sich die Ergebnisse der Clusteranalyse von Bühlmann et al. nicht replizieren ließen, wird für den Mittelwertvergleich auf die sich in unserer Clusteranalyse ergebende Gruppeneinteilung zurückgegriffen. Es werden ebenfalls vier Cluster betrachtet: Gruppe 1: Island, Norwegen, Belgien, Dänemark, Niederlande, Slowenien, Finnland, Schweden, Neuseeland, Australien, Kanada, Österreich, Deutschland, Spanien; Gruppe 2: Tschechische Republik, Ungarn, Italien, Costa Rica, Polen, Malta, Südafrika; Gruppe 3: Luxemburg, Schweiz, Zypern; Gruppe 4: Frankreich, Großbritannien, Irland, Portugal, Japan, USA.

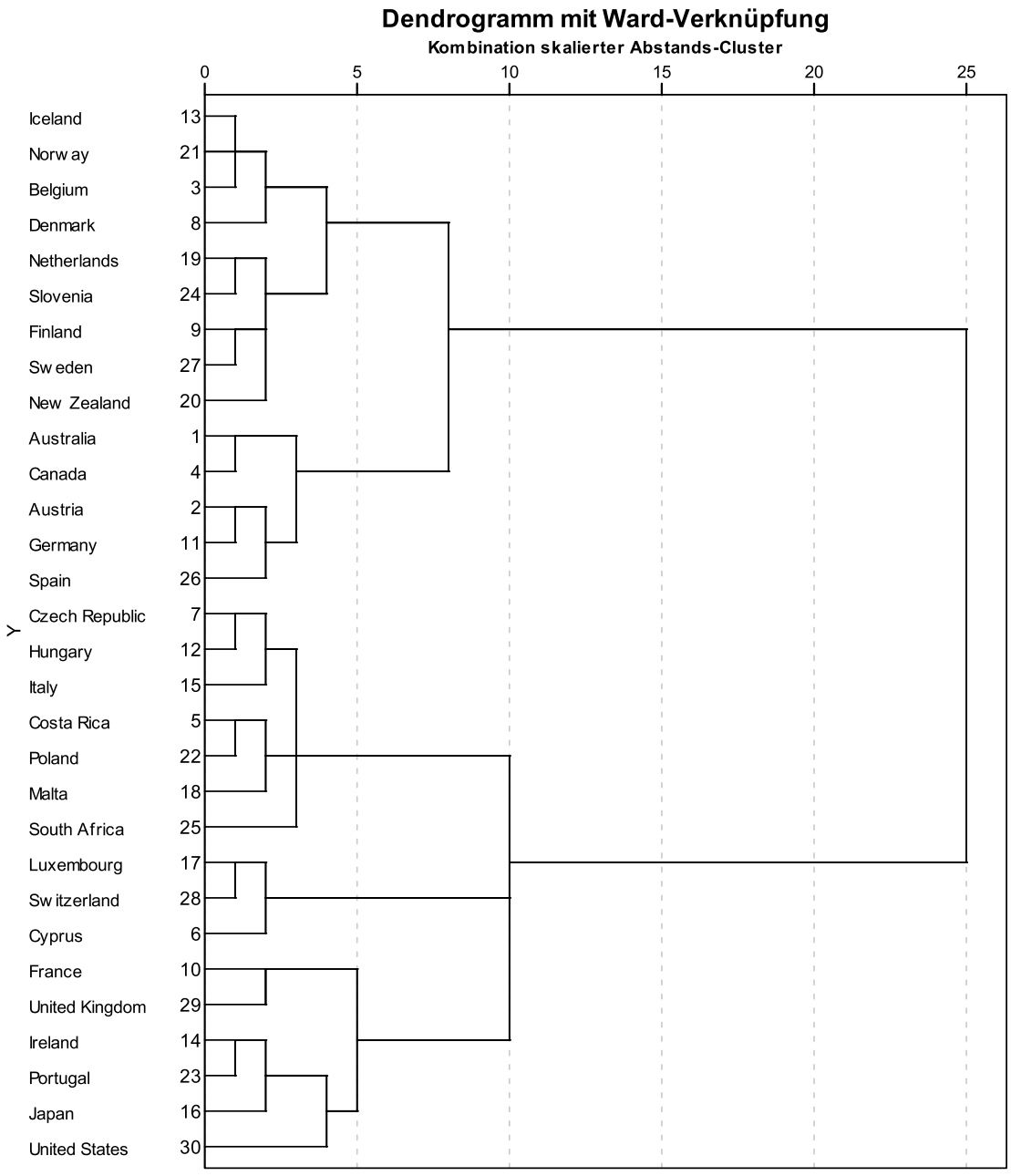


Tabelle 4 zeigt, dass nur 21 der 54 Gruppenvergleiche auf der Ebene der Funktionen signifikant unterschiedliche Mittelwerte aufweisen und auch auf den höher aggregierten Ebenen lassen sich die Gruppen zum Teil nicht signifikant differenzieren. Insofern muss auch die Konstruktvalidierung als tendenziell gescheitert angesehen werden.

