European Journal of Political Research ●●: ●●–●●, 2016
doi: 10.1111/1475-6765.12162

What drives the polarisation and moderation of opinions? Evidence from a Finnish citizen deliberation experiment on immigration

MARINA LINDELL,[1] ANDRÉ BÄCHTIGER,[2] KIMMO GRÖNLUND,[1] KAISA HERNE,[3] MAIJA SETÄLÄ[4] & DOMINIK WYSS[2]

[1]*Social Science Research Institute, Åbo Akademi University, Finland;* [2]*Institute of Social Sciences, University of Stuttgart, Germany;* [3]*School of Management, Politics, University of Tampere, Finland;* [4]*Department of Philosophy, Contemporary History and Political Science, University of Turku, Finland*

Abstract. In the study of deliberation, a largely under-explored area is why some participants polarise their opinion after deliberation and why others moderate them. Opinion polarisation is usually considered a suspicious outcome of deliberation, while moderation is seen as a desirable one. This article takes issue with this view. Results from a Finnish deliberative experiment on immigration show that polarisers and moderators were not different in socioeconomic, cognitive or affective profiles. Moreover, both polarisation and moderation can entail deliberatively desired pathways: in the experiment, both polarisers and moderators learned during deliberation, levels of empathy were fairly high on both sides, and group pressures barely mattered. Finally, the low physical presence of immigrants in some discussion groups was associated with polarisation in the anti-immigrant direction, bolstering longstanding claims regarding the importance of *presence* for democratic politics.

Keywords: enclave deliberation; deliberative democracy; opinion polarisation; opinion change; immigration attitudes

Introduction

Empirical studies of citizen deliberation suggest that participants often change opinions (and also quite radically; see, e.g., Fishkin 2009). Luskin et al. (2002) claim that knowledge gain is an important mechanism of opinion change, whereas Sanders (2012) was unable to identify any robust predictor of opinion change in a study based on a pan-European deliberative poll (Europolis). A largely under-studied area in this regard is why some participants become more extreme – that is, polarise their opinions due to deliberation – and why others moderate them. Moderation is normally seen as a desirable outcome of a deliberative process: by carefully listening to others, participants with extreme opinions realise that there is merit in another's positions and arguments. By contrast, polarisation is frequently considered to be a suspicious outcome. According to Sunstein (2002), group polarisation reflects a dynamic psychological process, whereby groups move to the extreme on the basis of biased information processing and biases in the argument pool. Recent lines of theorising put a question mark on this interpretation, arguing that polarisation may not necessarily be a bad thing since it may simply reflect preference clarification – that is, people become aware of what they really want (Knight & Johnson 2011). In other words, the reinforcement of existing opinions may have deliberative dimensions (or, is at least not anti-deliberative).

Despite some recent contributions on polarisation and moderation (Sunstein 2009; Jones 2013; Grönlund et al. 2015), we still know very little about what explains these tendencies at the individual level. This article addresses this gap in the literature by focusing on participants whose opinion change deviates from the average change in a deliberative event, either in a more extreme or a more moderate direction. The article explores what drives polarisation and moderation by focusing on a batch of individual-level and group-related variables. We focus on age, gender, education, knowledge of the topic, empathy, social trust, ideology, attitudinal openness toward immigrants, clarification of opinion as well as group pressures and group composition, and explore how they affect polarisation and moderation. Furthermore, we engage in a normative assessment of polarisation and moderation: the cognitive, affective and group-related factors that we analyse also carry normative significance, and we evaluate whether the moderation or polarisation of opinions occurs in accordance with normatively desired pathways or not.

We assess polarisation and moderation in the context of an experiment where 207 citizens deliberated on immigration in Turku (Finland). This experiment is particularly useful for our research purpose, since it manipulated the group context, by assigning participants in like-minded groups and diverse opinion groups (mixed groups). The experiment was designed in order to test what happens when groups of individuals with similar initial opinions discuss with deliberative discussion rules and the presence of a trained moderator. Even though group polarisation did not occur to a large extent (Grönlund et al. 2015), both polarisation and moderation are observed at the individual level. Surely, polarisation at an aggregate level might be problematic from a democratic point of view, but our results show that polarisers and moderators did not fundamentally differ with regard to epistemic, ethical and group-related factors. This suggests that polarisation may reflect preference clarification in that individuals better understand what they really want.

The structure of the article is as follows. First, we provide a theoretical discussion on polarisation and moderation of opinions, identifying its individual-level and group-related antecedents. Second, we present the procedures of the experiment. Third, we report the results, followed by a discussion. Concluding remarks are presented at the end of the article.

The polarisation and moderation of opinions: A theoretical framework

Theories of deliberative democracy contain many empirical assumptions about preference and opinion formation. The thrust of deliberative approaches is that citizens are often not adequately informed about political issues and have not sufficiently engaged in weighing reasons for different policy positions (Muhlberger & Weber 2006; Fishkin 2009). Existing empirical findings are mixed in terms of the extent of opinion change in the aggregate. Some studies show major and radical changes in opinions at group level (Fishkin & Luskin 1999; Luskin et al. 2002; Goodin & Niemeyer 2003; Blais et al. 2008), whereas others show only minor changes (Denver et al. 1995; Merkle 1996; Hall et al. 2011). But most of the time, there are opinion changes at the *individual level* with movements in different directions that go undetected at the group level (Barabas 2004; Andersen & Hansen 2007). To understand the dynamics of opinion formation and opinion change, it is necessary to better understand individual-level processes underlying opinion change. In particular, why do some participants polarise their opinions, while others moderate them? This is not only

interesting from an analytical and empirical point of view, but it also carries significant normative meaning.

To date, the concept of 'polarisation' is mainly used at the aggregate level. According to Sunstein (2002: 176), 'group polarization means that members of a deliberating group predictably move toward a more extreme point in the direction indicated by the members' pre-deliberation tendencies'. This article is not about examining group polarisation; its focus is on individual polarisation in groups, and we explore the polarisation and moderation of opinions at the individual level. The term 'polarisation' is used to indicate that individuals move towards the extreme by strengthening their initial opinion (see, e.g., Wojcieszak (2011) for a similar interpretation). Reflecting on polarisation studies in psychology, Miller et al. (1993) strongly emphasise 'the importance of considering individual differences in the direction of reported attitude change'.

The standard assumption in the group polarisation literature is that the polarisation of opinions is due to group context (an aspect that we will discuss in more detailed fashion below). Sunstein (2002) has repeatedly argued that if participants enter deliberation with only like-minded people they become more extreme. Biased assimilation of information and biases in the argument pool are assumed to be the underlying causes for opinions becoming more extreme. More recently, however, the literature has offered contradictory findings about deliberation in enclaves. Findings from Karpowitz et al. (2009) and Grönlund et al. (2015) contradict the polarisation hypothesis of enclave deliberation since they find no clear evidence of group polarisation and amplification of cognitive errors when like-minded groups deliberated with moderators and rules.

Many deliberative scholars consider polarisation as a problematic outcome and moderation as a desired outcome of a deliberative process. O'Flynn (2007: 738), for instance, argues that moderation represents a marker of political equality:

> [I]f we enter into a democratic process in a spirit that recognizes that other citizens have equal standing with ourselves, we shall be ready to moderate our claims, since this is what equality requires in the face of the different and competing views of our fellow citizens.

Empirically, a properly set-up deliberative process where people are exposed to information and arguments that differ from their initial understanding might open up for (even strong) opinions to change and help participants to reach consensus and enlightenment (Barabas 2004: 689–690). Individuals with strong opinions may also become more uncertain and ambivalent when they interact with others and get to know the 'other side' (Ackerman & Fishkin 2004: 53; Mutz 2006: 102–105). Indeed, there are several studies indicating that opinions converge or move to the middle after deliberation (Lang 2007; French & Laver 2009; Farrar et al. 2010; Schweigert 2010).

