



Measuring Political Deliberation: A Discourse Quality Index

Marco R. Steenbergen^a, André Bächtiger^b, Markus Spörndli^b and Jürg Steiner^{a,b}

^aDepartment of Political Science, University of North Carolina, Chapel Hill, NC 27599-3265, USA.
E-mails: Marco_Steenbergen@unc.edu, jsteiner@email.unc.edu

^bInstitut für Politikwissenschaft, Universität Bern, Lerchenweg 36, CH-3000 Bern, Switzerland.
E-mails: baechtiger@ipw.unibe.ch, spoerndli@ipw.unibe.ch

In this paper, we develop a discourse quality index (DQI) that serves as a quantitative measure of discourse in deliberation. The DQI is rooted in Habermas' discourse ethics and provides an accurate representation of the most important principles underlying deliberation. At the same time, the DQI can be shown to be a reliable measurement instrument due to its focus on observable behavior and its detailed coding instructions. We illustrate the DQI for a parliamentary debate in the British House of Commons. We show that the DQI yields reliable data and we discuss how these data could be used in subsequent analysis. We conclude by discussing some limitations of the DQI and by identifying some areas in which it could prove useful.

Comparative European Politics (2003) 1, 21–48. doi:10.1057/palgrave.cep.6110002

Keywords: deliberation; discourse; measurement; parliamentary debate

Introduction

Over the past decade, deliberative politics has moved to the forefront of political theory.¹ Deliberation implies that political decision-making is or should be 'talk-centric' rather than 'vote-centric' (Bohman and Rehg, 1997; Chambers, 1999). Thus, decision-making is not concerned with the aggregation of pre-existing, fixed preferences. Rather, decision-making is a process in which political actors listen to each other, reasonably justify their positions, show mutual respect, and are willing to re-evaluate and eventually revise their initial preferences through a process of discourse about competing validity claims (Habermas, 1981, 1990, 1991, 1992, 1995, 1996; Chambers, 1995, 1999; Gutmann and Thompson, 1996). Deliberative theorists claim that such a process of discourse will lead to better-informed preferences (Fishkin, 1995) and will produce more legitimate decisions (Cohen, 1989).

Despite the advances in the political theory of deliberation and despite deliberative theorists' claims about the importance of discourse, empirically



focused political scientists have been slow to consider deliberation and discourse as research themes.² This is partially a reflection of the continued predominance of social choice theory and other theories that treat preferences as given and view decision-making as a process of preference aggregation. However, a larger issue may be the dearth of measurement instruments that allow researchers to operationalize and quantify the quality of discourse, and that open up deliberation for empirical research.

In this article, we develop a measurement instrument of deliberative quality — the discourse quality index (DQI). This measurement instrument has the advantage that it is theoretically grounded, finding its origins in Habermas (1981, 1990, 1991, 1992, 1995, 1996) as well as other theorists. At the same time, the DQI can be applied easily and reliably to a wide range of deliberative contexts. Thus, the DQI opens up deliberation for empirical research, allowing this research to interface with political theory.

We should point out at the outset that this is a measurement paper. We lack the space to investigate actually the impact of discourse quality on political outcomes. Although many theorists believe that deliberative politics leads to better outcomes, there is no consensus on this matter, with some arguing vociferously that an automatic connection between deliberation and just outcomes cannot be presumed (Sanders, 1997; Gabardi, 2001; Hauptmann, 2001; Young, 2001; Gutmann and Thompson, 2002). Our own view is that the linkage between deliberation and political outcomes should not be prejudged, but instead should be investigated empirically. We are currently undertaking such an investigation (Steiner *et al.*, 2001, 2004; Bächtiger *et al.*, 2002; Spörndli, 2002). The goal of this article is to describe the measurement instrument used in this research and to make it available to other scholars with an empirical interest in political deliberation.

We organize this article as follows. First, we discuss the criteria that a measure of the nature of discourse should satisfy. Next, we discuss two past efforts at measuring deliberation. Third, we discuss the theoretical foundation of the DQI. Having laid the groundwork, we then provide a detailed discussion of the DQI. This is followed by an empirical illustration that shows the coding procedures as well as the measurement properties of the DQI. We then discuss how the DQI can be employed in empirical research. We conclude by discussing some limitations of the DQI and opportunities for its use.

Measurement Criteria

A measure of the nature of discourse can serve as a bridge between political theory and empirical scholarship only if it does justice to the former and provides guidance to the latter. We believe, therefore, that such a measure



should meet four different criteria: (1) it should be theoretically grounded, (2) it should tap into observable phenomena, (3) it should be general, and (4) it should be reliable. The first criterion is essential because it concerns the validity of the discourse measure. An improperly grounded measure lacks construct validity, would be of little practical use and would fail to convince theorists. The complication here is the lack of agreement among political theorists about what constitute deliberation and discourse. One could attempt to develop a measure that captures all of the different conceptions of these concepts, but this would result in an instrument that is too complex to be of practical value and one that lacks internal consistency (since the different theories are not necessarily compatible). Our approach is different; we have selected a particular theory of deliberation, one that is most closely associated with Habermas' (1981, 1990, 1991, 1992, 1995, 1996) discourse ethics, and have used this as the foundation of the DQI. This ensures internal consistency, as well as greater simplicity of the resulting measure. Moreover, Habermas' ideas have inspired much of the interest in and debate over deliberative politics, which make them a good starting point for the development of a measurement instrument.

The second to fourth criteria speak to the empirical power of a discourse measure. Most importantly, the measurement instrument should tap into observable discourse behavior. This is essential if the measurement instrument is to produce reliable data and if it is to convince empirical scholars. In addition, an ideal measure of discourse quality should be general, so that it can be transported from one research domain to the next. Finally, such a measure should be reliable. This requires not only that it is based on observable behavior, but also that its coding instructions are specific, and that its coding categories are sufficiently clear that different coders could agree on the classification of the same discourse. Of course, reliable measurement is never guaranteed, so that empirical reliability assessments should be a standard practice in discourse analysis.

Below, we shall argue that the DQI meets these criteria. However, before outlining the logic of this measure, we should review past efforts at measuring discourse. As we shall see, these efforts leave considerable room for improvement.

Past Measurement Efforts

Attempts to measure discourse quality are scarce. The first such attempt that we are aware of is Gerhards' (1997) analysis of the discourse on abortion in two German newspapers. Gerhards identifies four components of discourse quality. First, he considers the representativeness of the actors covered in the newspaper articles. Second, he measures the degree of respect expressed towards other participants in the debate, using a five-point scale ranging from 'very positive' to 'very negative.' Third, he measures the degree of justification of



claims, which is captured through a binary indicator (justification *vs* no justification). Finally, he measures the rationality of discourse quality, which focuses on the number of values that were invoked and the integration of those values.

More recently, Holzinger (2001b) has performed a quantitative analysis of deliberative politics, focusing on a mediation process concerning the construction of an incinerator in the German district of Neuss. Holzinger distinguishes between bargaining and arguing, which she uses as a synonym for deliberation (see also Risse, 2000). Using speech act theory (Searle, 1969), she classifies individual words and groups of words as indicating bargaining or arguing (see also Holzinger, 2001a). She then counts the instances of each in the nine meetings that took place between March and August 1992, as a part of the mediation process.³

As pioneering measurements efforts of discourse, these two projects have made important contributions to political science. But how do they measure up against the criteria that we have laid out? Our first criterion states that these measures should be theoretically grounded. This is true for both measures, since both are based on Habermas' discourse ethics. However, neither measure fully captures Habermas. Gerhards' (1997) measure, for example, captures respect and justifications only crudely, while it confounds consensus building with justification and leaves out considerations of the common good altogether. As we shall see, these are important aspects of Habermas' discourse ethics. Holzinger (2001b) focuses on justification and consensus, paying less attention to the other aspects of a discourse ethics.