Tabelle 4: Mittelwertvergleich zwischen den gefundenen Clustern

	Signifikante Unterschiede zwischen den Gruppen vorhanden? (T-Test für unabhängige Stichproben, 2-seitig, 95% Signifikanzniveau)					
	Gruppe 1 vs. Gruppe 2	Gruppe 1 vs. Gruppe 3	Gruppe 1 vs. Gruppe 4	Gruppe 2 vs. Gruppe 3	Gruppe 2 vs. Gruppe 4	Gruppe 3 vs. Gruppe 4
Individuelle Freiheiten	Nein	Nein	Nein	Nein	Nein	Nein
Rechtsstaatlichkeit	Ja	Nein	Nein	Ja	Ja	Nein
Öffentlichkeit	Nein	Nein	Nein	Ja	Nein	Nein
Wettbewerb	Ja	Nein	Ja	Nein	Nein	Ja
Gewaltenkontrolle	Nein	Nein	Nein	Ja	Nein	Nein
Regierungs- und Implementationsfähigkeit	Ja	Nein	Ja	Ja	Nein	Nein
Transparenz	Ja	Ja	Nein	Nein	Ja	Ja
Partizipation	Ja	Ja	Ja	Nein	Nein	Nein
Repräsentation	Ja	Ja	Ja	Nein	Nein	Nein
Freiheit	Ja	Nein	Nein	Ja	Nein	Nein
Kontrolle	Ja	Nein	Ja	Nein	Nein	Nein
Gleichheit	Ja	Ja	Ja	Nein	Ja	Nein
Demokratiequalität	Ja	Ja	Ja	Nein	Nein	Nein

4.2 Externe Validität

Bei der Überprüfung der externen Validität (die unverständlicherweise als endogene Validität bezeichnet wird) greifen die Autoren des Demokratiebarometers auf die Worldwide Governance Indicators (WGI) der Weltbank zurück und zeigen, dass generell starke Korrelationen zwischen einem aus den sechs Dimensionen aggregierten Governance-Gesamtindex und dem Demokratiebarometer bestehen, wobei dessen einzelne Funktionen unterschiedlich stark mit den Governance Indicators zusammenhängen. Insbesondere die Kontrollfunktionen würden weniger stark korrelieren, was laut Bühlmann et al. dafür spricht, dass die Governance Indicators diese Dimension ähnlich wie andere Demokratiemaße und im Gegensatz zum Demokratiebarometer allzu stiefmütterlich behandeln würden (Bühlmann et al. 2011c: 30-31). Fraglich ist allerdings, weshalb für den Test der externen Validität auf ein Maß zurückgegriffen wird, das gar nicht beabsichtigt Demokratie bzw. demokratische Qualität zu messen. Ein weiteres weniger theoretisch sondern praktisch gelagertes Problem der externen Validierung über die WGI ist zudem in der potentiellen Quellenüberschneidung der beiden Messungen zu sehen. Der direkte Vergleich der Quellen des Demokratiebarometers und der WGI deutet auf die Existenz deutlicher Überschneidungen hin. Von den 30 Quellen die in die WGI einfließen finden sich 10 auch in den Quellen des Demokratiebarometers.[22] Zwar lässt sich aufgrund der komplexen Indikatorenzuordnung und Konstruktion des Demokratiebarometers nicht genau rekonstruieren welche Indikatoren aus anderen Messungen übernommen wurden. Basierend auf dem Codebuch lässt sich jedoch zumindest für die 9 Funktionen die Quellenüberschneidung aufschlüsseln (vgl. Tabelle 5). In

[22] Darüber hinaus werden die WGI selbst als Quelle im Demokratiebarometer ausgewiesen.

der Gesamtbetrachtung liegt der Anteil der sich überschneidenden Quellen bei knapp einem Sechstel, und in manchen Funktionen – z.B. Rule of Law – bei über einem Drittel. Der gefundene Zusammenhang zwischen den beiden Maßen kann vor diesem Hintergrund weniger verwundern und ist daher als externe Validierung nurmehr bedingt geeignet.

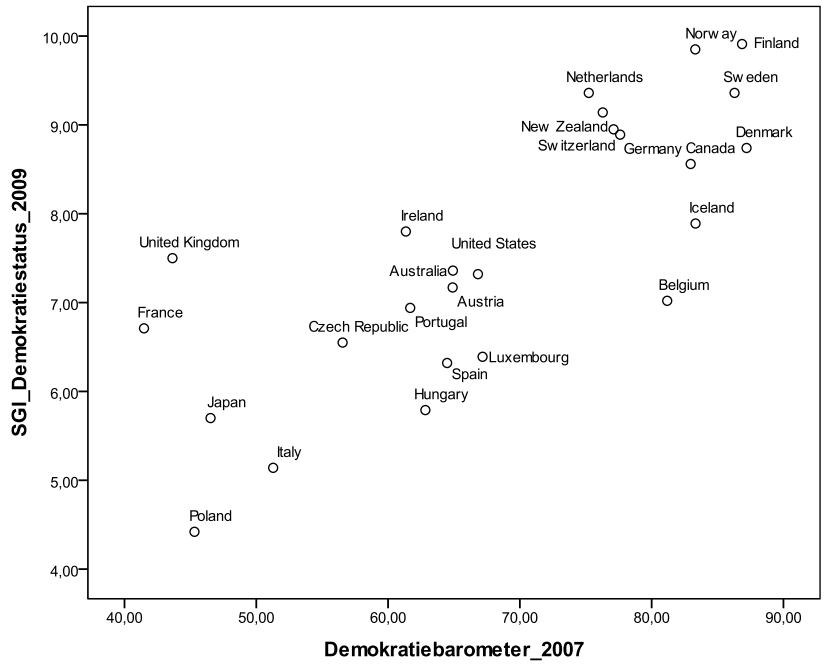
Tabelle 5: Identische Quellen von Demokratiebarometer und Governance Indicators