In recent years, however, deliberative theory has opened itself up to a broader variety of opinion transformations, going beyond moderation and the search for consensual solutions. For instance, participants in a deliberative process might initially think that their preferences are reconcilable (or not too distant), but find out in discussion that the opposite is actually true. In a similar vein, Knight and Johnson (2011: 145) argue that 'even if as a result of the increased information that political argument makes available, individuals come to hold their preferences more reflectively, it in no way follows that this will lead to greater

substantive agreement at the aggregate level'. Knight and Johnson consider clarification and 'structured disagreement' more important than opinion change *per se*; and clarification may well encompass polarisation, moderation, or stability of opinions.

Similar trends are observable in psychology. Classic studies assumed that opinion polarisation is a product of biased information assimilation or 'motivated reasoning', and hence 'irrational' (see, e.g., Lord et al. 1979). Recent research challenges this view. Using Bayesian network analysis, Jern et al. (2014) show that polarisation can indeed be consistent with what they call 'a normative account of belief revision'. Belief divergence (or, polarisation) is not irrational when people have different prior beliefs about how specific factors affect an outcome. Suppose a test result is most probable when either factor A combines with a high level of B or factor C combines with a low level of factor B. Then two people with different prior beliefs about the level of factor B may come to opposite conclusions about the causal relevance of factor A and C being confronted with the same test result (see Jern et al. 2014: 209). This is a fairly common situation for many political issues, such as immigration, the focus of our study, since 'hypotheses are rarely considered in isolation, and inferences about one hypothesis typically depend on additional hypotheses and beliefs' (Jern et al. 2014: 209). As Benoit and Dubra (2014) concur,

> when a person is presented with equivocal evidence, that is, evidence that can reasonably interpreted as being either in favor or against a proposition, his beliefs can reasonably move either towards or away from accepting the proposition, or not move at all, and by that very fact, the harmonization, moderation, and polarization of two individuals are all reasonable outcomes.

Surely, both deliberative democrats and psychologists would insist that if polarisation (or, moderation) occurs, it should not be the product of undesirable group dynamics or on other non-deliberative pathways. We shall return to this issue in due course.

We will discuss a number of individual-level and group-related factors that may affect opinion change in general and the polarisation and moderation of opinions in particular. There are, of course, a myriad of factors that may drive the polarisation and moderation of opinions; we shall concentrate on those factors that are not only particularly relevant from an empirical-analytical point of view, but also from a normative perspective. We will focus on cognitive and socioeconomic variables, empathy and trust, ideology and openness towards immigrants, clarification of opinion and group effects.

Individual-level factors

Cognitive and socioeconomic variables. The general assumption is that higher levels of education and knowledge make more sophisticated deliberators: better reasoning skills and higher knowledge levels involve a better ability to consider reasoned arguments of others (Rosenberg 2007: 342–343). With regard to the polarisation and moderation of opinions, one hypothesis is that high education and high knowledge on the issue fosters the moderation of opinions. This is for two reasons. First, more educated and more knowledgeable persons may have more profound democratic values leading to higher levels of tolerance and respect and hence lead them towards opinion moderation (Mendelberg 2002). A second mechanism

is that more educated and more knowledgeable persons may also possess higher levels of cognitive complexity (Tausczik & Pennebaker 2010; Wyss & Beste 2014).

Cognitive complexity captures the degree to which an individual perceives, distinguishes and integrates various dimensions of an issue under discussion. High scores of cognitive complexity indicate that an individual is able to accommodate conflicting goals or values (Gruenfeld & Preston 2000) as well as recognise that different opinions are legitimate and can be held simultaneously. Consequently, one can expect that more educated and more knowledgeable persons possess higher levels of cognitive complexity and moderate their opinions when they are confronted with other viewpoints and counterarguments.

By contrast, persons with less education or knowledge may possess lower levels of cognitive complexity, which makes them more conducive to polarisation. However, a counter-hypothesis derived from Zaller (1992) suggests that highly educated and knowledgeable persons more strongly hold onto their attitudes. They might also be better at finding support for their own perspectives (see Taber & Lodge 2006). Thus, they may be more resistant both to moderation and polarisation trends than individuals with low education and low knowledge who are not so confident in their opinions and thus find it easier to assimilate new information and subsequently change their minds. Barabas (2004) found that participants with high knowledge who debated social security policy changed their views the least, and, when deliberation was not consensual, they even strengthened their prior opinions.

Other sociodemographic factors, such as gender and age, might also be influential in shaping opinion formation. However, findings are contradictory when it comes to the impact of gender and age on opinion change. In a citizen deliberation event in Finland, the oldest participants changed their opinions the most (towards lower support for nuclear power), while education had no effect on opinion change (Setälä et al. 2010). French and Laver's (2009) experiment found that participants with low education changed their opinion the most while age had no effect. Some studies indicate that women are more easily persuadable than men and respond more to debate and influence (Karpowitz et al. 2012). However, much seems to depend on the topic of deliberation since the gender effect can be reversed for topics where women have strong attitudes (Suiter at al. 2014: 3). Overall, it seems difficult to make clear-cut predictions for socioeconomic factors and polarisation or moderation trends. We leave it to the empirical analysis to unravel relevant effects.

Empathy and trust. We expect that empathy and trust are important drivers of moderation. The term 'empathy' has been defined in a number of ways and various elements have been connected to it (e.g., Hoffman 2000; Preston & De Waal 2002; Walter 2012). Roughly speaking, empathy is the ability to put oneself in another's position. Some scholars (including many deliberative democrats), emphasise the cognitive side of empathy (i.e., taking others' perspectives), which leads to understanding what others think (Walter 2012). Other scholars stress the emotional side of empathy which is about *role-taking* (i.e., to feel what others feel) (Morrell 2010: 58–62). Both variants converge in predicting that higher levels of empathy decrease the likelihood that people attribute negative motives to others whereas positive perceptions of out-groups increase (Morrell 2010: 107, 114). There is general agreement that this should lead to the moderation of opinions (see Rosenberg 2007: 355). As Morrell (2010: 126) concurs, 'in order to decrease biases and polarization, and increase cooperation

and reciprocity, deliberators must demonstrate predispositions to both perspective taking and empathic concern'. Since the cognitive and emotional aspects of empathy are strongly correlated empirically, we analyse empathy as a compound phenomenon in the empirical analysis.[1]

By 'social trust' we understand generalised interpersonal trust as opposed to particularised interpersonal trust. According to Newton (2007: 344–345), particularised interpersonal trust means trusting only individuals you know (i.e., trust one's own family or social group), while generalised interpersonal trust entails general feelings that most people can be trusted. Trust is said to contribute to social integration, cooperation, personal life satisfaction and optimism (Uslaner 2001). Most importantly, trust is also about understanding others' opinions. Similar to empathy, we expect people with high values of generalised interpersonal trust to approach disagreement with understanding and cooperative attitudes, thus leading them to moderate their opinions.

Ideology and attitudinal openness toward immigrants. Since we take the direction of opinion change into account – namely the polarisation and moderation towards pro- and anti-immigration directions – we also need to consider ideological factors. Modern cleavage literature suggests the existence of a two-dimensional map of economic and cultural integration or demarcation (see Kriesi et al. 2006). Consequently, we focus both on left-right ideology and on attitudes on openness/closedness towards immigrants. One expectation is that persons with clear ideological profiles are more confident about their positions and thus might either keep their opinions on immigration or polarise in one or the other direction. Consequently, a clear leftist and openist or progressive stance may facilitate post-deliberative polarisation in pro-immigration directions, whereas a clear rightist and conservative stance facilitates post-deliberative polarisation in anti-immigration directions. Another expectation is that people with anti-immigration attitudes – who frequently have less education, less empathy and less social trust than people with pro-immigration attitudes – are especially prone to polarisation tendencies.

Group effects

As mentioned above, opinion polarisation and moderation might be affected by the composition of groups. Deliberation in like-minded groups – or, enclaves – can breed groupthink since the argument pool might be limited and the group makes poor decisions based on incomplete and biased information. By contrast, in groups with opinion diversity, groupthink mechanisms should not set in, thus stifling opinion polarisation.