One can understand these limitations in part as a consequence of the empirical focus of these studies. This is especially true of Holzinger's work, which focuses on bargaining situations where quality of arguments and consensus building are indeed the proper focus. Our objective is to develop a measurement instrument that can be applied in a multitude of contexts, including those that do not entail bargaining. Given this broader focus, there is a need for a measure that captures discourse ethics in a comprehensive fashion.

From an empirical standpoint too, the two measurement approaches suffer from limitations, although both satisfy our second criterion by using observable behavior as their basis. The most important of these limitations is the lack of reliability testing, which makes it difficult to assess the quality of the measurement instruments. We believe that reliability testing is essential, and therefore it is an essential element for the development of our DQI.

Theoretical Foundations

The theoretical foundation of our measure of discourse quality is Habermas' (1981, 1990, 1991, 1992, 1995, 1996) discourse ethics.⁴ The starting point of



discourse ethics is the principle of universalism, which holds that a norm is valid only if everyone who is potentially affected by the norm accepts its consequences, including any anticipated negative side effects. The acceptance of norms cannot be imposed in an authoritarian manner. Rather, individuals ought to consent to those norms, and this is done through a process of argumentation and persuasion. This process of discourse constitutes ‘communicative action:’ individuals give and criticize reasons for holding or rejecting particular validity claims, so that universally valid norms can be discovered through reason.

According to Habermas and other theorists, the discourse ethics should *ideally* follow several rules.⁵ First, there should be open *participation*. Every competent individual should be free to take part in the discourse. Thus, everyone should be allowed to introduce any assertion into the debate. Moreover, everyone should be able to express his/her attitudes, desires, and needs. No one should be prevented from exercising these rights due to internal or external coercion. Even the rules and procedures of the discourse should be open for discussion (Cohen, 1989; Habermas, 1992, 370–372; Chambers, 1995; Benhabib, 1996).

Second, fruitful discourse requires the *justification* of assertions and validity claims. That is, assertions should be introduced and critically assessed through ‘the orderly exchange of information and reasons between parties’ (‘den geregelten Austausch von Informationen und Gründen zwischen Parteien,’ Habermas, 1992, 370). Justifications at once can fuel and resolve disputes, thus stimulating the deliberative process. A critical aspect of these justifications is their logical coherence. From semiotics, we know that argumentation is a process in which ‘someone tries to convince someone of something by citing evidence and drawing, or suggesting, inferences from this evidence and from other beliefs and assumptions (hypotheses)’ (Sebeok, 1986, 50–51). An inference means a ‘semiotic process in which from something given (the premises), something else (the conclusion) is derived on the basis of certain relations between premises and conclusion’ (Sebeok, 1986, 51). The tighter the connection between premises and conclusions, the more coherent the justification is and the more useful it will be for deliberation. However, the connections between premises and conclusions do not always have to be stated explicitly, using terms such as ‘since,’ ‘for,’ ‘so,’ ‘therefore,’ and ‘because.’ Indeed, ‘economies of speech’ may cause individuals to leave out entire parts of arguments, since they may be so obvious that it is unnecessary to state them (Angell, 1964).

Third, the participants in the discourse should consider the *common good*. That is, there should be a sense of empathy, other-directedness, or solidarity that allows the participants to consider the well-being of others and of the community at large. This does not mean that self-interest should be excluded



from argumentation. However, someone using self-interest must demonstrate that it is compatible with or contributes to the common good. Appeals to the common good can take several forms. On the one hand, the common good may be stated in utilitarian terms, that is, as the best solution for the greatest number of people (Mill, 1998). However, we believe that the common good may also be expressed through the difference principle: the common good is served if the least advantaged in a society are helped (Rawls, 1971).

Fourth, the participants in a discourse should treat each other with *respect* (Müller, 1995; Gutmann and Thompson, 1996; Macedo, 1999). Respect is a prerequisite for serious listening, which in turn is essential for deliberation. Several dimensions of respect play a role in discourse. One of these dimensions is respect toward *groups*, which is a reflection of Habermas' emphasis on empathy and solidarity. Respect in this sense implies that participants, either implicitly or explicitly, acknowledge the needs and rights of different social groups. Another dimension is respect for the *demands* under discussion, at least as long as they can intersubjectively be seen as justified. A third dimension is respect toward *counterarguments*, that is, arguments raised by opponents that contradict one's own conclusion with regard to the demand (Gutmann and Thompson, 1996). These last two dimensions pertain to the treatment of other participants in the debate and are especially important for deliberation. In particular, respect toward counterarguments is a necessary condition for the weighting of alternatives, which some view as an essential element of deliberation (Chambers, 1999; Luskin and Fishkin, 2002).⁶

Fifth, in the words of Cohen (1989, 23), 'ideal deliberation aims to arrive at a rationally motivated consensus.' This is what we call *constructive politics*. It should be noted that consensus is merely an aim and not an absolute necessity. In the real world of politics, consensus is often not possible. Important, however, is that the participants in a discourse should at least attempt to reach mutually acceptable compromise solutions, since this is the only way in which universalism can be attained.

Finally, Habermas' discourse ethics requires authenticity, which is the absence of deception in expressing intentions (Habermas, 1981, 149). In political terms, the stated preferences should be sincere rather than strategic so that the discourse can develop in an open and honest manner. While we acknowledge the importance of authenticity for deliberative theory, it causes the greatest difficulties from a measurement perspective. To judge if a speech act is authentic is to make a judgment about a person's true *vs* stated preferences. This is exceedingly difficult, since the true preferences are not directly observable. The speculative nature of such a judgment is bound to introduce large amounts of (possibly systematic) measurement error, and for this reason we shall not further consider authenticity. All other elements of Habermas' discourse ethics, however, find a place in our DQI.



DQI

Now that we have laid out the theoretical foundation of the DQI, it is time to discuss our measure in detail. We proceed in three steps. First, we discuss the unit of analysis, next we discuss the coding categories, and finally we discuss the index.

Unit of analysis

The unit of analysis of the DQI is a *speech*, that is, the public discourse by a particular individual delivered at a particular point in a debate. Thus, the entire discourse is broken down into smaller speech units. If an individual delivers multiple speeches, each is coded separately, even if the codes are the same as those for an earlier speech. If an individual is interrupted, then the interruption itself is also considered a speech.

For each speech, including interruptions, we distinguish between relevant and irrelevant parts, and only the relevant parts are coded. A relevant part is one that contains a *demand*, that is, a proposal on what decision should or should not be made.⁷ Irrelevant parts make no demands; these could be clarifying questions or remarks unrelated to the debate.⁸ Our emphasis on demands stems from the fact that they constitute the heart of the deliberation. That is, demands stipulate what ought to be and what ought not to be, and this normative character puts them at the center of discourse ethics.

Coding categories

If a speech contains relevant parts, then the demands are noted and the speech is coded for its discourse quality. Here we rely on seven coding categories, which closely follow the principles of Habermas' discourse ethics that we discussed earlier. Taken together, these categories reflect how well a discourse corresponds to the principles outlined by Habermas. In this sense, the DQI is a measure of discourse *quality*.

Participation

This refers to a speaker's ability to participate freely in a debate. We use two codes for participation:

- (0) *Interruption of a speaker*
- (1) *Normal participation is possible*

The first code is reserved for situations in which a speaker explicitly states that he/she is disturbed by an interruption and for situations in which the interruption occurs through a formal decision. This does not include situations in which speakers are interrupted because their speaking time is up.



Level of justification

This refers to the nature of the justification of demands. Here we judge to what extent a speech gives complete justifications for demands. The completeness of the justifications is judged in terms of the inferences that are made. There are four levels of justification:

- (0) *No justification*: A speaker only says that X should or should not be done, but no reason is given.
- (1) *Inferior justification*: Here a reason Y is given as to why X should or should not be done, but no linkage is made between X and Y — the inference is incomplete. This code also applies if a conclusion is merely supported with illustrations.
- (2) *Qualified justification*: A linkage is made as to why one should expect that X contributes to or detracts from Y. A single such complete inference already qualifies for code 2.⁹
- (3) *Sophisticated justification*: Here at least two complete justifications are given, either two complete justifications for the same demand or complete justifications for two different demands.