Komponente	Identische Quelle (Gesamtanzahl der Quellen)
Individuelle Freiheiten	6/18
Rechtsstaatlichkeit	11/30
Öffentlichkeit	2/45
Wettbewerb	0/31
Gewaltenkontrolle	0/19
Regierungs- und Implementationsfähigkeit	7/22
Transparenz	6/26
Partizipation	12/76
Repräsentation	3/35
Gesamtanteil	**47/302**

Eigene Darstellung basierend auf Bühlmann et al. (2011). Democracy Barometer. Codebook for Blueprint Dataset Version 2. Aarau: Zentrum für Demokratie.) und WGI (http://info.worldbank.org/governance/wgi/sources.htm)

Ein Rückgriff auf etablierte Maße (FH, Polity IV usw.) stellt zwar aus offensichtlichen Gründen auch keine Alternative dar, aber ein Vergleich mit den Sustainable Governance Indicators, die zumindest vorgeben ebenfalls ein feineres Messinstrumentarium für Demokratie zu verwenden, wäre durchaus angebracht. Abbildung 6 trägt den SGI Status der Demokratie für 2009 gegen den Demokratiebarometerwert für 2007 ab. Das Ergebnis dürfte die Macher des Demokratiebarometers durchaus erfreuen. Ein Zusammenhang zwischen den beiden Maßen lässt sich nicht leugnen. Auch wenn aufgrund unterschiedlicher Länderabdeckung nur 25 Staaten in den Vergleich einbezogen werden können so ergibt sich doch ein Spearman Rangkorrelationskoeffizient von 0,744 (signifikant auf dem 99%-Niveau).

Abbildung 6: Demokratiebarometer vs. SGI-Demokratiestatus



Allerdings zeigt das Streudiagramm auch an, welche Fälle beim Demokratibarometer im Vergleich zum Demokratiestatus des SGI unterbewertet sind. Dies sind vor allem Frankreich und das Vereinigte Königreich, die man als lange etablierte Demokratien nicht auf den beiden letzten Plätzen vermutet hätte (vgl. auch Tabelle 1). Dagegen ist Belgien beim Demokratiebarometer im Vergleich zum SGI überbewertet. Interessant wäre bei einem solchen Abschneiden von zentralen Fällen, die externe Validität der Indikatoren zu überprüfen, die für diesen Befund verantwortlich sind.

4.3 Adäquate Präsentation der Daten im Internet fördert Nachvollziehbarkeit

Als grundsätzlich gelungen, insbesondere im Vergleich zu etablierten Indizes wie Freedom House oder Polity IV, ist die Präsentation der Daten zu bezeichnen. Auf der Homepage des Demokratiebarometers wird nicht nur der Gesamtindex dargestellt, sondern gleichzeitig auch die Methodik beschrieben, die Ergebnisse für alle Staaten in einzelnen Länderstudien verdeutlicht sowie die Datenbasis des Demokratiebarometers sowohl in standardisierter als auch in unstandardisierter Rohdatenform zum Download zur Verfügung gestellt. Zudem bietet sie die Möglichkeit eigene auf spezifische Fragestellungen angepasste Vergleichsgrafiken direkt zu erstellen. Die umfassende Dokumentation und Datenoffenlegung erhöht zweifellos die Nachvollziehbarkeit. Einzig die Zuordnung der Daten zu den ursprünglichen Quellen gestaltet sich schwieriger als nötig. Die kurzen Beschreibungen, die zusammen mit der Angabe aus welcher Sekundärquelle die Daten stammen in den Tabellen des Methodologie-Papiers (Bühlmann et al. 2011c: 20-42) aufgeführt sind, erlauben zumeist keine einfache Zuordnung. Würde in einer weiteren Spalte stehen, wie der Indikator in seinem ursprünglichen Datensatz bezeichnet wird, würde dies die Nachvollziehbarkeit nochmals erhöhen. Zudem könnte der geneigte Leser so auch einfacher für sich entscheiden, ob der verwendete Indikator wirklich passend ist.

4.4. Objektivität unklar und nicht thematisiert

Das dritte Kriterium der Objektivität bezeichnet, dass der Indikator unabhängig sein sollte von normativen Verzerrungen, insbesondere etwa vom Einfluss derjenigen, welche die Datenerhebung durchführen bzw. das Konzept des Indikators definieren. Bei Umfragen kann man diese „Unabhängigkeit vom Testleiter“ durch Intercoderreliabilität prüfen, d.h. wie übereinstimmend Länder bewertet werden. Objektivität ist jedoch auch von Bedeutung, wenn statistische Daten externer Quellen herangezogen werden. Prima vista gilt oftmals die Überlegenheit solcher Sekundär- und Aggegatdaten gegenüber Individualdaten und eigenen Primärdaten. Doch auch im Falle von Sekundärdaten ist die Qualität der Daten nicht immer gesichert (Elklit 1994). Bei 100 Variablen im Datensatz kann eine Qualitätsprüfung nur stichprobenartig erfolgen, aber anhand der direktdemokratischen Beteiligungsvariablen DIRDEM soll das Problem verdeutlicht werden. Die Autoren verwenden einen zusammengesetzten Indikator, der aus Daten für das obligatorische und fakultative Referendum besteht (die wiederum standardisiert werden). Am Ende liegt die Schweiz wie