Two types of mechanisms behind polarisation have been identified: social comparison and persuasive arguments (Isenberg 1986; Sunstein 2002: 179–180; Farrar et al. 2009: 616). The former refers to the tendency of individuals to act in order to win social acceptance from other members of the group. In order to be accepted, individuals need to process information of how other people present themselves, and adjust their own behaviour accordingly (Isenberg 1986: 1142). Individuals may act in different ways in order to be perceived favourably by other group members. For instance, they may try to adjust their opinions according to the dominant view in the group. Social psychology experiments have also demonstrated that group pressures work in the way that people tend to conform to

the views of the majority (Asch 1951). If someone agrees with you, you are apt to like that person more and since everybody wants to be liked, this imposes a lot of pressure on people with views inconsistent with the group's consensus (Sunstein 2006: 68).

The other mechanism behind group polarisation – the deployment of persuasive arguments – is based on the idea that individuals are convinced by the contents of arguments put forward in the group. Consequently, if arguments heard in a group are biased in one direction, there is likely to be a further shift to this direction. Group polarisation is likely to be reinforced by biases in information processing. 'Confirmation bias' is a well-established phenomenon, which means that people are inclined to seek information confirming their prior beliefs and to disregard information against them (Mercier & Landemore 2012: 251). More generally, 'motivated reasoning' refers to a variety of cognitive and affective mechanisms that lead individuals to arrive at the conclusions at which they want to arrive (Kunda 1990). In a group of like-minded people, individual biases in information processing and reasoning are not checked by arguments put forward by individuals supporting conflicting views. Opinions are likely to polarise because individuals only hear arguments supporting their own prior position – in fact, they may even hear new arguments in support of it.[2]

With regard to polarisation and moderation, we also focus on the effects of group pressure. One plausible expectation is that group pressure might be conducive both to the polarisation *and* the moderation of opinions: in enclave groups with like-minded people, participants might feel pressure to go the 'extremes', whereas in non-enclave groups with diverse opinions, participants might also experience group pressure to move to the 'middle'. However, group dynamics might go unnoticed by participants, especially in enclave groups. By asking participants about their discussion experiences, we make a first stab at shedding some light on these questions.

Finally, another group composition effect – the presence of the *other* – might affect opinion polarisation or moderation as well. There are longstanding claims in the literature that the *physical presence* of less privileged or marginalised groups is not only a democratic predicament (Philips 1995), but matters for outcomes as well. According to social identity theory, members of a group might have a tendency to emphasise their similarities (i.e., strengthen in-group identity) and thus seek to find negative aspects of out-groups. The physical presence of out-groups may be an important factor in reducing such tendencies (Hogg 1993). As with empathy, we expect that more positive evaluations of out-groups or minorities are conducive to the moderation of opinions.

Normative considerations

Finally, our study goes beyond a simple inventory of the drivers behind the polarisation and moderation of opinions. The factors we introduced above also carry important normative significance. Assume that we find that opinion polarisation was based on low knowledge, low levels of empathy and group pressure. We then would conclude that the polarisation of opinion is indeed an undesirable outcome. But if the converse were true, that opinion polarisation is associated with high knowledge, high levels of empathy and no group pressure, we would conclude that polarisation, even though it is considered to be a suspicious outcome by many deliberative democrats, is at least not anti-deliberative. Or, similarly,

the moderation of opinions even though considered to be generally desirable by many deliberative democrats becomes questionable when it is based on low knowledge, low levels of empathy and group pressure.

We will normatively judge the polarisation and moderation of opinions on the basis of four factors (see also Sanders 2012; Baccaro et al. 2016): epistemic advancement and capacity (knowledge and education); empathy and understanding, which closely relate to the ethical dimension of deliberation (see Mansbridge et al. 2012); group pressure (both real and perceived); and (perceived) 'deliberativeness' of the discussion process. While the first three factors have been discussed before, a word on the fourth factor – (perceived) 'deliberativeness' – is in order. By focusing on perceived 'deliberativeness', we try to capture whether participants found the discussion process clarifying and in accordance with ethical deliberative principles (such as civil and inclusive discussion). If participants rate the process 'deliberative' according to these standards, polarisation (or moderation) trends – in conjunction with the presence of epistemic advancement, empathy and understanding, as well as the absence of group pressure – take on deliberative dimensions. Under such conditions, polarisation (or moderation) trends can also be seen in accordance with the clarification function stressed by Knight and Johnson (2011).

As well as such cognitive, ethical and group-related criteria, there may also be substantive considerations. When it comes to issues with a humanist or humanitarian dimension (such as immigration), Neblo (2007) shows that there is a tradition in deliberative theory holding that deliberation should be conducive to an expanded sense of community. Neblo (2007: 548) calls this 'progressive vanguardism': 'On this understanding, deliberative democracy is intrinsically and primarily an emancipatory project with strong substantive content, more or less tracking leftist political concerns.' This means that positions should become more progressive after deliberation (here: pro-immigration), whereas those people who already possess progressive positions should accentuate (or keep) these positions.

Clearly, 'progressive vanguardism' is a highly contestable position, and many deliberative theorists are wary of making a statement regarding the directionality of opinion change. Indeed, a communitarian variant of deliberation argues that 'good reasons' cannot be equated with liberal and progressive ideas and concepts (even if these frequently reflect the dominant position among philosophers), but arise on the basis of communal values and self-understandings that mirror local and temporal circumstances (Forst 2001). Therefore, our prime standard of normative evaluation will concern cognitive, ethical and group-related factors.

Data

The experiment

The topic of the deliberation experiment in Turku 2012 was immigration – a salient political issue in Finland. The purpose of the experiment was to compare deliberation processes and effects in two settings: deliberation in like-minded groups; and deliberation in groups consisting of people with clearly different opinions. Based on their initial opinions, the respondents were first classified as anti- or pro-immigrant (i.e., assigned to two opinion enclaves). After that, the participants were randomly allocated within their enclaves to

like-minded groups, mixed groups and a control group. Notice, however, that our study uses data from this experiment without being an experiment in itself.

A short survey (T1) was sent to a random sample of 12,000 adults in the Turku region. 39 per cent responded to the survey of 14 questions that measured immigration attitudes. Those whose value for the sum variable was > 8.3 were included in the pro-immigration enclave and those whose value was < 6.7 were included in the anti-immigration enclave.[3] The second survey (T2) was sent to 2,601 persons. This survey also included an invitation to take part in a discussion about immigration. A gift certificate of €90 was offered to each participant in the actual event. A total of 805 people volunteered and 366 were invited to take part in the deliberation; 207 people eventually showed up, with some bias towards the pro-immigration camp. The research team formed ten pro-immigration groups, five anti-immigration groups and 11 mixed groups.

Participants took part either on Saturday, 31 March or Sunday, 1 April 2012. The day started with a quiz measuring immigration-related and general political knowledge (T3). This was followed by a short briefing event containing unbiased and basic facts about immigration; then the small group discussions began. Every group had a facilitator and discussed for four hours. The rules emphasised respect for another's opinions, the importance of justifying one's opinions and keeping an open mind towards other arguments and positions. The deliberation day ended with a survey (T4) repeating questions in T1, T2 and T3, as well as questions about how participants experienced the event.