We should point out that the completeness of a justification does not depend on whether it is explicit. Implicit inferences can qualify as complete inferences. However, it must be beyond a reasonable doubt for the coder that the meaning of the implicit linkage is well understood by all the participants in the debate.

Content of justifications

This coding category captures whether appeals are made in terms of narrow group interests, in terms of the common *good*, or in terms of both. We employ four codes:

- (0) *Explicit statement concerning group interests*: If one or more groups or constituencies are mentioned in a speech, then a code of 0 is assigned.
- (1) *Neutral statement*: There are no explicit references to constituency/group interests or to the common good.
- (2a) *Explicit statement of the common good in utilitarian terms*: There is an explicit mention of the common good and this is conceived in utilitarian terms, that is, with reference to the ‘greatest good for the greatest number’ (Mill, 1998).
- (2b) *Explicit statement of the common good in terms of the difference principle*: There is an explicit mention of the common good and this is conceived in terms of the difference principle, that is, with reference to helping the least advantaged in a society (Rawls, 1971).



Note that codes (0), (2a), and (2b) are not mutually exclusive. In many cases, one will find references to group interests as well as the common good, and a speech is coded for all of these. The balance in such appeals is often of particular interest, since it suggests the relative emphasis that is placed on the common good *vs* more narrowly defined interests.

Respect

The DQI contains three indicators of respect. First, there is respect for the *groups* that are to be helped through particular policies. Here we use three different codes:

- (0) *No respect*: This code is reserved for speeches in which there are only negative statements about the groups.
- (1) *Implicit respect*: We use this code if there are no explicitly negative statements, but neither are there explicit positive statements.
- (2) *Explicit respect*: This code is assigned if there is at least one explicitly positive statement about the groups, regardless of the presence of negative statements.¹⁰

The next indicator is respect toward the *demands* of others. This indicator uses the same codes as the group respect indicator. However, respect toward demands is not always coded. This occurs when there is only one demand on the agenda and the speaker supports it. In this case, we assume that the speaker respects the demand and we do not explicitly code respect.

Our final indicator of respect concerns *counterarguments*. This type of respect is coded only if there are counterarguments on the table or if a speaker anticipates such arguments. If there are multiple counterarguments, then the indicator serves as a summary judgment of the respect toward all these arguments. We employ four codes to measure respect toward counterarguments:

- (0) *Counterarguments ignored*: There are counterarguments but the speaker ignores these.
- (1) *Counterarguments included but degraded*: This code applies when a speaker acknowledges a counterargument, but then explicitly degrades it by making a negative statement about it or the individuals and groups that propose the argument. A single negative statement is sufficient to assign code 1, unless the speech also contains positive statements about a counterargument (in which case a code of 3 applies). If neutral statements accompany a negative statement (and there are no positive statements), a code of 1 also applies.
- (2) *Counterarguments included — neutral*: We use this code if a counterargument is acknowledged and if there are no explicit negative or positive statements about it.



- (3) *Counterarguments included and valued*: This code applies if the counter-argument is acknowledged and is explicitly valued. We assign this code even if there are also negative statements.

Constructive politics

Our final indicator concerns consensus building, or what we call constructive politics. We capture this via three codes:

- (0) *Positional politics*: Speakers sit on their positions. There is no attempt at compromise, reconciliation, or consensus building.
- (1) *Alternative proposal*: A speaker makes a mediating proposal that does not fit the current agenda but belongs to another agenda. In such cases, the proposal is really not relevant for the current debate, although it may be taken up in a different debate.
- (2) *Mediating proposal*: A speaker makes a mediating proposal that fits the current agenda.

The index

Our assumption is that the seven components of the DQI are, at least in principle, scalable. That is, we expect the coding categories to hang together reasonably well that a subset (or perhaps all) of them can be combined to form a scale that can serve as an overall measure of discourse quality. We do not require that all of the components can be combined all of the time, since much depends on the specific circumstances of the discourse. For instance, there may be no variation on one or more components, which makes it uninteresting to add them into a composite scale. Alternatively, in the context of a particular debate, one of the components may be negatively associated with the other components. Looking across a variety of contexts, however, this tendency should disappear. Methods for creating the composite will be discussed in the next section.

Empirical Example

As our illustrative example for this article, we consider a parliamentary debate in the British House of Commons. A parliament is a particularly important deliberative institution, a ‘Congress of Opinions,’ as Mill called it,

where every person in the country may count upon finding somebody who speaks his mind as well or better than he could speak it himself — not to friends and partisans exclusively, but in the face of opponents, to be tested by adverse controversy; where those whose opinion is over-ruled feel satisfied that it is heard, and set aside not by a mere act of will, but for what are thought superior reasons (Mill, 1991, 116).



We now show how the DQI may be used to code such parliamentary deliberative activity. Using the debate in the House of Commons, we illustrate the coding categories of the DQI, show that high intercoder reliability may be achieved in applying those categories, and describe the resulting index.

Description

On 27 February 1998, the House of Commons held a plenary debate on women's issues, under the official title 'Women (Government Priorities).' The objective of the debate was to discuss ways in which to improve the lives of British women. Women's issues had been one of Labour's main themes during the 1997 election campaign and had become an important priority in the Labour government of Tony Blair after the election. The 27 February 1998 debate was one of the first occasions for the government to discuss its ideas and for the Conservatives — in the opposition for the first time since 1979 — to react. The Secretary of State for Social Security and the Minister for Women, Harriet Harman, started the debate, which lasted approximately 5 h.¹¹

Coding procedure

There were two coders for this particular debate: Jürg Steiner (coder 1) and Marco Steenbergen (coder 2). Coding proceeded in two steps. First, each coder read through the debate individually and coded the relevant speeches. These independent codings will serve as the basis for the reliability statistics reported in this paper. Next, the coders came together to compare codes. In cases where there was a disagreement, the coders read through the speech again and discussed the merits of the rival codes. At the end of this deliberative process, they settled on a particular code, but not until each coder had been convinced of the accuracy of that code.

In the debate on women's issues, the two coders identified a total of 56 relevant speeches ($N = 56$), that is, speeches that made a demand. There was no disagreement between the coders on the identification of these speeches. In addition to coding those speeches, the coders also wrote down comments that justified the codes. These comments were also used to resolve disagreements between the coders.

Illustrations of coding categories

Participation

No abnormal interruptions of the speakers occurred during the debate. This is not to say there were no interruptions, only that none of the speakers



complained about them. Thus, all 56 speeches received code 1 — normal participation is possible.¹²

Level of justification

The debate contained instances of each of the categories of our level of justification indicator. The lowest level of this indicator is 0 — no justification. An example is the following demand made by Cheryl Gillan (Conservatives, Chesham Amersham):

I am pleased that the hon. Lady [Julie Morgan, Labour, Cardiff North] praises the work of Chwarae Teg, on which the fair play for women exercise was built by the previous Conservative Government. Does she share my hope that the Government will continue to support fair play for women in Wales and the rest of the country? We want a firm commitment from the Minister to back that (Col. 666).¹³

Gillan demands the continued support for fair play, but she does not justify why this is desirable. Since we doubt the rationale would have been obvious to other MPs, we coded the statement as a 0 on the level of justification.

An example of inferior justification can be found in a statement made by Jacqui Lait (Conservatives, Beckenham):

Does my hon. Friend [Eleanor Laing, Conservatives, Epping Forest] agree that, if the rumours are true that people will not need receipts to claim the child care allowance, they could indeed spend the money on washing machines? (Col. 660)

The implicit demand here is that people should have receipts in order to claim childcare allowance. However, the justification is incomplete. The suggestion is that without receipts, people will spend their allowances frivolously but this is not backed up by an argument or evidence. Thus, it is doubtful that this argument could serve as the foundation of serious deliberation, as would be required by discourse ethics.