erwartet vorne, auf Platz 2 jedoch Spanien, in dem seit 1990 nur ein einziges nationales Referendum stattfand. Damit befindet sich Spanien in der Gruppe der Länder, die am wenigsten direktdemokratisch abstimmen („Rules in Use", Quelle C2D Aarau). Auch bei den konstitutionellen Möglichkeiten ist Spanien bestenfalls im Mittelfeld der europäischen Länder. Diese „Rules in Form" werden vom IRI-Index (Kaufmann und Waters 2004) gemessen und Spanien gehört hier in die Gruppe der Vorsichtigen. So liegt Italien, welches bei beiden Indikatoren vor Spanien (die aus den gleichen Datenquellen wie bei Demokratiebarometer kommen) liegt im Demokratiebarometer hinter diesem Land. Solche Fehler können nur durch Qualitätskontrolle ausgemerzt werden, was durch die lobenswerte Transparenz immerhin für Externe möglich ist. Problematisch ist jedoch das Negieren der Objektivitätsproblematik bei Bühlmann et al. (2011a; 2011b).

4. Fazit: Ein vielversprechender Ansatz – mit deutlichen Defiziten

Betrachtet man das Problem der mangelnden Varianz innerhalb etablierter Demokratien als eine der zentralen Herausforderungen der Demokratiemessung, kommt das Demokratiebarometer dem eigenen Anspruch nach zu urteilen einer Komplettlösung gleich. Zum einen sollen methodische Defizite bisheriger Ansätze umgangen und zum anderen eine wesentlich präzisere Analyse etablierter Demokratien erreicht werden. Sowohl die Vorgehensweise bei der Entwicklung und Konstruktion des Messvorschlages als auch die eigentlichen Ergebnisse des Demokratiebarometers lassen hierbei zunächst vermuten, dass dieses neue Instrument seine hehren Ziele auch erfüllt. Im Hinblick auf das „Design" muss man dem Demokratiebarometer insbesondere im Vergleich zu vielen „alten" Messansätzen eine stärkere Sensitivität für die Problematik und Relevanz der theoretischen Verankerung und entsprechender Konzeptualisierung zu Gute halten. Dies wird vor allem in der relativ breiten Erläuterung des Demokratiekonzepts deutlich. Ein weiterer wichtiger Aspekt den das Demokratiebarometer explizit adressiert, ist die bis dato eher wenig beachtete Diskrepanz zwischen de jure und de facto Indikatoren in der Demokratiemessung (vgl. Lauth 2004). Ebenso positiv schneidet die Messanlage im Hinblick auf die Verfügbarkeit und Präsentation der Daten über die Projekthomepage ab. Schliesslich lässt der Blick auf die bisherige Probemessung den Schluss zu, dass das Demokratiebarometer seine eigentliche Kernaufgabe, die Binnendifferenzierung etablierter Demokratien, scheinbar erfüllt.

Auf den zweiten Blick treten jedoch zahlreiche problematische Aspekte zu Tage, die uns trotz der genannten Vorzüge dazu veranlassen die Leistungsfähigkeit des Demokratiebarometers kritisch zu bewerten. Neben des bereits von Kaina (2008) und Lauth (2010) adressierten Problems, dass das grundsätzliche, vom Demokratiebarometer verwendete Konzept von(demokratischer) Qualität nur mangelhaft expliziert wird und dabei dessen inhärentes Spannungsverhältnis zwischen den drei Demokratieprinzipien verloren geht, lassen sich sowohl auf der Ebene der Konzeption, der Messung als auch der Aggregation Defizite feststellen.

Auf der konzeptionellen Ebene lässt sich festhalten, dass die Wahleiner maximalistischen und damit zwangsläufig anspruchsvolleren Demokratiedefinition zu wählen zwar als zielführend anzusehen ist, Sie jedoch in Verbindung mit der gewählten Operationalisierung zu einem potentiellen Problem wird. Erstens, verliert die logische Verbindung der verschiedene Konzeptebenen umso weiter wir uns im Konzeptbaum hinauf bewegen an Argumentationstiefe und Überzeugungskraft. Gerade das Vorgehen in der Verbindung der Subkomponenten und der eigentlichen Indikatoren hätte hierbei mehr Raum einnehmen sollen und müssen. Zweitens, und dies erscheint auch als das schwerwiegendere Problem, wird die Binnendifferenzierung über einen potentiell überladenen Indikatorenapparat erreicht. Gerade in Verbindung einer mangelnden Erläuterung der Indikatorenauswahl drängt sich der Verdacht auf, dass das Ziel der Varianzerzeugung zumindest teilweise höher als theoretische Geeignetheit gewichtet wurde. Denn auch wenn der Auwahlprozess wie von den Autoren selbst hervorgehoben „basically theory driven" erfolgt, lassen Sie den geneigten Leser an diesem Prozess nicht ausreichend teilhaben um die theoretische Geeignetheit individuell zu bewerten und nachzuvollziehen. Die Messanlage läuft hierbei Gefahr Indikatoren aufzunehmen, die fraglich im Hinblick auf ihre „cross-systemische" Äquivalenz sind und teilweise nur mit viel gutem Willen als Proxy für demokratische Qualität angesehen werden können.