Operationalisation

Dependent variable: Opinion change

The dependent variable is based on a sum variable of 14 questions measuring the opinions on immigration (listed in the online appendix, Cronbach's alpha $= 0.94$). Each question was recoded into a scale from 0 to 1, whereby 1 indicates a pro-immigration opinion. We focus on those participants who changed their minds on immigration between T1 (before deliberation) and T4 (after deliberation) in a deviant manner compared to the group mean, either in more extreme or more moderate directions. In order to identify participants whose opinions did not change according to the mean change within their discussion group, we used a multistage process. As mentioned before, individuals with a clear positive (pro-immigration) or negative (anti-immigration) view about immigration were initially included in the experiment. They were randomly assigned either to a like-minded treatment, where discussion took place in groups whose initial views were similar on the topic of immigration or a mixed treatment, where discussion took place in groups with an equal number of participants from both opinion enclaves (four from each opinion enclave). We thereby obtain four experimental groups: pro-immigrants in like-minded opinion groups, pro-immigrants in mixed opinion groups, anti-immigrants in like-minded opinion groups and anti-immigrants in mixed opinion groups. Based on the aggregate opinion change in these four experimental groups, we identified participants whose opinions did not change according to the mean change within the equivalent group. We did so by first identifying the individual distance from the group mean change.[4]

Table 1. Four categories of deviant opinion change

Initial opinion and direction of change	N	Direction of opinion change
Category 1: Positive initial opinion toward immigration and positive change	22	Polarisation
Category 2: Negative initial opinion toward immigration and negative change	12	Polarisation
Category 3: Positive initial opinion toward immigration and negative change	17	Moderation
Category 4: Negative initial opinion toward immigration and positive change	13	Moderation
Total	64 (of 207)	

In the next stage we introduced the standard deviations, ranging from 1.01 to 1.49 in the four groups. Individuals with a distance from the group mean that was larger than one standard deviation were coded as deviant. Having taken the direction of deviance into account we could identify polarisers and moderators. However, this procedure is not without problems. For participants with an anti-immigrant opinion in a mixed opinion group, the mean change at the group level is +1.80, and the SD is 1.44. If an individual's opinion change is +0.33, the individual distance from the group mean (0.33 – 1.80 = –1.47) is more than one standard deviation, but the individual change is only 0.33. This individual would be coded as polarised in an anti-immigrant direction due to the difference between his change and the group's mean change (his personal value would be –1.47).[5] We manually checked all the codings and the values, and the problem occured only twice in the anti-immigrant mixed opinion group (not in any of the other groups), and it is due to the big mean change in that group. We manually recoded these two individuals (from polarised in anti-immigrant direction to the 'average' change (per group)) to correct the error (available from the authors upon request).

Consequently, we focus on individuals who have changed one standard deviation or more than the mean change in their experimental group. If the change is in the same direction as the initial opinion, the opinion change is considered to be polarisation. If the change is in the opposite direction (towards 'the other side'), the change is considered to be moderation. A total of 34 individuals (16 per cent) polarised their opinions – that is, they reinforced their opinions in the same direction as their initial opinion; and 30 individuals (15 per cent) depolarised their opinions – that is, they changed their opinions away from their initial opinion. When we combine polarisation and moderation with the direction of the opinion change – towards pro or anti positions – we obtain four groups of 'deviant changers', as displayed in Table 1.

Figure 1 shows the development of opinions on immigration in the four categories in comparison with the 'average' changers (per group).[6] While the graph displays clear differences in the magnitude of opinion change between polarisers, moderators and 'average' changers (per group), we also see that polarisation and moderation generally occur within 'ideological camps': only in category 3, where participants moderated their opinion

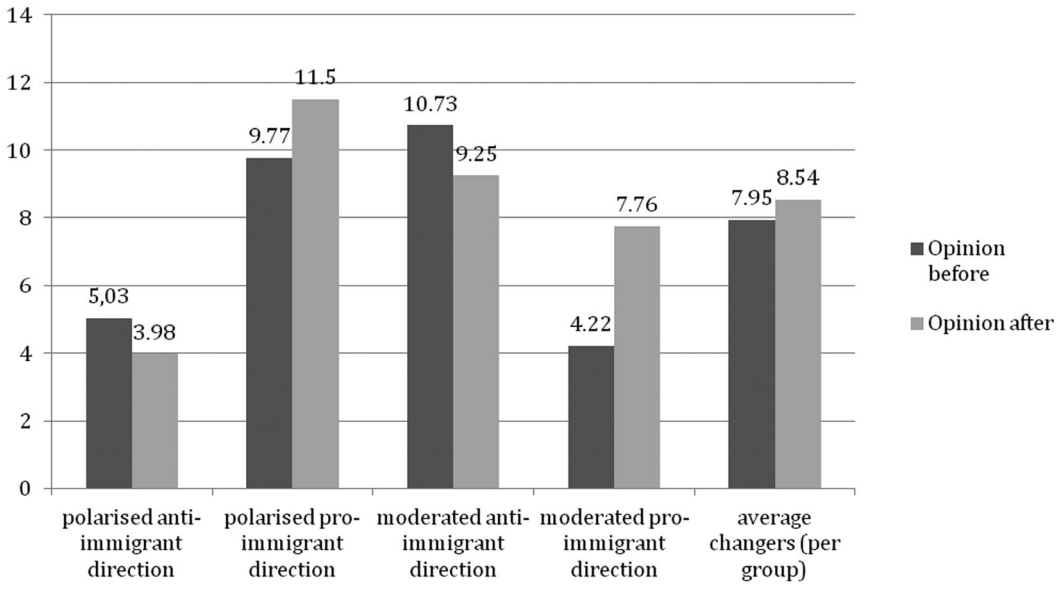

Figure 1. Opinion change for polarisers and moderators.

in pro-immigration directions, did opinion shift over the mid-point of the scale (7) towards the other 'ideological camp'.

Several comparisons among these four categories will be made. First, we will link opinion change in the four categories to individual-level and group-related variables. Second, we will compare the four categories in the context of polarising individuals and moderating individuals.

Operationalisation of the predictor variables

First, we consider education and knowledge on immigration. With regard to education, there is no standardised definitions of 'low' and 'high' education. We define 'low education' as elementary, vocational and secondary school, and 'high education' as upper college-level, polytechnic, and lower and upper university degree.

Immigration knowledge is measured through a quiz consisting of ten questions about immigration in Finland. In the analysis, we use the average percentage of correct answers. In order to capture both initial levels of knowledge and learning, we employ a measure for immigration knowledge before and after discussion.

Next, we focus on empathy and trust, which are sum variables recoded into a scale from 0 to 100 where 0 indicates the lowest, and 100 the highest levels. As with knowledge, we check for the variables before and after discussion.

With regard to ideology, we focus on left-right ideology. The participants were asked to place themselves on a scale, where 0 means the left and 10 means the right. We also consider how the participants feel about immigrants. This is measured in terms of openness and closedness towards immigrants. We use a sum variable, consisting of five questions.

The variable can vary between 0 and 100 where 0 indicates closedness and 100 indicates maximum openness.

Clarification of opinion is operationalised via the survey item 'I am now more certain about my opinion on immigration than before the discussion'. High scores indicate that participants feel more aware of what they think, whereas low levels indicate no clarification of opinion.

Perceived 'deliberativeness' is a sum variable, consisting of three survey items exploring whether the participants felt the discussion was pleasant and inclusive: (a) 'discussing in a group was a pleasant experience'; (b) 'it was easy for me to express my opinion during the discussion'; and (c) 'I would be happy to participate in a similar discussion again'. The variable can range from 0 to 100 with 0 indicating the lowest score on perceived 'deliberativeness' and 100 indicating the highest score.

Group context includes variables with regard to the amount of immigrants in the discussion group and participants' subjective experiences of the discussion. The variable 'experienced group pressure' is a sum variable of three items: (a) 'some participants dominated the discussion too much'; (b) 'other participants interrupted when it was my turn to speak'; and (c) 'I found it difficult to listen to people who disagree with me'. The variable can vary between 0 and 100 with 0 indicating the highest level of experienced group pressure and 100 indicating the absence of group pressure.

The variable 'like-minded treatment' is the percentage of individuals who were part of a like-minded group compared to a mixed group.

For more information on the variables, see the online appendix.