An example of a qualified justification comes from David Rendel (Liberal Democrats, Newbury):

Does the hon. Lady [Caroline Spelman, Conservatives, Meriden] agree that there is a further point on the separate taxation of men and women? Women who are abused in the household sometimes find it difficult to get away from the home. Separate taxation helps women to have the courage to move out on an abusive household (Cols. 624–625).

The demand is that there should be separate taxation of men and women. A complete justification is provided: such a policy would give women the courage to escape from an abusive household.



Finally, let us consider an example of a sophisticated justification. For this, we turn to Beverly Hughes (Labour, Stretford Urmston). Hughes demands family-friendly employment policies, which should extend to fathers. She goes on to provide three complete justifications for this demand:

There are several reasons why that is important. First, some parents want to look after their own children. Secondly, it is not so bad looking after one child, but when there are two or three children, the difficulty of getting them to different arrangements at different times and juggling child care with work increases exponentially.

Thirdly, and I believe this to be the most important reason, for many children there are advantages in experiencing daily daytime care from their father. Although in practice it is a women's issue, we must try to redefine the issue of who cares for children as an issue for men and women. In demanding an end to the segregation between work and mothering, we must extend the argument to ending the segregation between work and fathering (Col. 652).

Content of justification

There were no instances of neutral statements. Instead all of the speeches made explicit reference to group or constituency interests. In addition, quite a few speeches also appealed to the common good, stated either in utilitarian terms or in terms of the difference principle.

As one would expect in a debate of this sort, most of the group interest references pertained to women in general or particular groups of women (e.g., poor women or abused women). However, there were some notable exceptions. For instance, Conservative MP Damian Green (Ashford) tries to focus the debate on both men and women, arguing that

if they [the Labour government] are not careful, the old-fashioned male chauvinism, which is that some areas of life and public policy are of no interest to women, with a new form of chauvinism — equally regrettable and reactionary — which is that some areas of life are specially reserved for women (Col. 654).

We considered a speech an appeal to the common good if it satisfied certain conditions. To count as an appeal to the common good in utilitarian terms, we looked for explicit references to such terms as 'the good for the county,' 'the best for society,' 'best for the most people,' or any other macrolevel statement about benefits or costs. To count as an appeal to the common good stated in terms of the difference principle, we focused on explicit references to the most disadvantaged. These references did not have to say 'most disadvantaged' or something similar; if it was clear that a particular group is among the least advantaged (e.g., poor women), then a reference to that group would be sufficient.



As an example, consider the speech made by Jackie Ballard (Liberal Democrats, Taunton). She raises questions about the Labour government's welfare reform policies, in particular, the decision to cut the single-parent benefit. Ballard begins by focusing on the least advantaged of society. Citing the situation of a constituent on income support who was negatively affected by the cut, she makes a broader statement:

I hope that the Government will consider back-to-work benefits for *the many people in that situation*. Those of us who have been single parents on income support know that people do not have savings to fall back on in such situations (Col. 629, *italics* are ours).

This focus on the least advantaged qualifies as an appeal to the common good stated in terms of the difference principle. Ballard then focuses on the implications of welfare reform for society as a whole:

Any welfare system with paid work as its primary goal has serious implications for women and for society. I do not believe that it is in our best interests as a society to force carers out to work (Col. 629).

This focus to the best interest of society qualifies as an appeal to the common good stated in utilitarian terms.

Respect toward groups

In this debate there were no explicitly disrespectful statements about groups.¹⁴ Most of the statements showed implicit respect, although a sizable number was explicit in their respect. Consider, for example, the following speech delivered by Lorna Fitzsimons (Labour, Rochdale):

I warmly welcome the announcements made by my right hon. Friend the Secretary of State in the first momentous debate under our Government celebrating the achievements of women, both inside and outside the House (Col. 619) [...] I pay tribute to the women's organisations, the unsung heroes of our communities; we should be better off if we listened to them (Col. 622).

Clearly, this speech contains a great deal of explicit respect for women and women's organizations and should be coded a 2 on our 'respect toward groups' indicator.

Respect toward demands

Respect toward the demands of other speakers in the debate spanned the entire range of our indicator. First, let us consider some examples of disrespectful statements. Early in the debate, Labour MP Margaret Hodge (Barking) complained that the previous Conservative government had a man



representing women's issues. The demand implicit in the complaint was that female politicians should be responsible for women's issues. This demand elicited the following negative reaction of Bernard Jenkin (Conservatives, North Essex):

Would we not be going backwards if, while we were trying to abolish woman-free zones in public life, we started to create man-free zones? (Col. 617)

The sarcasm of Jenkin's remark is a clear example of disrespect toward a demand. Such disrespect can also manifest itself through negative statements about the person or groups making the demand. Consider, for example, a speech made by Debra Shipley (Labour, Stourbridge), which contains the following statements:

How far out of touch the Opposition are with the 40-nation Council [of Europe] was clearly demonstrated when the British Conservative delegates blocked the establishment of a full standing committee on equality [...] The Tories are as out of touch in Europe as they are in Britain (Col. 626).

Some of the speeches showed explicit respect for the demands of others. Consider, for example, a speech made by Robert Syms (Conservatives, Poole):

All issues affect women, and it is important that we tackle those issues across Government. The previous Government recognised that fact by ensuring that a range of Ministers understood the needs of women, and I believe that this Government are doing the same. I think that that is a *positive step forward* (Col. 668, *italics* are ours).

This statement received the highest code for our 'respect toward demands' indicators.

Not all speeches were coded for respect toward demands. If there was obvious agreement with a demand, then a speech received the code 'inapplicable.' Speeches that did not fall into this category, or in the explicitly respectful or disrespectful categories, received a code of 1 — 'implicit respect.'

Respect toward counterarguments

This dimension of respect was coded only if a counterargument was on the table or if a speaker anticipated one. Thus, the first coding decision was to identify the presence of a counterargument. The second decision was to determine if a speaker was ignoring a counterargument. An example of this occurred in the speech by Jane Griffiths (Labour, Reading East). She describes how a whole generation of women has grown up under Conservative governments, 'who treated whole groups in society with contempt, effectively excluding them' (Col. 657). This evokes an interruption by Robert Syms



(Conservatives, Poole) and a response by Griffiths:

Syms: Does the hon. Lady not consider that the position of women in society today is better than it ever has been, and is improving? It has certainly improved over the past 20 years, especially under Conservative Governments.

Griffiths: If the hon. Gentleman believes that there is reason for complacency about the position of women, I do not agree with him (Col. 657).

While it may appear that Griffiths responds to the counterargument, she does not actually address its claim, stating only that one should not be complacent.

Once it is established that a counterargument is acknowledged, the next step is to code for the tone of the reaction to the argument. In this debate, many of the reactions were negative. Consider, for example, the following exchange between Michael Fabricant (Conservatives, Lichfield) and Barbara Follett (Labour, Stevenage):

Fabricant: The point I was making is that I support — as I believe we all support — equal opportunities. What we do not support is positive discrimination and quotas, which tend to lower standards.

Follett: Positive discrimination is illegal in this country. Positive action is legal, and quotas come under positive action. I would point out to the hon. Gentleman, whom I thank for his courtesy in giving way, that positive discrimination and positive action have acted in favour of men for centuries, and I agree with him that it has lowered standards.

Fabricant: The hon. Lady is absolutely right. She knows everything about positive discrimination being illegal, as her own party was ruled illegal in positively discriminating for women when it tried to introduce women-only shortlists for Parliamentary candidates (Col. 634).

