American Psychological Association

Journal of Experimental Psychology: General

0096-3445/18/$12.00

2018, Vol. 147, No. 9, 1382–1399
http://dx.doi.org/10.1037/xge0000489

False Beliefs and Confabulation Can Lead to Lasting Changes in Political Attitudes

Thomas Strandberg, David Sivén, and Lars Hall
Lund University

Petter Johansson
Lund University and Swedish Collegium for Advanced Study

Philip Pärnamets
Lund University and Karolinska Institute

In times of increasing polarization and political acrimony, fueled by distrust of government and media disinformation, it is ever more important to understand the cognitive mechanisms behind political attitude change. In two experiments, we present evidence that false beliefs about one's own prior attitudes and confabulatory reasoning can lead to lasting changes in political attitudes. In Experiment 1 ($N = 140$), participants stated their opinions about salient political issues, and using the Choice Blindness Paradigm we covertly altered some of their responses to indicate an opposite position. In the first condition, we asked the participants to immediately verify the manipulated responses, and in the second, we also asked them to provide underlying arguments behind their attitudes. Only half of the manipulations were corrected by the participants. To measure lasting attitude change, we asked the participants to rate the same issues again later in the experiment, as well as one week after the first session. Participants in both conditions exhibited lasting shifts in attitudes, but the effect was considerably larger in the group that confabulated supporting arguments. We fully replicated these findings in Experiment 2 ($N = 232$). In addition, we found that participants' analytical skill correlated with their correction of the manipulation, whereas political involvement did not. This study contributes to the understanding of how confabulatory reasoning and self-perceptive processes can interact in lasting attitude change. It also highlights how political expressions can be both stable in the context of everyday life, yet flexible when argumentative processes are engaged.

Keywords: attitude change, confabulation, false beliefs, political psychology, reasoning

Supplemental materials: http://dx.doi.org/10.1037/xge0000489.supp

In an increasingly polarized political landscape, as exemplified by the dramatic U.K. decision to leave the European Union and the acrimonious 2016 U.S General Election, it is ever more important to understand the sources and dynamics of political attitude change. On the one hand, social psychological experiments have indicated that political attitudes can be flexible and sensitive to contextual influences, and that these attitudes either may be constructed in the moment (Bishop, 2005; Converse, 1975, 1964; Haidt, 2001; Zaller, 1992), or easily altered by the deliberation of the respondents (Hall, Johansson, & Strandberg, 2012; Hall et al., 2013). This perspective has long prompted a concern about the power of corporate capital and the political elite to shape the public

Thomas Strandberg, David Sivén, and Lars Hall, Lund University Cognitive Science, Lund University; Petter Johansson, Lund University Cognitive Science, Lund University, and Swedish Collegium for Advanced Study; Philip Pärnamets, Lund University Cognitive Science, Lund University, and Division of Psychology, Department of Clinical Neuroscience, Karolinska Institute.

Author contributions was as follows: Thomas Strandberg, Philip Pärnamets, and Petter Johansson developed the study concept. All authors contributed to the study design. Testing and data collection were performed by Thomas Strandberg, David Sivén, and Petter Johansson, Philip Pärnamets, Thomas Strandberg, and David Sivén performed the data analysis. Thomas Strandberg, David Sivén, and Philip Pärnamets drafted the manuscript, Petter Johansson and Lars Hall provided critical revisions. Philip Pärnamets and Petter Johansson jointly supervised the project. All authors approved the final version of the manuscript for submission.

Results reported as Experiment 1 in this article were partially presented at the 38th Annual Conference of the Cognitive Science Society in Philadelphia, PA, and as part of the conference proceedings. This article supersedes previous dissemination. We thank Anders Lindén at Andvision for implementing the design of the STS and Rosanna Ljungren for assisting with the data collection in Experiment 2. We are also grateful to Melissa M. Kibbe for editing the article. The data presented in this article and a complete list of target political statements used in the study (translated from Swedish to English) can be found at the Open Science Framework: https://osf.io/x5cmq/.