Im Hinblick auf die eigentliche Messung erscheint die weitgehende Ablehnung auf Expertenbefragungen beruhender Indikatoren teilweise willkürlich. Der Fokus auf *hard data* im Sinne einer gesteigerten Reliabilität überlagert hierbei die Sicherstellung der Validität der Messung. Die Ebene der Aggregation bietet aus unserer Sicht die größte Angriffsfläche und zugleich die wesentlichste Schwachstelle des Demokratiebarometers. Erstens ist die gewählte Reskalierung als problematisch einzustufen, Alternativen sollten zumindest durchgespielt und diskutiert werden, da sie wie gezeigt die Rangfolge deutlich ändern können. Zum einen steht Sie in Widerspruch zu der ursprünglichen Konzeption des Demokratiebarometers, die hohe Demokratiequalität durch unterschiedliche Kombinationen aus Gleichheit, Freiheit und Kontrolle verwirklicht sieht, wenn ein möglichst hoher Wert bei jedem Indikator als besonders gut für die Demokratiequalität betrachtet wird. Zum anderen, wird die Frage der Gewichtung nicht thematisiert. Schließlich erscheint die vorgenommene Reskalierung als reine Aufspreizung mangelnder Unterschiede, die darüber hinaus im ungünstigsten Fall sogar zu einer Verschiebung der Rangfolge der Länder im Demokratiebarometer führen kann. Vor diesem Hintergrund entpuppt sich die erreichte höhere Differenzierung als fabrizierte Varianz. Schliesslich steht die verwendete Aggregationsregel in einen potentiellen Widerspruch mit der zugrunde liegenden Demokratiekonzept und die Verwendung des Arkustangens bedeutet eine unnötige Verkomplizierung ohne besonders großen Mehrwert, die dabei die Nachvollziehbarkeit des Aggregationsprozesses deutlich erschwert. Die Alternative wäre das geometrische Mittel.

Die von den Autoren vorgenommenen Prüfungen bzw. Überprüfungen der Güte des eigenen Messvorschlages sind zunächst begrüßungswürdig, doch bietet das gewählte Vorgehen erneut Raum

für Kritik. Neben der prinzipiellen Vernachlässigung der Inhaltsvalidität, erweist sich die Konstruktvalidität nur als bedingt gewährleistet. Plausibilitätschecks, statistische Sensitivitätsanalysen und die nicht diskutierte Objektivitätsproblematik geben Raum zur Verbesserung. Die Heranziehung der World Governance Indicators im Sinne einer externen Validierung betrachten wir nicht nur aufgrund der fragwürdigen theoretischen Geeignetheit sondern auch aufgrund der teils deutlichen Quellenüberschneidungen als wenig zielführend und schlagen daher die Verwendung alternativer Maße vor, die wirklich den Anspruch haben wie das Demokratiebarometer demokratische Qualität zu messen, beispielsweise die SGI.

In der Gesamtschau ziehen wir somit ein überwiegend skeptisches Resümee, denn auch wenn das Demokratiebarometer vieles besser als vorhergegangene Messvorschläge macht, überwiegen letztlich die Schwächen sowohl in theoretisch-konzeptioneller als auch in praktischer Hinsicht. Um ein letztes Mal die Metapher zu bemühen, die uns durch den bisherigen Artikel begleitet hat: Der Konzeptbaum des Demokratiebarometers mag zwar über stärkere Wurzeln verfügen als bisherige Demokratiemaße. Diese können aber nicht verhindern, dass er aufgrund eines fragilen Stammes und einer vor Blättern strotzenden Baumkrone Gefahr läuft umzuknicken.