Statistical analysis

The statistical analysis is conducted in a Bayesian framework (see, e.g., Gelman et al. 2003; Gill 2007). For our data, Bayesian inference has one distinctive advantage compared to the more widely used frequentist methods (i.e., the so-called 'Neyman-Pearson framework'): it makes inferences conditional on the actual sample, which is in contrast to frequentist statistics where inference is made to some hypothetical super-population. Given the fact that the various categories entail low numbers of observations, inference to a super-population is misplaced and Bayesian inference enables us to say whether observed differences in the actual sample is significant.

Given our categorical dependent variables – involving three and five categories (see below) – we run Bayesian multinomial probits. Due to the relatively low numbers in the polarising and moderating categories, we refrained from inserting all predictor variables into the models. For each of the dependent variables, we have probed for various specifications and report a model that is both theoretically relevant and computationally efficient. While a Bayesian framework permits incorporating substantive prior knowledge, we refrain from doing so. For the context of our study, there is very little knowledge from prior studies; hence, we employ diffuse multivariate normal priors for the coefficients (Jackman 2009). For the multinomial probit models, we ran four chains for 100,000 iterations initiating from overdispersed starting points. We discarded the first 10,000 iterations as burn-in and assessed convergence using the Brooks-Gelman version of the Gelman-Rubin test statistic. In addition, neither the Heidelberg-Welch nor Raftery-Lewis nor Geweke diagnostics

Table 2. Comparison between polarisers and moderators based on the Bayesian multinomial probit model

	Polarisers Reference: Moderators	Polarisers Reference: Average changers (per group)	Moderators Reference: Average changers (per group)
Education	–0.432^ (0.293)	–0.305^ (0.229)	0.135 (0.235)
Age	–0.304 (0.314)	0.379* (0.211)	0.683** (0.274)
Gender	0.481 (0.552)	1.075** (0.428)	0.601 (0.488)
Ideology	–0.751** (0.292)	–0.286^ (0.218)	0.468* (0.248)
Knowledge	0.512* (0.307)	0.325^ (0.234)	–0.204 (0.256)
Social trust	–0.269 (0.312)	–0.279 (0.218)	–0.013 (0.259)
Empathy	0.295 (0.304)	0.147 (0.218)	–0.141 (0.247)
Likeminded	0.494 (0.549)	0.335 (0.41)	–0.141 (0.454)
Intercept	–0.432^ (0.293)	–0.305^ (0.229)	0.135 (0.235)
Deviance	361	361	361
DIC	349	349	351
N	205	205	205

Notes: The table reports mean and standard deviation of coefficients' posterior distributions. For reasons of convenience, we additionally report whether a significant part of the distribution is settled either on the positive or the negative side. The reported significance levels are as follows: ***p < 0.001; **p < 0.01; *p < 0.05; ^p < 0.1

indicated any sign of non-convergence. We have also re-checked our findings in frequentist multinomial probit models, using cluster-corrected standard errors at the level of discussion groups; results are very similar (available from the authors upon request). In addition to the Bayesian framework, we illustrate the differences between the groups by using an analysis of means.

Results

First of all, like-minded and mixed groups drive neither polarisation nor moderation of individual preferences. This is suggested by a logit model in which we estimated the effect of group composition on the likelihood to polarise. The dependent variable of the model is whether or not participants reported an individual preference shift towards more extreme positions. On this basis, we are able to check whether like-minded groups foster opinion polarisation as well as whether mixed groups foster opinion moderation. We find no statistically significant and substantively large effects either for polarisation or for moderation (see the online appendix). These results bolster our claim that it is essential to focus on the individual level in order to explain polarisation and moderation trends.[7]

As displayed in Table 2, there are some differences between polarising and moderating individuals and the category of 'average' changers (per group). The statistically significant differences among the groups are age, gender, left-right ideology, education and knowledge (whereby for the last two variables, the effects are only marginally significant), while trust, empathy and like-minded versus mixed groups do not drive polarisation and moderation.[8]

In comparison to the category of 'average' changers (per group), polarising tendencies are slightly more common among women. However, since we lack a clear theoretical rationale for this finding, we do not strongly interpret it. Ideology matters as well: compared to polarisers, left-wing people have a higher tendency to moderate their opinions. Again, we do not strongly interpret this result, even though a new study on American political blog posts shows that liberal bloggers have a higher level of cognitive or integrative complexity compared to conservative bloggers, which the authors relate to deliberative quality (Brundidge et al. 2014). There is also a slight tendency for polarisers to have a lower level of education, which would be in line with one postulated mechanism – namely that more educated persons have more profound democratic values leading to opinion moderation. Yet, the effect is only marginally significant and the exact chain of mechanisms unclear. Intriguingly, immigration knowledge is slightly higher among polarisers, which is in line with the suggestion that more knowledgeable persons more strongly hold onto their attitudes, which can be conducive to polarisation trends.

Overall, the pattern of differences among the three categories is 'patchy'. Yet based on previous research on deliberation and group polarisation, we expected that those moderating their opinions would have more empathy, more social trust and that enclaves would fuel polarisation. But that is not the case: various individual-level and group-related factors are not good predictors of polarisation and moderation. In the regression models, we also probed for various interaction effects, such as like-minded/mixed groups and education, immigration knowledge and empathy, but found none (results available from the authors upon request).

Next, we focus on different types of polarising and moderating individuals by drawing comparisons between those who reinforced their negative opinion towards immigration, those who reinforced their positive opinion towards immigration, and those who moderated their opinions towards pro- and anti-immigrant positions. Given the fact that the number of individuals in these four categories is quite small, we cannot draw sweeping generalisations. Again, we use a Bayesian multinomial probit model with a reduced set of significant and theoretically relevant control variables. We have also run models with all of the variables that we use in the normative evaluation below – namely empathy, knowledge (at T4), like-minded versus enclave groups, civil discussion and clarification of opinions. All of these variables did not yield statistically significant differences among the four groups (results available from the authors upon request).

For our normative evaluation, we are also interested in 'profiling' different types of polarising and moderating individuals. Therefore, we consider the raw figures (see Table 3), complemented by non-parametric tests. With four groups in the analysis, eight pairwise comparisons have to be made. In general, a Mann-Whitney-Test with Bonferroni correction is employed in such a situation. However, with so many comparisons, Bonferroni can be seen as a relatively strict test. Therefore, we have chosen to use Conover's Post Hoc test that is more permissive and better suited to test significances for more than four groups. Significant differences between the categories are reported in the rightmost columns of Table 3. In order to compare the four groups in a straightforward way, we focus on Figure 2 (the model is presented in the online appendix; notice that the model compares five groups, including the 'average' changers; for reasons of clarity, however, we only report the most interesting group comparisons).

Table 3. Normative evaluation of different groups of polarisers and moderators

	(1) Polarisation toward anti-immigration positions (N = 12)	(2) Polarisation toward pro-immigration positions (N = 22)	(3) Moderation toward pro-immigration positions (N = 13)	(4) Moderation toward anti-immigration positions (n = 17)	1–2	2–3	2–4	1–3	1–4	3–4
Individual-level factors										
Highly educated (%)[1*]	17	41	17	65					*	*
Education (1–8)[**]	3.8	4.6	3.4	5.7		^	^		**	**
Immigration knowledge – pre	42	48	35	42		**				
Immigration knowledge – post[^]	58	62	52	64		^				^
Empathy – pre	64	72	64	66						
Empathy – post	69	73	65	67						
Feelings about immigrants[***]	27	65	36	66	^	*				*
Civil and inclusive discussion	67	66	61	62						
Clarification of opinion	75	71	69	67						
Group context										
Enclaves (%)[1]	50	77	39	65						
Immigrant in group (%)[1^]	8	59	46	47	*				^	
Experienced group pressure[*]	29	17	33	30	^	*	*			

Notes: [1]Significance test: Fisher's Exact Test, other variables: Kruskal Wallis Test: ^p < 0.10; *p < 0.05; **p < 0.01; ***p < 0.001. Columns to the right show significant pairwise differences (Conover's Post Hoc Test)