Correspondence concerning this article should be addressed to Thomas Strandberg, Lund University Cognitive Science, Lund University, Box 192, S-221 00, Lund, Sweden or to Philip Pärnamets, Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Solnavägen 1, 17177 Solna, Sweden. E-mail: thomas.strandberg@lucs.lu.se or philip.parnamets@ki.se

agenda (Bullock, 2011; Burke, 1774). More recently, it has led to a common recognition of the malicious persuasive potential of fake news spreading through social networks and media outlets (McNair, 2017). On the other hand, longitudinal studies have demonstrated a remarkable stability in political attitudes over the life span, and traced their genesis to developmental context and personality traits (Gerber, Huber, Doherty, & Dowling, 2011; Hatemi et al., 2009; Hooghe & Wilkenfeld, 2008; Lewis, 2018). One large-scale study found that partisan affiliation remained unchanged when measured over the course of almost four decades (Sears & Funk, 1990). They also found that only a minority of the individual attitudes fluctuated, and that these fluctuations occurred in incremental and consistent ways (see also Alwin, 1994; Sears, 1983). Similarly, much work within political science has underlined stability and resistance to change as central characteristics of political attitudes (Bartels, 2002). In light of this, when a recent study of door to door canvassing showed how 10 min of induced perspective taking could change participants' attitudes toward transgender persons (Broockman & Kalla, 2016), it was widely seen as a political sensation (Ledford, 2016).

But how can these differing perspectives, one focusing on attitude stability and the other on attitude flexibility, be reconciled? Here we use the choice blindness paradigm (CBP) to contribute to these questions. In the original CBP study (Johansson, Hall, Sikström, & Olsson, 2005), participants decided which face they found most attractive in a pair, but sometimes the opposite alternative was presented as their actual choice. The results showed that participants often failed to notice these manipulations, and instead accepted the false feedback as their preferred choice. In addition, participants readily gave verbal explanations of why they preferred the manipulated outcome, thus confabulating reasons for a choice they did not make. These results indicated a striking dissociation between the act of making a choice and its later justification and highlight the perils of assuming infallible self-knowledge about preferences, as is common in cognitive and economic models of decision making (Johansson et al., 2005).

The CBP, and its underlying methodology of creating dissociations between action and outcome, has since been widely replicated in a variety of different domains. These include taste preferences in a supermarket setting (Hall, Johansson, Tärning, Sikström, & Deutgen, 2010), financial decisions (McLaughlin & Somerville, 2013), eye-witness testimony (Sagana, Sauerland, & Merckelbach, 2016), haptic feedback (Steenfeldt-Kristensen, & Thornton, 2013), and speech intentions (Lind, Hall, Breidegard, Balkenius, & Johansson, 2014). Recent work has also demonstrated interesting downstream effects of accepting the false feedback in the CBP, both on later memories for past choices (Pärnamets, Hall, & Johansson, 2015), and for later preferences themselves (Johansson, Hall, Tärning, Sikström, & Chater, 2014; Luo & Yu, 2016; Taya, Gupta, Farber, & Mullette-Gillman, 2014). In these latter experiments, not only are the participants' ratings of alternatives influenced, but also their later choices so that they become more likely to choose an alternative they previously received false feedback about choosing.

The format of the decisions in the CBP, which includes both deliberation and explanation, makes it well suited for application to political attitudes, where this type of explicit reasoning often is highlighted as an important ideal (Anand & Krosnick, 2003; Druckman, 2004; Taber & Lodge, 2013). In previous work, we have demonstrated that salient moral (Hall et al., 2012) and political attitudes (Hall et al., 2013) are susceptible to false feedback manipulations. In these studies, participants' responses were reversed to indicate the opposite of what they had answered, and more than half of these manipulated responses were accepted by the participants as being their original attitudes. Yet, it is unclear whether CBP can induce lasting attitude change, as the participants in these studies were debriefed about the false feedback soon after the study and were reacquainted with their original answers. Using faces as stimuli, Taya et al. (2014) found preference change resulting from the false feedback in the short-term, but no effect when measured a week later. However, the influence of the false feedback in Hall et al. (2013, 2012) was considerable, and it is likely it might have been sustained if the debrief had been postponed and the participants queried at a later time. Thus, the first aim of the current study is to investigate whether false feedback about one's own survey responses can result in lasting change to one's political attitudes.