The counterargument of Follett, which is quite sarcastic in its own right, is followed up by a very negative reaction by Fabricant, thus clearly calling for a code of 1 — ‘counterarguments included but explicitly degraded.’¹⁵

Not all speeches are so negative. Consider, for example, the speech by Teresa Gorman (Billericay), a fellow Conservative of Fabricant’s. When Gorman criticizes the House of Commons because it ‘often gives the impression that young women are an underclass or an underdog group’ (Col. 642), Laura Moffatt (Labour, Crawly) interrupts:

Does the hon. Lady agree that there has been no suggestion of painting women as victims, as the debate has been wide-ranging — about all sorts of women and the contributions that they make? Does she also agree that many of the women she met were of postgraduate age and that it is a little later, when one has to face child care issues and all that those bring to bear, that women start to run into trouble? (Col. 642).



Gorman acknowledges Moffatt's argument in a neutral manner, simply by stating 'I hear what the hon. Lady says' (Col. 642). Thus, Gorman's speech was coded 2 — 'counterarguments included — neutral.'

The opening speech by Secretary Harriet Harman (Labour) serves as an example of a positive appraisal of counterarguments. After Conservative MP Jacqui Lait (Beckenham) raises a concern about the equal funding of pensions for men and women, which serves as a counterargument to the Secretary's proposal of a second pension to be added to the state pension, Harman responds:

The hon. Lady raises an important issue, about which the pensions review is liaising with actuaries (Col. 608).

This speech received a code of 3 — 'counterarguments included and valued.'

Constructive politics

This debate showed no variation on our constructive politics indicator. All the speakers stood by their initial positions. Even at the end of the debate, there were no attempts at reconciliation. Thus, all speeches received a code of 0 — 'positional politics.'¹⁶

Reliability of the DQI indicators

The above examples show how a discursive text, such as the debate on women's issues, can be quantified using the indicators of the DQI. However, there remains an important question: can two different coders, both familiar with the context of the debate, agree on the codes that should be given to the text? Put differently, is the DQI a reliable measurement instrument?

The reliability of the DQI hinges on two types of judgment. First, do coders agree in their judgment that a particular indicator is applicable? Second, if the indicator is deemed applicable, do coders agree on the code that a speech or other discursive text should receive? Agreement on both of these judgments is essential for obtaining a reliable measure.

The following analysis shows that the DQI is indeed a reliable measure. Using four different indicators of intercoder agreement — the ratio of coding agreement (Holsti, 1969), Cohen's (1960) κ ('kappa') and, where appropriate, Spearman's rank correlation (Siegel, 1956) and the standardized α (a reliability statistic, see Holsti, 1969) — we demonstrate that the DQI has outstanding measurement properties for the debate on women's issues.¹⁷ We consider both the overall performance of the DQI and its performance for each of the coding categories.

Overall agreement

In total, the two coders rendered 504 judgments for the debate.¹⁸ They agreed on 461 of these judgments. This included agreements on specific codes, as well



Table 1 Reliability scores by coding category

<i>Category</i>	<i>RCA</i>	κ	<i>s.e.</i>	<i>r</i>	α
Participation	1.000				
Level of justification	0.732	0.615**	0.085	0.716**	0.834
1st content of justification	1.000				
2nd content of justification	0.875	0.775**	0.080		
3rd content of justification	0.964	0.837**	0.113		
Respect toward groups	0.875	0.746**	0.090	0.747**	0.855
Respect toward demands	0.893	0.844**	0.060	0.855**	0.922
Respect toward counterarguments	0.893	0.559**	0.170	0.791	0.883
Constructive politics	1.000				

Note: $N=56$ with the following exceptions: (1) $N=29$ for the computation of r and α for 'respect toward demands'; (2) $N=5$ for the computation of r and α for 'respect toward counterarguments'. ** $p < 0.01$.

as on judgments of whether a particular indicator was applicable. Thus, $RCA=0.915$; the coders agreed 91.5% of the time, which is an excellent reliability score.

Agreement on specific categories

Table 1 shows the reliability statistics for specific coding categories. First, consider participation. Both coders agreed that normal participation was possible for all the speakers. Thus, $RCA=1.00$, a perfect score.¹⁹

Considering the level of justification, $RCA=0.732$, which is respectable. Taking into consideration that the coders may have agreed by chance alone, we also computed Cohen's κ , which equals 0.615. This is both significant and indicative of substantial agreement (Landis and Koch, 1977). Since level of justification is an ordinal indicator, it is also useful to consider Spearman's rank correlation, which takes into consideration the difference in the rank orderings of speeches between coders. For this debate, Spearman's $r=0.716$, which produces a standardized item α of 0.834. This may be considered a very good reliability.

For the content of justification, we performed three reliability analyses, since up to three codes could be given. All speeches contained references to group or constituent interests, and the coders were in perfect agreement about this; $RCA=1.00$. In addition, some speeches also contained one or more references to the common good. The coders had to judge first if this was the case, and next what kind of reference to the common good was being made (utilitarian or difference principle). For the second content of justification coding, the coders agreed 87.5% of the time ($RCA=0.875$).²⁰ Taking chance agreement into account, $\kappa=0.775$; this is statistically significant and indicates substantial



agreement. For the third content of justification coding, $RCA = 0.964$ and $\kappa = 0.837$, indicating excellent reliability.²¹

The reliability of the indicators for respect was also excellent. The coders agreed 87.5% of the time ($RCA = 0.875$) in their judgment of a speech's respect toward groups. Further, $\kappa = 0.746$, which is again significant and indicative of substantial agreement. Since the indicator for respect toward groups is ordinal, we also computed Spearman's rank correlation: $r = 0.747$, which results in an impressive α of 0.855.

The results for respect toward demands are even better. Here $RCA = 0.893$ and $\kappa = 0.844$, which is significant and indicative of near-perfect agreement (Landis and Koch, 1977).²² The 'respect toward demands' category was deemed applicable by both coders for 29 speeches. For those speeches, Spearman's r is an impressive 0.855, which implies $\alpha = 0.922$.

The last dimension of respect concerns counterarguments. Here $RCA = 0.893$, which again reflects excellent agreement.²³ Since this level of agreement does not differ much from that expected by chance, κ is only 0.559. While this is the lowest reliability in Table 1, it still corresponds to 'moderate agreement' (Landis and Koch, 1977). Since counterarguments were rare in the debate, there are only five cases in which both coders deemed this indicator applicable. Focusing on those cases, the rank correlation between the codes was 0.791, producing $\alpha = 0.883$, which is again outstanding.

Our final coding category is constructive politics. Here both coders agreed that all speeches reflected positional politics. Hence, $RCA = 1.00$, which is perfect. These results indicate that reliable measurement of discourse through the DQI is possible. Even the worst reliability scores are still respectable, suggesting that different coders looking at the same discursive text will be able to agree on the DQI and its components. This is an important result, as it greatly increases the confidence one can place in the DQI.

Creating the index

So far, our discussion of the DQI has focused on its components, but not on the index itself. How would one construct this index? The foundation should be that the components form a coherent set, as measured by their correlations. This is a necessary condition for unidimensionality, that is, the requirement that a set of indicators measure one and only one thing (Steenbergen, 2000).

Before embarking on a check of the correlations between the DQI components for the debate on women's issues, we need to take two preparatory steps. First, we exclude components that showed no variation in the debate (i.e., participation, constructive politics, and appeals to group interests), since these components do not contribute to the index (except as constant terms). Second, we created a dummy variable measuring if there had been an appeal to



Table 2 Three-DQI component correlation matrix

	<i>L</i>	<i>CG</i>	<i>G</i>
Level of justification (<i>L</i>)	1.000		
Common good (<i>CG</i>)	0.821	1.000	
Respect toward groups (<i>G</i>)	0.781	0.781	1.000

Note: Table entries are polychoric correlation coefficients. $N = 56$.