Tabelle A1: Demokratiemaße im Vergleich

Name	Länderzahl	Zeitraum	Autoren	Demokratiedefinition	Operationalisierung
Bertelsmann Transformations Index	128	2003-2012	Bertelsmann-Stiftung unter der Führung von Werner Weidenfeld, Wolfgang Merkel, Aurel Croissant und Hans-Jürgen Puhle	BTI misst Transformation zu Demokratie und Marktwirtschaft, Demokratieindex daher Teil des Statusindexes „Das Demokratieverständnis des BTI geht weit über andere Definitionen von Demokratie hinaus, die sich vorrangig auf elementare Bürgerrechte und die Durchführung von freien Wahlen beschränken. Es umfasst eine Untersuchung der Staatlichkeit mit Einzelfragen zum staatlichen Gewaltmonopol und zu den Verwaltungsstrukturen eines Landes als Voraussetzung für politische Transformation ebenso wie die Analyse der Rechtsstaatlichkeit mit Blick auf Gewaltenteilung oder die Ahndung von Amtsmissbrauch. Zudem wird untersucht, inwieweit das demokratische System im Hinblick auf seine Akzeptanz, Repräsentativität und die politische Kultur konsolidiert ist.“	5 Kriterien des Demokratieindexes: Staatlichkeit, Politische Partizipation, Rechtsstaatlichkeit, Stabilität demokratischer Institutionen und Politische und gesellschaftliche Integration. Je 4 Fragen pro Kriterium (außer bei Stabilität, dort 2). Demokratieindex wird durch Mittelwert der 5 Kriterien gebildet (diese wiederum jeweils durch Mittelwert der Fragen) Öffentlich zugängliche Gutachten von je 2 Länderexperten (einer lokal, einer auswärtig), danach Anpassung der Werte durch Regionalkoordinatoren und Projektteam Ein Staat muss 6 Minimalbedingungen erfüllen um als Demokratie zu gelten: freie und faire Wahlen (mind. 6 Punkte), effektive Reg.gewalt, Organisat.- und Versammlungsfreiheit, Meinungs- und Pressefreiheit, Gewaltenteilung, Bürgerl. Freiheitsrechte (je mind. 3 P.)
Democracy and Dictatorship (DD)	202	1946-2008	José Antonio Cheibub, Jennifer Gandhi, James Raymond Vreeland	"[T]he minimalist conception of democracy we adopt here is procedural in the sense that it sees democracy simply as a method for choosing rulers."	DD: basiert auf 'objektiven Kriterien Austragen von Wahlen, die Existenz von mehr als einer politischen Partei und Veränderung in der Regierungsbesetzung; Alle Merkmale müssen vorliegen, damit es als Demokratie klassifiziert wird; Die Klassifizierung von Ländern ist entsprechend dichotom: Demokratie oder Diktatur.
Democracy Barometer	30	1990-2007	Daniel Bochsler, Wolfgang Merkel	"We define democracy as a political system that continuously redefines and alters itself, depending on ongoing political as well as societal deliberation"	Demokratiequalität wird gemessen über die drei Dimensionen Freiheit, Kontrolle und Gleichheit; Jede der Dimensionen beinhaltet drei unterfunktionen, diese wiederum Unterkategorien, welchen schließlich wiederum Indikatoren zugeordnet sind; insgesamt setzt sich das Demokratiequalitätsmaß aus 100 Indikatoren zusammen; Bei der Aggregierung werden die empirischen Minima und Maxima von Ausprägungen für die Skalierung der Werte herangezogen; dazu werden bei der Normierung sogenannte blueprint Demokratien betrachtet die Freedom House Werte 1.5 oder niedriger und Polity IV Werte von mindestens 9 zwischen 1995 und 2005 aufweisen; bei der Aggregierung wird innerhalb der jeweiligen Ebene/Gruppe von Elementen das gleiche Gewicht angenommen.
Democracy-Index	167	2006-2011	Economist Intelligence Unit	Stützt sich auf das Konzept von Freedom House und erweitert dieses um mehrere Faktoren. Über die Angabe der fünf Basiskategorien „electoral process and pluralism; civil liberties; the functioning of government;political participation; and political culture“ wird keine weitere theoretische Fundierung angegeben., Es wird keine theoretische Fundierung angegeben.	Additive Aggregation und Mittelwertbildung der fünf Kategorien (Skala von 1-10)
Freedom House Index (Freedom in the World)	195	1972-2011	Freedom House, unter Führung von David Kramer	Freedom="Freedom House measures freedom according to two broad categories: political rights and civil liberties. Political rights enable people to participate freely in the political process through the right to vote, compete for public office and elect representatives who have a decisive impact on public policies and are accountable to the	Dimensionen Civil Rights und Political Freedoms, beiden gemessen auf einer Skala von 1 (am meisten frei) bis 7 (am wenigsten frei); Die Einstufung basiert auf einer Gesampukntzahl bis 100 für jedes Land auf der Grundlage von 10 Fragen zu "political rights" und 15 Fragen zu "civil liberties";

				electorate. Civil liberties allow for the freedoms of expression and belief, associational and organizational rights, rule of law, and personal autonomy without interference from the state. Freedom House's definition of freedom is derived in large measure from the Universal Declaration of Human Rights that was adopted by the United Nations General Assembly in 1948. The rights enumerated in the Universal Declaration include freedom of religion, expression, and assembly; freedom from torture; and the right to take part in the government of his or her country. "	Die Vergabe von Punkten beruht auf "a multilayered process of analysis and evaluation by a team of in-house and consultant regional experts and scholars"; Freiheit wird letztlich verstanden als der Durchschnitt der beiden genannten Dimensionen (freedom rating): Free (1.0 - 2.5), Partly Free (3.0 - 5.0), Not Free (5.5 - 7.0)
Neuer Index der Demokratie (NID)	57 bzw. 27	1996-2002 (bis 2004 für die EU-Länder)	Hans-Joachim Lauth	Das Demokratiekonzept des NID stützt sich auf vorangehende Überlegungen bei Freedom House, Vanhanen, Dahl und Polity. Es beansprucht eine Ergänzung durch die Berücksichtigung von "horizontaler *Accountability* und Rechtstaatlichkeit" vorzunehmen.	Variable NID3D: beruht auf Polity: DEMOC (0 bis +10), der Freedom House Political Rights Skala (auf 0 bis 10 skaliert) und dem Weltbank governance indicator Rule of Law"; NID3D = dritte Wurzel des Produktes der drei Komponenten; Variable Staatlichkeit=Indikator "Political Stability" der Weltbank governance indicators; NID=Wurzel aus dem produkt von NID3D und Staatlichkeit
Polity IV Polity Score	164	1800-2010	Monty G. Marshall, Keith Jaggers, Ted R. Gurr (Initiator); untersützt von Political Instability Task Force, Societal-Systems Research und Center for Systemic Peace	"Democracy is conceived as three essential, interdependent elements. One is the presence of institutions and procedures through which citizens can express effective preferences about alternative policies and leaders. Second is the existence of institutionalized constraints on the exercise of power by the executive. Third is the guarantee of civil liberties to all citizens in their daily lives and in acts of political participation. Other aspects of plural democracy, such as the rule of law, systems of checks and balances, freedom of the press, and so on are means to, or specific manifestations of, these general principles."	DEMOC (0 bis +10): Institutionalisierte Demokratie gemessen durch Indikatoren für Wettbewerb bei politischer Partizipation, Offenheit und Wettbewerb bei der Bestellung der Exekutive und Beschränkungen der Spitze der Exekutive; AUTOC (0 bis +10): Autokratiegrad gemessen durch Indikatoren für Wettbewerb bei politischer Partizipation, Regulierung politischer Partizipation, Offenheit und Wettbewerb bei der Bestellung der Exekutive und Beschränkungen der Spitze der Exekutive; POLITY (-10 bis +10): Kombinierter Wert durch Subtraktion AUTOC von DEMOC
Polyarchie and Contestation Scales	196	1985, 2000	Michael Coppedge, Wolfgang Reinecke	Stützt sich auf den Polyarchiebegriff von Robert Dahl; "Polyarchy is defined as the set of institutional arrangements that permits public opposition and establishes the fight to participate in politics. In these two respects-public contestation and inclusiveness-polyarchy is similar to the concept of democracy. However, polyarchy is not, and was not intended to be, exactly equivalent to democracy" (Coppedge/Reinecke 1990: 51).	Indikatoren: Faire Wahlen, Organisationsfreiheit, Meinungsfreiheit und Medienpluralismus; Kodierung durch Autoren und Studenten gemäß Vorgabe vpn Codes.
Sustainable Governance Indicators	31 (OECD)	2009, 2011	Martin Brusis, Jörg Siegmund; Bertelsmann Stiftung	"The SGI's concept of democracy […] includes not only the rights of political participation and electoral competition, but also the rule of law […] There are a series of questions designed to address whether citizens face discrimination in the electoral process, how citizens can access public information, the degree to which the media are independent and diversified, how well states protect civil rights, and whether the government and administration act predictably and in accordance with the law."	Dimensionen: Wahlprozess (vier Indikatoren), Informationszugang (drei Indikatoren), Bürgerrechte (drei Indikatoren), Rechtsstaatlichkeit (vier Indikatoren); diese von Experten beurteilen Dimensionen/Indikatoren werden ergänzt durch objektive Variablen für Rechtsstaatlichkeit und Korruption; Alle Indikatoren (teils nach linearer Transformation) messen auf einer Skala von 1 bis 10 und werden gleichgewichtet addiert.
Vanhanen-Index	187	1810-2000	Tatu Vanhanen	Demokratie="[A] political system in which ideologically and socially different groups are legally entitled to compete for political power and in which institutional power holders are elected by the people and are responsible to the people."	Indikator Competition=100-Stimmenanteil der größten Partei; Indikator Participation=Wahlbeteiligte als Bevölkerungsanteil; Demokratie=(Competition x Participation) / 100