Second, if this is the case, what might the mechanisms be? In a classic study, Janis and King (1954) used role playing as a manipulation and had participants actively arguing for hypothetical future events, such as an estimation of the amount of movie theaters still open in three years' time. They found that participants who expressed verbal arguments in favor of an estimate were more likely to change their attitude to correspond with it, compared with a passive control group that did not verbally engage with the issue. They also found that participants in the experimental condition reported a higher confidence in their attitude. Similar kinds of attitude change have also been reported for groups, for example, when groups' jointly decided attitudes toward specific issues were rated as more extreme compared with the mean original rating of each individual (Kogan & Wallach, 1967). In particular, the attitudes of actively discussing groups changed more compared with groups that only listened to recordings of another group's discussion (Kogan & Wallach, 1967; Isenberg, 1986). In another more recent line of work, Clarkson, Tormala, and Leone (2011) found that if participants get to think about an object for up to 300 s compared with 60 s, their confidence regarding their own attitudes directed at this object was increased and their attitudes became more extreme. In Barden and Tormala (2014), participants' attitude strength was similarly influenced by how they experienced their own arguments: the more arguments the participants expressed in favor of a cause, the stronger their proattitude for that cause became. These findings illustrate that the perception and verbalization of one's own reasoning processes can largely impact one's attitudes (Knowles & Linn, 2004; Tormala & Petty, 2002).

Reasoning is a core element in the CBP, because participants are asked to verbally explain their (putative) choice (Johansson, Hall, Sikström, Tärning, & Lind, 2006). What is interesting is that we can be certain that these explanations are confabulatory, because the participants give reasons for a choice they in fact did not make (Johansson et al., 2005). The majority of previous research on confabulation has described it as a clinical spectrum disorder (Fotopoulou, Conway, & Solms, 2007; Hirstein, 2009). Confabulation has also been implicated in (false) memory formation (Bernstein, Laney, Morris, & Loftus, 2005; Loftus & Zanni, 1975), and there are indications that it might be prevalent in typical peoples' everyday lives (French, Garry, & Loftus, 2009). This possibility is strengthened by the lack of semantic and emotional differences

found in CBP contrast analysis between the nonmanipulated and manipulated verbal reports (for detailed analyses of such reports, see Johansson et al., 2005, 2006 and Hall et al., 2012). Potentially, the process behind all introspective reports might be confabulatory at its core (Dennett, 1987). However, without a wedge like CBP to get between the decisions of the participants and their reports, it is difficult to question the subjective authority of the participants. Consequently, the impact of confabulatory reasoning on attitude change has not been studied at all. Because confabulatory reasoning has been found to strengthen false beliefs, and because depth of reasoning in general can influence attitudes, we hypothesized that the amount of confabulation a participant engages in when justifying a false feedback response will increase the self-induced attitude change, as well as its persistence over time.

To investigate this as well as the longevity of attitude change following false feedback, we conducted two experiments. In Experiment 1 our participants filled out a political attitude survey on several specific political issues in the areas of health care, education, and environment. They then received false feedback about some of their responses to these issues (see Figure 1). Half of the participants were assigned to the *Acknowledge* condition, and asked to merely acknowledge their responses, whereas the other half was assigned to the *Confabulation* condition and asked to give verbal explanations behind some of their responses. We then asked participants to state their attitudes to the same issues a second time, a few minutes after having been confronted with the false feedback. Participants were also invited to a third attitude survey one week later. In Experiment 2 we sought to replicate the findings of Experiment 1, as well as adding additional measures to investigate some possible moderators of the reported effects.