Table 3 Five-DQI component correlation matrix

	<i>L</i>	<i>CG</i>	<i>G</i>	<i>C</i>	<i>D</i>
Level of justification (<i>L</i>)	1.000				
Common good (<i>CG</i>)	0.983	1.000			
Respect toward groups (<i>G</i>)	0.994	0.997	1.000		
Respect toward counterarguments (<i>C</i>)	1.000	0.983	0.995	1.000	
Respect toward demands (<i>D</i>)	-0.428	-0.584	-0.832	-0.503	1.000

Note: Table entries are polychoric correlation coefficients. $N = 9$.

the common good (stated either in utilitarian terms or in terms of the difference principle). This variable is based on the second and third content of justification measures; we refer to it as the *common good* indicator.

We are now in a position to compute the correlations between the DQI components. Table 2 shows these correlations for the three components — level of justification, respect toward groups, and common good — that were measured for all speeches. To take into account the categorical nature of these components, we report polychoric correlations. As can be seen, the correlations between these three components are impressive: the average correlation is 0.794, which produces a reliability (standardized α) of 0.920.²⁴ These results suggest that level of justification, common good, and respect toward groups can be combined into an additive index.

We can extend the correlational analysis by also including respect toward demands and respect toward counterarguments. Since these attributes were not coded for all speeches, the sample size dwindles to a mere nine speeches. Table 3 shows the polychoric correlation between level of justification, common good, and the three respect indicators. We observe very high positive correlations between all these indicators, except for respect toward demands, which is negatively correlated with the other components. Thus, while the average correlation between the remaining indicators is 0.992, producing an outstanding α of 0.998, respect toward demands does not appear to fit into the index.²⁵



Table 4 DQI descriptive statistics

<i>Index</i>	<i>Mean</i>	<i>Median</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>
3-component DQI (<i>N</i> = 56)	3.982	4.000	1.700	1	6
4-component DQI (<i>N</i> = 10)	6.600	7.000	1.647	3	9

We can only speculate on the reasons why respect toward demands does not scale well with the other components of the DQI for this debate. Perhaps this is an outgrowth of the competitive British political system; had we performed our analysis in a consociational system, we might have found something different. Perhaps this finding indicates something about the nature of the times. In the aftermath of an election that returned the Conservatives to the opposition benches for the first time since 1979, both parties may have attempted to define their differences as strongly as possible, and this may have decreased the overall respect toward each other's demands. However, the important point is that the DQI could help settle these questions if we were to collect longitudinal and cross-national data.

Having demonstrated that at least a subset of the DQI indicators hang together, we can now construct the index. The simplest way in which this can be done is to simply add the indicators. If one would want to, one could also create factor scores to form the scale (Gorsuch, 1983), although this makes little difference in the present case because the factor loadings for the three items are almost identical.²⁶ Table 4 shows the descriptive statistics for the three-component DQI, consisting of level of justification, common good, and respect toward groups. For comparative purposes, the table also shows the descriptive statistics for the four-component index, which includes respect toward counterarguments along with the other indicators. As this table shows, the observed DQI scores cover the entire range; although they tend toward the higher end of the scale, there is considerable variation across speeches. One could attempt to explain this variation, or use it to explain policy outputs, as we shall discuss in the next section.

Discussion

Our empirical illustration has shown that reliable measurement of the DQI components is possible. Furthermore, at least a subset of these components hangs together, which allows us to turn them into an index. One should keep in mind, of course, that these findings only pertain to the House of Commons debate on women's issues. Future research will have to demonstrate the measurement properties of the DQI in other contexts. However, the present



results are encouraging and suggest not only that empirical measurement via the DQI is feasible, but also that the resulting measures can be of very high quality.

Using the DQI

Now that we have seen that the DQI can be used to measure discourse quality, of what use will this be? That is, how would researchers integrate the DQI into a program of empirical research on individual and institutional decision-making? We envision two roles for the DQI — as a predictor and as a dependent variable.

Elsewhere, we have described how the DQI can be used to predict substantive policy outputs. Here we distinguished between a formal dimension of decisions, that is, the presence of a genuine rational consensus, and a substantive dimension, that is, the presence of distributive justice. We are presently involved in an extensive research program exploring the implications, if any, of discourse quality for these dimensions (Steiner *et al.*, 2001; Spörndli, 2002).

We can also use the DQI as a dependent variable, trying to explain why the quality of discourse varies. Focusing on favorable (and unfavorable) institutional and other contextual conditions, we are currently investigating how the nature of the political system (consensual *vs* competitive), the type of arena (closed *vs* open), and the type of issue influence the DQI score of a debate (Steiner *et al.*, 2001; Bächtiger *et al.*, 2002). In addition, political psychologists might be interested in individual differences in the DQI, much like they are presently interested in variation in integrative complexity (Levi and Tetlock, 1980). At this level, ideological preferences, dogmatism, and the level of stress experienced in the decision-making process may all play a role.

We envision that research based on the DQI will traverse multiple units of analysis. At the lowest level, one can analyze a particular speech. One can ask how the DQI score of that speech is related to the previous speech, and how it affects the next one. At a higher level, one can aggregate DQI scores at the level of speakers and then correlate this with attributes of those speakers.²⁷ At yet a higher level of analysis, one could consider the mean DQI scores for political parties and other groups. One could then look at aggregate DQI scores for an entire debate. This could be followed by aggregate analyses of DQI scores across a number of debates over time in the same country. Finally, one could consider cross-national differences in DQI scores. Of course, in moving up on this hierarchy it becomes ever more important to assess the measurement properties of the DQI and its comparability across units. However, in principle, the DQI can be used for micro-, meso-, and macrolevel political analyses, making it a flexible measurement instrument.



Conclusion

In this paper, we have presented an instrument for measuring the quality of deliberation — the discourse quality index or DQI. This instrument satisfies the measurement criteria that we outlined. First, the instrument is rooted in political theory, capturing Habermas' discourse ethics more completely than previous measures (Gerhards, 1997; Holzinger, 2001b). Second, this measure is rooted in observable behavior such as that reflected in parliamentary debates. Third, the measure is general, in that it can be used in a variety of contexts — a point we shall elaborate below. Finally, the measure is reliable, as our empirical illustration has demonstrated. In conclusion, the DQI is the kind of measurement instrument that could help bridge the gap between political theory and empirical research.

This is not to say that the DQI is without limitations. One limitation arises within the Habermasian theoretical framework. Habermas has placed considerable emphasis on the authenticity of claims, an aspect of discourse ethics that the DQI ignores completely because of its unobservable nature. However, we believe this omission does not detract from our claim that the DQI is a very good fit to discourse ethics.

Another limitation of the DQI arises because it is located within the Habermasian framework. Not all theorists accept this framework as the proper definition of deliberation. For instance, Basu (1999) (see also Gabardi, 2001) criticizes Habermas for leaving out humor from his conception of deliberation. Whereas Habermas appears to view humor as a vice, Basu views it as a virtue, since it may contribute to openness in a debate. Clearly, the DQI does not contain any coding categories for the humor in a speech.

Others criticize Habermas for being excessively procedural in his focus. As a result, there is insufficient attention for the substance of the arguments that are made (Gutmann and Thompson, 2002). The DQI does not incorporate these elements into its coding.

A third limitation of the DQI is that it is limited to discursive texts. Nonverbal communication is not coded as part of the DQI, nor is the tone of voice in which speeches are delivered. These omissions are of some potential importance, since psychological research shows that nonverbal cues influence the interpretation of messages (Walthier and D'Addario, 2001).

A final limitation concerns the application of the DQI to parliamentary debates. This application domain has some peculiar features, especially in terms of participation. Members of parliament face rather mild restrictions on participation, the worst of which is probably being cut short in a debate. Clearly, access to participation can be a considerably greater hurdle in other deliberative arenas, with some citizens being precluded from participating altogether. The present DQI does not capture such



restrictions, but they should be included if the instrument is to be used in other contexts.