Literatur

Abromeit, Heidrun 2004: Die Messbarkeit von Demokratie: Zur Relevanz des Kontextes, in: *Politische Vierteljahresschrift* 45 (1), 73 - 93.

Beetham, David, Carvalho, Edzia, Landman, Todd und Weir, Stuart 2009: *Assessing the Quality of Democracy - A Practical Guide,* Stockholm, International IDEA.

Berg-Schlosser, Dirk 1999: *Empirische Demokratieforschung - Exemplarische Analysen,* Frankfurt a.M., Campus Verlag.

Brusis, Martin 2009a: Designing Sustainable Governance Indicators - Assessment Criteria and Methodology, in: Stiftung, Bertelsmann (Hg.), *Sustainable Governance Indicators 2009*, Gütersloh, Verlag Bertelsmann Stiftung.

Brusis, Martin 2009b: Konzepte, Messansatz und Validierung der Sustainable Governance Indicators. Eine Replik auf Sebastian Jäckle und Rafael Bauschke, in: *Zeitschrift für Politikwissenschaft*, 4, 537 - 552.

Büger, Christian 2007: Beyond the gap: relevance, fields of practice and the securitizing consequences of (democratic peace) research, in: *Journal of International Relations and Development* 10 (4).

Bühlmann, M., Merkel, W., Müller, L., Giebler, H. und Wessels, B. 2011a: Demokratiebarometer – ein neues Instrument zur Messung von Demokratiequalität. http://www.democracybarometer.org/publications_de.html.

Bühlmann, M., Merkel, W., Müller, L., Giebler, H. und Wessels, B. 2011b: Democracy Barometer. Codebook for Blueprint Dataset Version 1, Aarau. http://www.democracybarometer.org/Images/Codebook_only_blueprints_JAN2011.pdf.

Bühlmann, M., Merkel, W., Müller, L., Giebler, H. und Wessels, B. 2011c: Democracy Barometer. Methodology Aarau. http://www.democracybarometer.org/Images/Methodical_Explanatory_Note_JAN_2011.pdf.

Campbell, David F.J. und Barth, Thorsten D. 2009: Wie können Demokratie und Demokratiequalität gemessen werden? Modelle, Demokratie-Indices und Länderbeispiele im globalen Vergleich, in: *SWS-Rundschau*, 49 (2), 208 - 233.

Castles, Francis G. und Mair, Peter 1984: Left-Right Political Scales. Some 'Expert' Judgements, in: *European Journal of Political Research*, 12 (1), 73-88.

Diamond, Larry Jay 2002: Thinking about Hybrid Regimes, in: *Journal of Democracy*, 13 (2), 21-35.

Diamond, Larry und Morlino, Leonardo 2005: *Assessing the Quality of Democracy,* Baltimore, Johns Hopkins University Press.

Elklit, Jorgen 1994: Is the degree of electoral democracy measurable? Experiences from Bulgaria, Kenya, Latvia, Mongolia and Nepal, in: Beetham, David (Hg.), *Defining and Measuring Democracy*, London/Thousand Oaks/New Delhi, Sage.