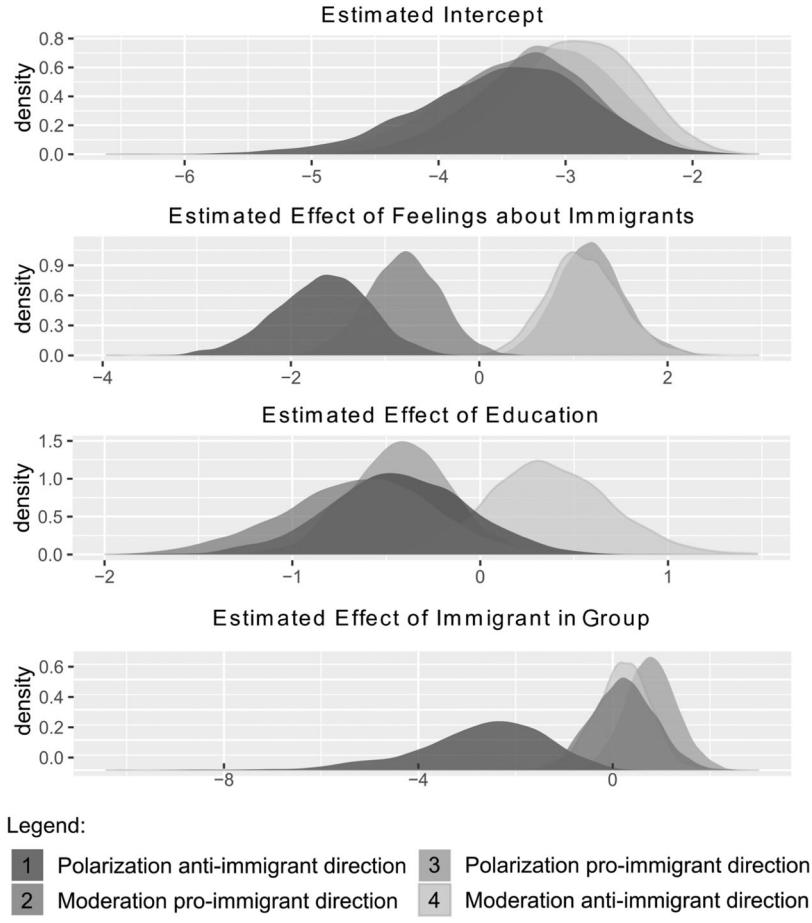

Figure 2. Bayesian estimates of the coefficients in the Multinominal Probit Model explaining the four categories of polarisers and moderators.

A first finding is that after the deliberative process there are clear and statistically significant differences between 'unexpected' categories: polarisers in anti-immigrant directions (1) and moderators in pro-immigrant directions (2) are different from polarisers in pro-immigrant directions (3) and moderators in anti-immigrant directions (4). What seems puzzling at first glance can be solved by taking *initial* attitudinal factors into account. Remember that people in categories 1 and 2 are in the anti-immigration camp at T1, whereas people in categories 3 and 4 are in the pro-immigration camp at T1 (see Figure 1). Even though there were polarisation and moderation trends in both ideological camps, on average the deviant opinion changers still ended up in the same ideological camp after deliberation (with the exception of individuals in category 3 moving to a post-deliberative score of 7.76). The key message from the graphs is that polarising and moderating trends can be observed in both ideological camps. This challenges the expectation that being in a specific ideological camp produces polarisation in that direction; it also challenges the expectation that especially less-educated people with anti-immigrant attitudes would be prone to polarisation tendencies. Corroborating this interpretation,

education is also not clearly stratified among the various categories of polarisers and moderators; we only find that high education is more likely among the anti-immigrant moderators.

However, there is one notable effect with important practical and normative ramifications. Those who polarised their opinions in anti-immigrant directions very rarely had an immigrant in their discussion group. The Bayesian multinomial probit shows the probability of being among anti-immigration polarisers is higher when no immigrant is present in the discussion group. This finding is in line with the argument that the presence of the *other* – here an immigrant in the discussion group – might have an important effect on how opinions change.

Let us now do some 'profiling' of the individuals in the four categories in connection with our normative considerations. First, we focus on those participants who moderated their views in pro-immigrant directions. As Table 3 shows, especially their low initial knowledge on immigration issues (35 per cent correct answers) raises some questions of whether their opinion changes are in line with a deliberative pathway to belief revision. However, they learned quite a bit during discussion (52 per cent correct answers at T4; for a statistical test of learning, see the online appendix). Moreover, perceived 'deliberativeness' and levels of empathy were not different from the other groups and with about 60 points (out of 100) substantively quite high as well; and there were no signs of group pressure. As such, moderation in this category is clearly associated with deliberatively desirable features.

Next, we consider the category of individuals moderating their opinions toward anti-immigration positions. The fact that some well-educated (65 per cent with high education) and strongly empathetic people (66 points out of 100) moderated their opinions toward anti-immigrant positions raises some interesting normative questions. One could say that moderation on such cognitive and ethical grounds is exactly what deliberative theorists vie for. Yet from the perspective of 'progressive vanguardism' (see Neblo 2007), this moderation trend is questionable. Individuals with more progressive opinions, as in column 4, should certainly listen to 'anti-immigrant' arguments but not necessarily shift their opinions in this direction, since humanitarian positions are expected to have a higher level of universalisability and should therefore have a higher persuasive capacity. As mentioned before, not only is 'progressive vanguardism' a contestable position, those moderating their opinions in anti-immigrant directions also ended up in the pro-immigration camp after deliberation (on the 0–14 immigration scale, their post-deliberative opinions were at 9.25). Thus, we would still qualify this moderation trend as largely consistent with deliberative ideals.

When it comes to the polarisation of opinions, the picture is not fundamentally different from the one we obtained for moderation. At first glance, those individuals who polarised towards anti-immigration positions might conform to the conventional view. As can be gleaned from Table 3, their level of education was lower compared to those individuals who moderated in anti-immigration directions. Yet, they learned quite a bit during discussion and equalised their knowledge compared to the other groups (the percentage of correct answers was 58 at T4; for a statistical test of learning, see the online appendix). Moreover, there were neither group pressures nor differing levels of empathy, nor differences in perceived 'deliberativeness' in comparison with the other groups. Under such conditions, it might be difficult to judge polarisation as normatively problematic. Interesting are also

those individuals who polarised towards pro-immigration directions. They experienced a bit more group pressure than others (an effect which is also corroborated by the Bayesian multinomial probit), but otherwise they were not different from moderators in pro-immigrant directions when it comes to learning, empathy and 'deliberativeness'. Hence, it might be difficult to discard this polarisation trend as normatively questionable.

In conclusion, we have mixed news for deliberative theorists. First, the story behind our results is a bit like the one presented by Sanders (2012) for a transnational deliberative poll (Europolis). Participants change minds after deliberation (and in our case, we explicitly focused on those who changed the most in the discussion groups), but we do not really know why that is, even though we have probed for a large variety of factors. One explanation is that the organised citizen deliberation (as in our case) leads to different dynamics and outcomes than settings where communicative processes are far less structured (as in most psychological experiments). A new line of research provides hints in this direction (Grönlund et al. 2015; Baccaro et al. 2016). Second, there is broad agreement in deliberative theory that normatively desirable opinion changes should at least reflect a high epistemic quality and respective capacities of participants, the absence of group pressure, some ethical aspects such as empathy and 'deliberativeness' in the discussion process. In this regard, our results reveal an intriguing picture: polarisers and moderators were not fundamentally different on epistemic grounds after deliberation – that is, they all learned, and those with low initial knowledge learned even more, group factors barely mattered, levels of empathy as well as perceived 'deliberativeness' were high and did not differ among moderators and polarisers either. This is good news for the growing number of deliberative democrats arguing that polarisation may reflect preference clarification in that participants better understand what they really want (e.g., Knight & Johnson 2011). In other words, polarisation (at the individual level) is not necessarily anti-deliberative.