For the moment, we believe that these limitations are best addressed by making extensive notes for each speech. Coders should make such notes anyway in order to justify their DQI codes. With a little extra effort they could note their impression of the authenticity of a speech or their impression of the role of humor. When appropriate, similar comments can be added concerning nonverbal cues and the substance of the arguments. While these notes are not codified into the DQI, they can help researchers to understand the context and character of the debate.

When combined with such qualitative documentation, the DQI can be a powerful research tool. In this paper, we have focused on its application to formal debates. However, we can see many other applications, including newspaper articles, editorials and commentary (Parker-Stephen, 2002), televised debates (e.g., debates between political candidates), informal policy discussions such as those occurring between US Presidents and their aides, and focus groups, including national issue conventions (Fishkin, 1995; Fishkin and Luskin, 1999; Luskin and Fishkin, 2002). Wherever there is deliberation of some sort and there is a record, the DQI can be applied. Thus, the DQI is truly a general measurement instrument, one that allows empirical researchers to peer into the real world of deliberation (Steiner *et al.*, 2004).

Notes

- 1 The authors would like to thank the Swiss National Science Foundation for supporting this research.
- 2 Exceptions are the qualitative studies of discourse by Chambers (1999) and Risse (2000), and the quantitative studies by Gerhards (1997) and Holzinger (2001a, b). We should also mention the research program on deliberative polling by Fishkin and his colleagues, which focuses on the consequences of deliberation for mass public opinion (Fishkin, 1995; Fishkin and Luskin, 1999; Luskin and Fishkin, 2002).
- 3 Luskin and Fishkin (2002) discuss a series of deliberative aspects that are similar to those discussed by Gerhards (1997) and Holzinger (2001a, b). However, they operationalize these aspects in terms of individual-level outcomes, rather than in terms of the actual discourse.
- 4 In the following discussion, we draw from all of these works. We shall refer to specific works only if we quote directly from them.
- 5 It is important to point out that these rules describe the 'ideal speech situation.' Habermas realizes that real political debates are usually far away from the ideal type, which should therefore be seen as the end of a discourse continuum that most likely will never be fully reached. The empirical question then is how far away specific political debates are from the ideal type, and this is what the DQI intends to measure.
- 6 A fourth dimension of respect is personal respect, which may be viewed in terms of politeness. Depending on the context, it may be worthwhile to make this a part of the



measurement of discourse quality. However, one should take extreme care in doing so because codes of conduct are often institutionalized and as such do not reflect much about genuine respect (Haase, 1994).

- 7 A series of related demands constitute an issue. A discourse may consider more than one issue. In this case, each issue is coded separately.
- 8 While we do not code irrelevant parts, we make note of them, in particular if they contain friendly or unfriendly remarks toward other participants.
- 9 If a speech contains other conclusions, but these are embedded in incomplete inferences, then a code of 2 still applies.
- 10 A more precise measure would consider the balance between positive and negative statements, perhaps using a five-point scale like Gerhards (1997) developed. However, we have found that the reliability of such an indicator can be problematic, especially if there are subtle differences between the positive and negative statements.
- 11 The text of the debate can be downloaded from the on-line edition of the House of Commons Hansard (<http://www.parliament.the-stationery-office.co.uk/pa/cm199798/cmhansrd/vo980227/debindx/80227-x.htm>). We shall refer to passages of this debate via the column numbers in this electronic edition.
- 12 An illustration of a disturbing interruption can be found in the 25 June 1998 debate in the House of Commons concerning Social Welfare. John Sweeney (Scottish National Party) is interrupted several times, which clearly annoys him. Finally, he exclaims: 'I shall not take any more interventions.'
- 13 Chwarae Teg is a Welsh organization promoting equal opportunities for women in the work force. The name translates as 'fair play.'
- 14 An example of a disrespectful statement can be found in the 10 March 1997, House of Commons debate concerning 'Public Responsibility for Social Justice.' In this debate, Labour MP Brian Wilson (Cunninghame) makes the following statement about the Scots (and the Welsh):

I wonder why we should be less concerned about people in England who are homeless than about those in Scotland and Wales. A proportionate number of people are homeless in England, many of whom are Scottish or Welsh in origin. Does that upset the primary nationalist argument that England is somehow a land flowing with milk and honey at Scotland's expense?

With the rhetorical question at the end, Wilson implies that the Scots unduly play the nationalistic card to receive more aid.

- 15 In the same speech, Fabricant reacts very negatively to the demands of the Labour party, accusing it of betraying 'their own principles' (Col. 638), failure, 'the total fiasco of the changes to single parent benefit' (Col. 639), and taking 'U-turns' (Col. 639). Thus, Fabricant's speech also scores low on respect toward demands.
- 16 For an example of an alternative proposal (code 1), we turn to an American example. During the 1996 debate in the House of Representatives on an increase in the minimum wage, Representative Campbell (Republican) opposed the increase but advocated an alternative proposal consisting of an increase in the earned income tax credit. Since the tax code was not under debate, this proposal did not fit the agenda and would receive a code of 1 on the constructive politics indicator.

For a mediating proposal (code 2), consider the 28 November 1997, debate in the House of Commons concerning the Wild Mammals (Hunting with Dogs) Bill whose goal is to outlaw the hunting of mammals with wild dogs. Labour MP Michael Foster (Worcester) expresses his strong wish to 'vote to ban hunting with dogs.' Despite this strong view, he offers a mediating proposal that takes account of the grievances of sheep farmers in Wales. Thus, he proposes to



make an exemption ‘that dogs could be used to flush foxes out of cover, where they could be quickly and humanly shot.’

- 17 A complete discussion of these reliability statistics can be found in our project website: www.ipw.unibe.ch/discourse.
- 18 They made 56 coding decisions for each of the following indicators: participation, level of justification, respect toward groups, respect toward demands, respect toward counterarguments, and constructive politics. Since there were no neutral codes for the content of the justification, the coders also rendered separate judgments of whether the codes (0), (2a), and (2b) applied for this indicator. (Remember that these codes are not mutually exclusive, so that a speech could receive up to three codes for the content of justification.)
- 19 Since there was no variation in the codes, Cohen’s κ and Spearman’s r cannot be computed.
- 20 There were seven disagreements. In four cases, the first coder believed there to be an appeal to the common good while the second coder did not. In two cases, the reverse pattern occurred. Finally, in one instance the two coders disagreed as to whether the appeal was stated in utilitarian terms or in terms of the difference principle.
- 21 The only disagreement that could occur on the third coding concerned the presence of a second appeal to the common good. The way the coding was set up, if a second appeal was present it had to be in terms of the difference principle. In just two cases did the first coder believe that there was a second appeal to the common good, while the second coder believed there was not.
- 22 There were six disagreements. In one case, the first coder believed that the category was applicable, while the second coder believed it was not. In three cases, this pattern was reversed. Finally, there were two disagreements about the code for speeches for which the category was deemed applicable by both coders.
- 23 There were six disagreements. In three cases, coder 1 deemed the category applicable, while coder 2 did not. In two cases, the reverse pattern occurred. Finally, there was one instance where both coders deemed the category applicable but disagreed on the code.
- 24 Using similarity coefficients (Steenbergen, 2000), the scalewise similarity is 0.998, which is almost perfect.
- 25 This result is also borne out in a similarity analysis (Steenbergen, 2000). The scalewise similarity across all five components is a mere 0.343. If we exclude respect toward demands, then the scalewise similarity jumps to 0.992, which is close to perfect.
- 26 The loadings on a single factor are 0.736 for level of justification, 0.727 for the common good indicator, and 0.717 for respect toward groups.
- 27 For the House of Common’s debate on women’s issues, for instance, we found that the average three-component DQI scores for main speakers were considerably higher than those for other speakers: $M = 5.478$ for main speakers and $M = 2.939$ for others, $t = 8.116$, $p < 0.01$. We also found that Conservatives had the lowest average three-component DQI scores ($M = 3.483$), followed by Labour ($M = 4.440$) and the Liberal Democrats ($M = 5.500$); these differences are statistically significant: $F = 17.081$, $p < 0.05$.