Foweraker, J. und Krznaric, R. 2000: Measuring Liberal Democratic Performance: A Conceptual and Empirical Critique, in: *Political Studies* 45 (3), 759 - 787.

Fuchs, Dieter 2004: Konzept und Messung von Demokratie. Eine Replik auf Heidrun Abromeit, in: *Politische Vierteljahresschrift*, 45 (1), 94 - 106.

Fuchs, Dieter und Roller, Edeltraud 2008: Die Konzeptualisierung der Qualität von Demokratie. Eine kritische Diskussion aktueller Ansätze, in: Brodocz, André, Llanque, Markus und Schaal, Gary (Hg.), *Bedrohungen der Demokratie*, Wiesbaden, VS-Verlag für Sozialwissenschaften.

Geis, Anna und Wagner, Wolfgang 2006: Vom „demokratischen Frieden" zur demokratiezentrierten Friedens-und Gewaltforschung in: *Politische Vierteljahresschrift*, 47 (2), 276 - 288.

Jäckle, Sebastian und Bauschke, Rafael 2009: Lässt sich Reformfähigkeit messen? Eine kritische Würdigung der Sustainable Governance Indicators, in: *Zeitschrift für Politikwissenschaft*, 3, 359 - 386.

Jäckle, Sebastian und Bauschke, Rafael 2010: Die Problematik bleibt bestehen. Antwort auf die Replik von Martin Brusis zu den Sustainable Governance Indicators, in: *Zeitschrift für Politikwissenschaft*, 1, 79 - 88.

Jahn, Detlef 2006: *Einführung in die Vergleichende Politikwissenschaft,* Wiesbaden, VS-Verlag.

Kaina, Viktoria 2008: Die Messbarkeit von Demokratiequalität als ungelöstes Theorieproblem, in: *Politische Vierteljahresschrift*, 49 (3), 518 - 524.

Kaufmann, Bruno und Waters, M. Dane (Hg.) 2004: *Direct Democracy in Europe. A Comprehensive Reference Guide to the Initiative and Referendum Process in Europe*, Durham.
Lauth, Hans-Joachim 2004: *Demokratie und Demokratiemessung. Eine Konzeptionelle Grundlegung für den Interkulturellen Vergleich,* Wiesbaden, VS-Verlag.
Lauth, Hans-Joachim 2008: Demokratieentwicklung und demokratische Qualität, in: Gabriel, Oscar W. und Kropp, Sabine (Hg.), *Die EU-Staaten im Vergleich: Strukturen, Prozesse, Politikinhalte*, Wiesbaden, VS-Verlag.
Lauth, Hans-Joachim 2010: Möglichkeiten und Grenzen der Demokratiemessung, in: *ZSE*, (4), 498-529.
Lauth, Hans-Joachim, Pickel, Gert und Welzel, Christian (Hg.) 2000: *Demokratiemessung*, Westdeutscher Verlag.
Laver, Michael und Hunt, W. Ben 1992: *Policy and Party Competition,* New York, Routledge.
Merkel, Wolfgang, Puhle, Hans-Jürgen und Croissant, Aurel (Hg.) 2003: *Defekte Demokratien. Bd.1, Theorien und Probleme*, Wiebaden, VS-Verlag für Sozialwissenschaften.
Merkel, Wolfgang, Puhle, Hans-Jürgen, Croissant, Aurel und Thiery, Peter (Hg.) 2006: *Defekte Demokratie. Bd. 2, Regionalanalysen*, Wiesbaden, VS-Verlag für Sozialwissenschaften

Müller, Harald 2002: Antinomien des demokratischen Friedens, in: *Politische Vierteljahresschrift*, 43 (1), 46 - 81.
Müller, Thomas und Pickel, Susanne 2007: Wie lässt sich Demokratie am besten messen? Zur Konzeptqualität von Demokratie-Indizes, in: *Politische Vierteljahresschrift* 48 (3), 511 - 539.
Munck, Gerardo L. und Verkuilen, Jay 2002: Conceptualizing and Measuring Demoracy. Evaluating Alternative Indices, in: *Comparative Political Studies*, 35 (1), 5 - 34.
Obinger, Herbert 2001: Demokratie und Wirtschaftswachstum. Theoretische Ansätze und empirische Befunde des quantitativen internationalen Vergleichs, in: *Zeitschrift für Internationale Beziehungen* 8(2), 321 - 344.
Pickel, Gert und Pickel, Susanne 2006: *Demokratisierung im internationalen Vergleich. Neue Erkenntnisse und Perspektiven,* Wiesbaden, VS-Verlag für Sozialwissenschaften.
Przeworsky, Adam und Teune, Henry 1970: *The Logic of Comparative Social Inquiry,* New york, Wiley.
Stoiber, Michael 2007: Eine neues, kontextualisierten Maß für Demokratie. Konzeptualisierung und Operationalisierung. https://www.dvpw.de/fileadmin/user_upload/sek_vp/Vortrag%20Stoiber%2011_11_07.pdf.
Stoiber, Michael 2011: *Die Qualität von Demokratien im Vergleich: Zur Bedeutung des Kontextes in der empirisch vergleichenden Demokratietheorie,* Baden-Baden, Nomos.
Sunde, Uwe 2006: Wirtschaftliche Entwicklung und Demokratie - Ist Demokratie ein Wohlstandsmotor oder ein Wohlstandsprodukt?, in: *Perspektiven der Wirtschafspolitik*, 7 (4), 471 - 499.