Conclusion

This article explored the drivers behind opinion polarisation and opinion moderation. Focusing on data from a citizen deliberation experiment on immigration in Finland, we analysed participants who have changed their minds more than the group average, either in a more extreme or a more moderate direction. The results are quite striking: neither individual-level nor group-related factors are good predictors for the polarisation and the moderation of opinions. There are, however, some differences between individuals polarising in pro- or anti-immigrant directions. But the differences we found are due to basic ideological pre-dispositions. Nonetheless, we found one important factor for opinion polarisation in anti-immigrant directions: the absence of immigrants in the small group discussion, which is in accordance with longstanding claims regarding the importance of presence for democratic politics (Philips 1995). The key finding of this study is, however, that polarisers and moderators did not fundamentally differ with regard to epistemic, ethical and group-related factors (after the deliberative process). As such, our study challenges conventional interpretations of opinion polarisation and moderation. While polarisation might be problematic from a democratic point of view *at the aggregate level* (since societies and polities tend to fare better when polarisation is not extreme), moderation and polarisation are not normatively good or bad *per se at the individual level*. As long

as these pathways involve epistemic advancement, clarification and ethical aspects, and are not heavily influenced by group dynamics, both polarisation and moderation are not anti-deliberative. This is fully in line with recent advances in philosophy and psychology to understand polarisation not as irrational behaviour, but – given circumstances and pathways – as an outcome, which may conform to normative conceptions of belief formation.

We acknowledge that this study has several limitations. While we have considered a large array of factors that could potentially affect the polarisation or moderation of opinions, our list is far from exhaustive. For instance, detailed analyses of the discussions could tell us more about what was really going on in the groups (see Gerber et al. 2014). We have applied such an analysis focusing on equality and rationality in discussion (Lindell 2015), but due to space considerations we can only provide a summary analysis. However, the analysis of the group discussions corroborate our findings that polarisation is not 'irrational' behaviour. We found no differences between the various groups of polarisers, moderators and a reference group of 'average' changers (per group) with regard to speech activity, suggesting that inequalities in presenting arguments was not a driving force behind polarisation or moderation. Contrary to standard expectations, polarising individuals were not more disrespectful than moderating individuals and 'average' changers (per group); interestingly, moderating individuals were even slightly less respectful than polarising ones. Moreover, polarising individuals even presented the largest proportion of rational arguments, even though this pattern only occurs for one of the polarising groups.

Overall, the fact that frequently mentioned sources for opinion polarisation or moderation – such as group dynamics or empathy – did not matter in this experiment has major implications for future research. Indeed, people do change their minds in deliberative events (and even massively), and we need to devise a more dedicated research programme to understand why that is the case. Otherwise, it remains unclear to what extent processes of citizen deliberation can approximate the normative goals of deliberation, such as well-considered judgments and decision legitimacy.

Acknowledgements

We wish to thank Jane Mansbridge for her insightful comments, and the other participants at the Association Française de Science Politique Congrès, 2015. We also want to thank participants at the ECPR Joint Sessions in Salamanca in April 2014, and participants at the ECPR General Conference in Glasgow in September 2014 for comments and suggestions on previous versions of this article. Finally, we thank the two anonymous reviewers for their constructive feedback.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Table 1. Logit model showing no group-based effects on polarization and moderation
Figure 1. Bayesian estimates of the coefficients in the Multinominal Logit Model explaining polarizers and moderators

Table 2. Comparison of the four categories based on Bayesian Multinominal Probit Model
Table 3. Coding of variables
Table 4. Knowledge gains for the five groups
Table 5. Descriptive statistics of polarizers, moderators, and "average" changers (per group)
Table 6. Difference-in-difference Model for testing for polarization and motorization effects While alternative tests (T-test and logit model explaining the probability for being among the polarizers) do not support the polarization hypothesis by likeminded groups, a difference in difference analysis shows a slightly different picture. Only for the participants with counter-immigration attitudes the interaction term between mixed group and time is significant on the 0.95 level

Notes

1. We also analysed cognitive and affective empathy as two separate variables, but the result was the same as for the compound variable.
2. There is a debate between Lord et al. (1979) and Taber and Lodge (2006) on the sources of opinion polarisation (Ross 2012): while the former understand opinion polarisation as a product of misattribution, the latter understand it as a consequence of motivated reasoning. In this study we cannot test which of the two mechanisms causes opinion polarisation; hence we only take notice of this debate.
3. Due to the experimental design, only people with clear views about immigration were included, moderates with scores between 6.7–8.3 were excluded (n = 631)
4. We counted the individual distance from group mean by taking the individual change minus group's mean change.
5. Here are some examples where the procedure works smoothly: (a) Mean change 0.67, SD 1.06, individual change 0.33: 0.33 – 0.67 = –0.34, this individual is not coded as deviant (which is correct); (b) Mean change 0.67, SD 1.06, individual change 3.0: 3.0 – 0.67 = 2.33, this individual is coded as a deviant (which is correct); (c) Mean change 0.67, SD 1.06, individual change –0.5: –0.5 – 0.67 = –1.17, this individual is coded as a deviant (which is correct, since (s)he has changed opinion in the opposite direction of the group mean).
6. 'Average' changers (per group) include individuals whose opinions have changed close to the group mean change.
7. A difference-in-difference analysis draws a slightly different picture, however. We implemented the difference-in-difference analysis in a standard regression model (see, e.g., Smets & Isernia 2014), modelling both treatment (like-minded versus mixed group) and time. The results reveal that, in the case of anti-immigration participants, the coefficient of the interaction term is significant (b = –1.1; SD = 0.54*), suggesting that there is a difference between anti-immigration individuals in the like-minded groups and in the mixed groups. Yet, what we observe in the like-minded groups is a moderation trend towards pro-immigration attitudes, and not a polarisation trend towards anti-immigration attitudes (as one would expect; see also Grönlund et al. 2015). This unexpected finding makes it even more interesting to focus on individual trends of moderation and polarisation.
8. We have also run models including the variables immigration attitudes, experience of group pressure, amount of immigrants in group, perceived 'deliberativeness' and clarification of opinion. None of these variables yielded statistical significance.

References

Ackerman, B. & Fishkin, J.S. (2004). *Deliberation day*. New Haven, CT: Yale University Press.

Andersen, V. & Hansen, K. (2007). How deliberation makes better citizens: The Danish Deliberative Poll on the euro. *European Journal of Political Research* 46(4): 531–556.

Asch, S.E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (ed.), *Groups, leadership and men*. Pittsburgh, PA: Carnegie Press.

Baccaro, L., Bächtiger, A. & Deville, M. (2016). Small differences that matter: The impact of discussion modalities on deliberative outcomes. *British Journal of Political Science*, forthcoming.

Barabas, J. (2004). How deliberation affects policy opinions. *American Political Science Review* 98: 687–701.

Benoit, J.-P. & Dubra, J. (2014). A theory of rational attitude polarization. Available online at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2529494

Blais, A., Carty, K.R. & Fournier, P. (2008). Do citizens' assemblies make reasoned choices? In M.E. Warren & H. Pearse (eds), *Designing deliberative democracy: The British Columbia Citizens' Assembly*. New York: Cambridge University Press.

Brundidge, J.A. et al. (2014). The 'deliberative digital divide': Opinion leadership and integrative complexity in the US political blogosphere. *Political Psychology*, forthcoming.

Denver, D., Hands, G. & Jones, B. (1995). Fishkin and the deliberative opinion poll: Lessons from a study of the Granada 500 television program. *Political Communication* 12(2): 147–156.

Farrar, C. et al. (2009). Does discussion group composition affect policy preferences? Results from three randomized experiments. *Political Psychology* 30(4): 615–647.

Farrar, C. et al. (2010). Disaggregating deliberation's effects: An experiment within a deliberative poll. *British Journal of Political Science* 40: 333–347.

Fishkin, J.S. (2009). *When the people speak: Deliberative democracy and public consultation*. Oxford: Oxford University Press.