References

- Angell, R.B. (1964) *Reasoning and Logic*, New York: Appleton-Century-Croft.
- Bächtiger, A., Spörndli, M. and Steiner, J. (2002) ‘The Consociational Theory and Deliberative Politics: A Conceptual Framework for a Cross-national Analysis’, in S. Brooks (ed.) *The Challenge of Cultural Pluralism*, Westport, CT: Praeger, pp. 77–94.
- Basu, S. (1999) ‘Dialogic ethics and the virtue of humor’, *Journal of Political Philosophy* 7(4): 378–403.



- Benhabib, S. (1996) 'Toward a Deliberative Model of Democratic Legitimacy', in S. Benhabib (ed.) *Democracy and Difference: Contesting the Boundaries of the Political*, Princeton, NJ: Princeton University Press, pp. 67–94.
- Bohman, J. and Rehg, W. (1997) 'Introduction', in J. Bohman and W. Rehg (eds.) *Deliberative Democracy: Essays on Reason and Politics*, Cambridge, MA: MIT Press, pp. ix–xxvii.
- Chambers, S. (1995) 'Discourse and Democratic Practices', in S.K. White (ed.) *The Cambridge Companion to Habermas*, Cambridge: Cambridge University Press, pp. 233–259.
- Chambers, S. (1999) 'Talking versus voting: legitimacy, efficiency, and deliberative democracy', Unpublished manuscript, Department of Political Science, University of Colorado, Boulder, CO.
- Cohen, J. (1960) 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement* 20(1): 37–46.
- Cohen, J. (1989) 'Deliberation and Democratic Legitimacy', in A. Hamlin and P. Pettit (eds.) *The Good Polity: Normative Analysis of the State*, Oxford: Blackwell, pp. 17–34.
- Fishkin, J.S. (1995) *The Voice of the People: Public Opinion and Democracy*, New Haven, CT: Yale University Press.
- Fishkin, J.S. and Luskin, R.C. (1999) 'Bringing Deliberation to the Democratic Dialogue', in M. McCombs and A. Reynolds (eds.) *The Poll with a Human Face: The National Issues Convention Experiment in Political Communication*, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 3–38.
- Gabardi, W. (2001) 'Contemporary models of democracy', *Polity* 33(4): 547–568.
- Gerhards, J. (1997) 'Diskursive vs liberale Öffentlichkeit: Eine empirische Auseinandersetzung mit Jürgen Habermas', *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 49(1): 1–34.
- Gorsuch, R.L. (1983) *Factor Analysis*, Hillsdale, NJ: Lawrence Erlbaum Associates.
- Gutmann, A. and Thompson, D. (1996) *Democracy and Disagreement*, Cambridge, MA: Belknap Press of Harvard University Press.
- Gutmann, A. and Thompson, D. (2002) 'Deliberative democracy beyond process', *Journal of Political Philosophy* 10(2): 153–174.
- Haase, M. (1994) *Respekt: Die Grammatikalisierung von Höflichkeit*, München: Lincom Europa.
- Habermas, J. (1981) *Theorie des kommunikativen Handelns*, Frankfurt am Main: Suhrkamp.
- Habermas, J. (1990) *Strukturwandel der Öffentlichkeit: Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft*, Frankfurt am Main: Suhrkamp.
- Habermas, J. (1991) *Erläuterungen zur Diskursethik*, Frankfurt am Main: Suhrkamp.
- Habermas, J. (1992) *Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats*, Frankfurt am Main: Suhrkamp.
- Habermas, J. (1995) 'Reconciliation through the public use of reason: remarks on John Rawls's political liberalism', *Journal of Philosophy* 92: 109–131.
- Habermas, J. (1996) *Die Einbeziehung des Anderen: Studien zur politischen Theorie*, Frankfurt am Main: Suhrkamp.
- Hauptmann, E. (2001) 'Can less be more? Leftist deliberative democrats' critique of participatory democracy', *Polity* 33(3): 397–421.
- Holsti, O.R. (1969) *Content Analysis for the Social Sciences and Humanities*, Reading, MA: Addison-Wesley.
- Holzinger, K. (2001a) 'Kommunikationsmodi und Handlungstypen in den internationalen Beziehungen: Anmerkungen zu einigen irreführenden Dichotomien', *Zeitschrift für internationale Beziehungen* 8(2): 243–286.
- Holzinger, K. (2001b) 'Verhandeln statt Argumentieren oder Verhandeln durch Argumentieren? Eine empirische Analyse auf der Basis der Sprechakttheorie', *Politische Vierteljahresschrift* 42(3): 414–446.
- Landis, J.R. and Koch, G.G. (1977) 'The measurement of observer agreement for categorical data', *Biometrics* 33(1): 159–174.



- Levi, A. and Tetlock, P.E. (1980) 'A cognitive analysis of Japan's 1941 decision for war', *Journal of Conflict Resolution* 24(2): 195–211.
- Luskin, R.C. and Fishkin, J.S. (2002) 'Deliberation and civic engagement', Workshop on Deliberative Democracy in Theory and Practice, ECPR Joint Sessions; 22–27 March 2002; Turin.
- Macedo, S. (1999) 'Introduction', in S. Macedo (ed.) *Deliberative Politics: Essays on Democracy and Disagreement*, New York: Oxford University Press, pp. 3–16.
- Mill, J.S. (1991) *Considerations on Representative Government*, Buffalo, NY: Prometheus Books.
- Mill, J.S. (1998) *Utilitarianism*, Oxford: Oxford University Press.
- Müller, H. (1995) 'Spielen hilft nicht immer: Die Grenzen des Rational-Choice-Ansatzes und der Platz der Theorie des kommunikativen Handelns in der Analyse internationaler Beziehungen', *Zeitschrift für internationale Beziehungen* 2(2): 371–391.
- Parker-Stephen, E. (2002) 'Commentary and deliberation: a discourse analysis of news commentary after 9-11-01', Unpublished manuscript, Department of Political Science, University of North Carolina, Chapel Hill, NC.
- Rawls, J. (1971) *A Theory of Justice*, Cambridge, MA: Belknap Press of Harvard University Press.
- Risse, T. (2000) 'Let's Argue! Communicative Action in World Politics', *International Organization* 54(1): 1–39.
- Sanders, L.M. (1997) 'Against deliberation', *Political Theory* 25(3): 347–375.
- Searle, J.R. (1969) *Speech Acts: An Essay in the Philosophy of Language*, London: Cambridge University Press.
- Sebeok, T.A. (ed.) (1986) *Encyclopedic Dictionary of Semiotics*, Berlin: Mouton de Gruyter.
- Siegel, S. (1956) *Nonparametric Statistics for the Behavioral Sciences*, New York: McGraw-Hill.
- Spörndli, M. (2002) 'Procedure or substance? discursive politics and outputs', Workshop on Deliberative Democracy in Theory and Practice, ECPR Joint Sessions, 22–27 March 2002; Turin.
- Steenbergen, M.R. (2000) 'Item similarity in scale analysis', *Political Analysis* 8(3): 261–283.
- Steiner, J., Bächtiger, A. and Spörndli, M. (2001) 'The real world of deliberation: favourable conditions and substantive outcomes', 1st ECPR General Conference; 6–8 September 2001; Canterbury.
- Steiner, J., Bächtiger, A., Spörndli, M. and Steenbergen, M.R. (2004) *Deliberative Politics in Action: A Cross-National Study of Parliamentary Debates*, Cambridge: Cambridge University Press.
- Walthier, J.B. and D'Addario, K.P. (2001) 'The impacts of emoticons on message interpretation in computer-mediated communication', *Social Science Computer Review* 19(3): 324–347.
- Young, I.M. (2001) 'Activist challenges to deliberative democracy', *Political Theory* 29(5): 670–690.