Experiment 1

Method

Participants. We recruited a total of 150 participants (91 female), with an average age of 22.7 years (SD = 3.0), at Lund University campus. Ten participants were excluded from the final analysis: of these four participants did not show up for the second session, and six experienced a malfunction with the experimental apparatus. One hundred forty participants were included in the final analysis. Participants received two cinema vouchers in exchange for their participation in two experimental sessions, roughly one week apart (average 6.3 days (SD = 1.8)). At the start of the experiment, we described the general purpose and the outline of the experiment, but without telling the participants that some of their answers would be manipulated. We also informed the participants that they could quit the experiment at any time and request their data to be erased. All participants were fully debriefed after the second follow-up of the experiment, before consenting to their anonymized data to be used by signing a consent form. The participants that did not show up for the second follow-up were

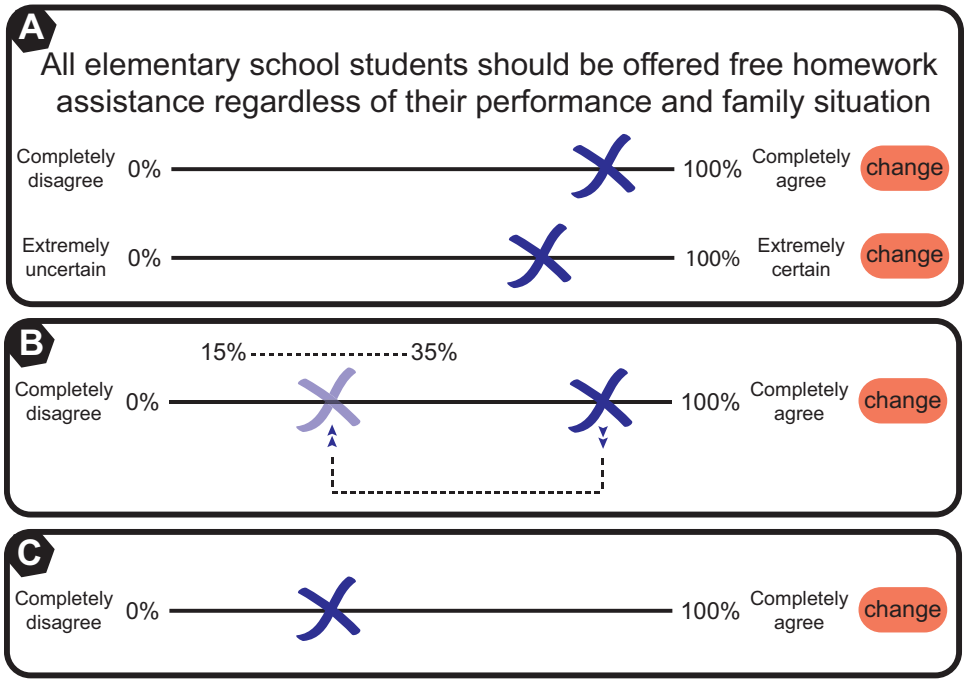

Figure 1. Manipulation. Participants rate to what extent they agree with a political statement as well as their level of confidence on a visual-analog scale ranging from 0% to 100% (A). After responding to all 12 statements, participants are asked to go over four of the responses together with the experimenter. At this stage, the application has moved two of their responses to the opposite side of the scale. The manipulation moves the responses across the midline and randomly place them between 15% and 35%, or 65% and 85% (B). In the acknowledge condition, participants are asked to just verify their responses. In the confabulation condition, they are also asked to explain the reasons behind each response (C). Participants can always change a response by clicking the change button (A–C). See the online article for the color version of this figure.

debriefed over the telephone. The experiment was approved by the Lund University Ethics board, D.nr. 2008–2435.

Materials and design. Three questionnaires were administered during the experiment. One questionnaire was a tablet application specifically developed for giving participants false feedback about their survey ratings, the Self Transforming Survey. It was developed in the programming language Python with Django framework as a backend on the server side. The front end was coded in HTML, CSS bootstrap, and the dynamical functionality in Javascript with the help of JQuery library. The remaining two surveys were regular pen and paper surveys. Further, an audio recorder was used to capture the verbal reports given by the participants.

The political statements were divided into three categories: health care, education, and environment. Six of the statements were used in all three questionnaires. Of these six, four were target statements that were randomly assigned as either manipulated or nonmanipulated, taken from the environment and education categories. All statements concerned salient political topics in Sweden at the time of the experiment and were constructed to state a proposed policy and give a brief explanation of that policy. One example of a target statement:

> The Swedish elementary school should be re-nationalized. Local municipalities would then lose some influence, and the state would become head of the school and assume the responsibility for resource allocation and quality assurance. [See OSF repository for complete list of statements.]