Fishkin, J.S. & Luskin, R.C. (1999). Bringing deliberation to the democratic dialogue. In M. McCombs & A. Reynolds (eds), *The poll with a human face: The national issues convention experiment in political communication*. Mahwah, NJ: Routledge.

Forst, R. (2001). The rule of reasons: Three models of deliberative democracy. *Ratio Juris* 14(4): 345–378.

French, D. & Laver, M. (2009). Participation, bias, durable opinion shifts and sabotage through withdrawal in citizens' juries. *Political Studies* 57: 422–450.

Gelman, A. et al. (2003). *Bayesian data analysis*, 2nd edn. Boca Raton, FL: Chapman & Hall.

Gerber, M. et al. (2014). Deliberative and non-deliberative persuasion: Mechanisms of opinion formation in EuroPolis. *European Union Politics* 15(3): 410–429.

Gill, J. (2007). *Bayesian methods for the social and behavioural sciences*, 2nd edn. Boca Raton, FL: Chapman & Hall.

Goodin, R.E. & Niemeyer, S. (2003). When does deliberation begin? Internal reflection versus public discussion in deliberative democracy. *Political Studies* 51: 627–649.

Gruenfeld, D.H. & Preston, J. (2000). Upending the status quo: Cognitive complexity in US Supreme Court Justices who overturn legal precedent. *Personality and Social Psychology Bulletin* 26(8): 1013–1022.

Grönlund, K., Herne, K. & Setälä, M. (2015). Does enclave deliberation polarize opinions?. *Political Behavior* 37(4): 995–1020.

Hall, T.E., Wilson, P. & Newman, J. (2011). Evaluating the short- and long-term effects of a modified deliberative poll on Idahoans' attitudes and civic engagement related to energy options. *Journal of Public Deliberation* 7(1): 1–30.

Hoffman, M.L. (2000). *Empathy and moral development: Implications for caring and justice*. Cambridge: Cambridge University Press.

Hogg, M.A. (1993). Group cohesiveness: A critical review and some new directions. *European Review of Social Psychology* 4(1): 85–111.

Isenberg, D.J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology* 50(6): 1141–1151.

Jackman, S.D. (2009). *Bayesian analysis for the social sciences*. New York: Wiley.

Jern A., Chang, K-M.K & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review* 121(2): 206–224.

Jones, D.A. (2013). The polarizing effect of a partisan workplace. *PS: Political Science and Politics* 46(1): 67–73.

Karpowitz, C.F., Raphael, C. & Hammond IV,A.S. (2009). Deliberative democracy and inequality: Two cheers for enclave deliberation among the disempowered. *Politics and Society* 37: 576–615.

Karpowitz, C.F., Mendelberg, T. & Shaker, L. (2012). Gender inequality in deliberative participation. *American Political Science Review* 106: 533–547.

Knight, J. & Johnson, J. (2011). *The priority of democracy: Political consequences of pragmatism*. Princeton, NJ: Princeton University Press.

Kriesi, H. et al. (2006). Globalization and the transformation of the national political space: Six European countries compared. *European Journal of Political Research* 45(6): 921–957.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin* 108(3): 480–498.

Lang, A. (2007). But is it for real? The British Columbia Citizens' Assembly as a model of state-sponsored citizen empowerment. *Politics and Society* 35: 35–69.

Lindell, M. (2015). *Deliberation och åsiktsförändring – en studie av individegenskaper och gruppkontext*. Åbo: Åbo Akademi University Press. Published doctoral thesis.

Lord, C.G., Ross, L., & Lepper, M.R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* 37: 2098–2109.

Luskin, R.C., Fishkin, J.S. & Jowell, R. (2002). Considered opinions: Deliberative polling in Britain. *British Journal of Political Science* 32: 455–487.

Mansbridge, J. et al. (2012). A systemic approach to deliberative democracy. In J. Parkinson & J. Mansbridge (eds), *Deliberative systems*. Cambridge: Cambridge University Press.

Mercier, H. & Landemore, H. (2012). Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology* 33(2): 243–258.

Mendelberg, T. (2002). The deliberative citizen: Theory and evidence. *Political Decision Making, Deliberation and Participation* 6: 151–193.

Merkle, D. (1996). The National Issues Convention Deliberative Poll. *Public Opinion Quarterly* 60: 588–619.

Miller, A.G. et al. (1993). The attitude polarization phenomenon: Role of response measure, attitude extremity and behavioral consequences of reported attitude change. *Journal of Personality and Social Psychology* 64(4): 561–574.

Morrell, M. (2010). *Empathy and democracy: Feeling, thinking and deliberation*. University Park, PA: Pennsylvania State University Press.

Muhlberger, P. & Weber, L.M. (2006). Lessons from the Virtual Agora Project: The effects of agency, identity, information and deliberation on political knowledge. *Journal of Public Deliberation* 2: 1–35.

Mutz, D. (2006). *Hearing the other side: Deliberative versus participatory democracy*. New York: Cambridge University Press.

Neblo, M. (2007). Family disputes: Diversity in defining and measuring deliberation. *Swiss Political Science Review* 13(4): 527–557.

Newton, K.E. (2007). Social and political trust. In R.S. Dalton & H.-E. Klingemann (eds), *The Oxford handbook of political behavior*. Oxford: Oxford University Press.

O'Flynn, I. (2007). Divided societies and deliberative democracy. *British Journal of Political Science* 37: 731–751.

Philips, A. (1995). *The politics of presence*. Oxford: Clarendon Press.

Preston, S.D. & De Waal, F.B.M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences* 25(1): 1–72.

Rosenberg, S. (2007). Rethinking democratic deliberation: The limits and potential of citizen participation. *Polity* 39: 335–360.

Ross, L. (2012). "Reflections on Biased Assimilation and Belief Polarization", *Critical Review*, vol. 24, no. 2, pp. 233–245.

Sanders, D. (2012). The effects of deliberative polling in an EU-wide experiment: Five mechanisms in search of an explanation. *British Journal of Political Science* 42(3): 617–640.

Schweigert, F.J. (2010). Strengthening citizenship through deliberative polling. *Journal of Community Practice* 18: 19–39.

Setälä, M., Grönlund, K. & Herne, K. (2010). Citizen deliberation on nuclear power: A comparison of two decision-making methods. *Political Studies* 58(4): 688–714.
Smets, K. & Isernia, P. (2014). The role of deliberation in attitude change: An empirical assessment of three theoretical mechanisms. *European Union Politics* 15(3): 389–409.
Suiter, J., Farrell, D.M. & O'Malley, E. (2014). When do deliberative citizens change their opinions? Evidence from the Irish Citizens' Assembly. *International Political Science Review* 37(2): 198–212.
Sunstein, C. (2002). The law of group polarization. *Journal of Political Psychology* 10: 175–195.
Sunstein, C. (2006). *Infotopia: How many minds produce knowledge*. New York: Oxford University Press.
Sunstein, C. (2009). *Going to extremes: How like minds unite and divide*. New York: Oxford University Press.
Taber, C. & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science* 50(3): 755–769.
Tausczik, Y.R. & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1): 24–54.
Uslaner, E. (2001). Producing and consuming trust. *Political Science Quarterly* 115: 569–590.
Walter, H. (2012). Social cognitive neuroscience of empathy: Concepts, circuits and genes. *Emotion Review* 4(1): 9–17.
Wyss, D. & Beste, S. (2014). Cognitive Complexity in a Deliberative Experiment. *Paper presented at the ECPR General Conference, Glasgow*.
Wojcieszak, M. (2011). Deliberation and attitude polarization. *Journal of Communication* 61: 596–617.
Zaller, J.R. (1992). *The nature and origins of mass opinion*. Cambridge: Cambridge University Press.

Address for correspondence: Marina Lindell, Social Science Research Institute, Åbo Akademi University, Asa A4, 20500 Åbo, Finland. E-mail: marina.lindell@abo.fi