Procedure. The experiment consisted of three sessions: initial rating and interaction with the manipulated and nonmanipulated responses (T1); a second rating session following their interaction with the experimenter and their initial ratings (T2); and a third rating session around one week later to measure lasting attitude change (T3). The participants were randomly assigned to one of two conditions: Acknowledge or Confabulation.

The experiment proceeded as follows: the participants were recruited from the common areas of a university building and asked whether they would be willing to answer a political questionnaire. If they accepted, participants were brought to a separate room, seated in front of a tablet, and explained the general outline of the procedure, but without mentioning the false feedback. The questionnaire ran on The Self-Transforming Survey (STS), a tablet application specifically developed for giving participants false feedback about their survey ratings. The questionnaire contained 12 political statements, presented one at a time, and the participants' task was to rate to what extent they agreed or disagreed with each statement by drawing a mark on a visual-analogue scale with end-points anchored at *completely disagree* to *completely agree*. Below each statement they also estimated how confident they felt about their attitude, on a similar scale but with endpoints going from *extremely uncertain* to *extremely certain* (Figure 1A). The participants were left to answer the questionnaire at their own pace. The attitude ratings obtained during this initial portion of the experiment are referred to as the T1 ratings and serve as the baseline to which later attitudes are compared.

Afterward, the experimenter reentered the room and informed the participants that the application would now randomly display four of the statements, one at a time, together with their ratings (but without the confidence rating). Here, the participants' ratings to two of the four displayed statements had been manipulated by the application (Figure 1B and 1C). The participants in both the Acknowledge and Confabulation conditions were instructed to read each displayed statement aloud, tell where on the scale their rating was, whether this implied that they agreed or disagreed with the statement, and to what extent (e.g., by saying "I agree with that to some extent"). Participants in the Confabulation condition were also instructed to explain their reasoning behind each response. After a participant had stated a position, the experimenter asked: "Why do you [to some extent] agree with that statement?" but avoided interacting with the participants while they were explaining. If a participant, for example, had questions the experimenter just mentioned that it was up to the participant to interpret the statement. Thus, all participants in the Confabulation condition received the same treatment and the experimenter was not involved in the reasoning task.

During a manipulated trial, the participants' rating was always moved across the midline of the 0–100% scale, thus shifting the participants stated attitude from agreeing to disagreeing with the statement (or vice versa). The manipulated rating was randomly placed between 15% and 35%, or between 65% and 85%, depending on the direction of the manipulation (see Figure 1). Additionally, each scale was coupled with a change button, so while filling out the survey, as well as when going over the ratings with the experimenter, the participants always had the option to change a rating should they feel that it did not reflect their attitude toward a particular issue. If the participants hesitated, or behaved like something was wrong, the experimenter informed them that they could change their response by clicking change and then draw another rating. A manipulation was automatically registered as corrected when the participants clicked the change button and drew a new rating on the scale.

After the tablet survey and the interaction with the four target statements was finished the participants were asked to fill out another questionnaire, this time on paper. These ratings are referred to as T2 ratings. The questionnaire also contained 12 political statements: six from the first questionnaire, including the two manipulated and the two nonmanipulated statements, as well as six new statements. The participants were told that it was possible that some of the statements that they had already responded to on the application might reappear, because they were all randomly drawn from the same bank of statements.

The participants were scheduled to return in one week for the second follow-up, which took place on average 6.3 days (SD = 1.8) later. These are referred to as T3 ratings. In this follow-up, the participants answered another paper survey containing 12 political statements, including the same six statements from the previous questionnaires (two manipulated, two nonmanipulated, and two filler statements) mixed with six new statements. Finally, the participants were debriefed in full, and signed data release statements.

Analysis. All ratings were converted to a 0–100 mm scale to facilitate comparisons between mediums (i.e., STS (T1) and paper-pen (T2 and T3)). For our analyses we used the ratings in two ways, outlined here.

First, we investigated whether *attitude strength* at T1 predicts correction in the task. To simplify the analysis, we converted the attitude ratings to a 0–50 scale. This was done by centering the scale, so it ranged from −50 to +50 and then used the absolute

resulting values. Thus, a rating of 0 (maximum disagree) and a rating of 100 (maximum agree) would both correspond to an attitude strength of 50 (maximum strength). A rating of 50 (no opinion or undecided) would be 0 on the attitude strength scale.

Second, for the main dependent measure, *attitude change*, we wanted to analyze changes to the participants' stated attitudes over time. To do this, we first needed to realign the attitude ratings, to make them comparable regardless of whether the participants agreed or disagreed with the statements. This was done at all time-steps of the experiment. We then used the realigned ratings to measure the difference between the original attitude (T1) and later attitudes (at T2 and T3). Both steps are described below.

Participants' ratings on the 0–100 mm scale were numerically realigned to facilitate comparison between participants who would otherwise have opposing opinions on an issue. For statements where the participants' T1 ratings were under the midline of the scale (<50), all ratings from that participant to that statement were flipped over the midline. For example, if the participants responded 25 at T1, 60 at T2, and 30 at T3 to some statement, these values were recoded to 75 at T1, 40 at T2, and 70 at T3. For statements where the participants' T1 ratings were over the midline of the scale (>50), no changes were made. All participants′ ratings at all time-steps of the experiment are shown on the same directional scale.

Because our main hypotheses concerned attitude change, the T2 and T3 ratings were analyzed as differences compared with the original T1 rating. A negative difference represents a movement in the attitude toward or beyond the midline, and for manipulated trials, in the direction of the false feedback. Referring back to our earlier example, if the participant's realigned rating at T1 was 75 and the rating at T2 was 40, this represents an attitude change score of −35. We refer to such changes as a *weakening* of the attitude. Conversely, if the participant's rating at T1 was 75 but the rating at T2 had been 80, this represents an attitude change score of +5, and is described as a *strengthening of the attitude.*

We analyzed our data using (generalized) linear mixed-effects models using the *lme4* package in *R*. Random-effects were modeled as per participant intercepts and slopes mirroring the full fixed-effects structure, or the maximally permitted structure that would converge (Bates, Maechler, Bolker, & Walker, 2015). Significance of fixed-effects was assessed using Wald chi-square tests as implemented in the *car* package (Fox & Weisberg, 2011). We report marginal model R^2 for the fitted models, describing the proportion variance explained by the fixed-factors, using the *piecewiseSEM* package (Lefcheck, 2016), which is a variance explained measure specific for mixed-effects models. For interpretation of effects we report unstandardized beta coefficients from our analyses and their standard errors, which can be interpreted on the 0–100 mm scale.

Results

Correction of manipulated responses. Of the 277 manipulated (M) trials, 134 (48.4%) were corrected by the participants, meaning 51.6% were accepted. Average by participant correction rate was 1.0 trials (SD = 0.8). Forty-five (32%) participants made no corrections, 56 (40%) made one correction, and 39 (28%) made two corrections. All participants and trials were included in the analyses.

Effects of confidence, attitude strength, and condition on correction. In the Confabulation condition, participants corrected 53.3% of manipulations, whereas participants in the Acknowledge condition corrected 43.6% of manipulations. Average attitude strength was M = 23.1, SD = 14, on a 0–50 scale where 0 represents the indifference point. Next to each political statement, the participants also rated how confident they felt about their response. Average confidence was high with an average of 63 of 100 (SD = 23). Confidence was higher for Corrected trials (M = 70, SD = 22) than for Accepted trials (M = 56, SD = 23; Welch t test $t(191.77) = 5.88$, $p = 1.78^{*}10^{-8}$). Confidence was highly correlated with attitude strength, r = .64, 95% CI [.59, .69].

We analyzed the effects of confidence, attitude strength, and confabulation condition on the probability of correcting the manipulation. Both confidence and attitude strength were standardized prior to analysis to aid model convergence, whereas condition was deviation coded (Confabulation = 0.5). We found a significant interaction between confidence and attitude strength, $\chi^2_{(1)} = 8.09$, p = .0044, but no other significant effects, with marginal model R^2 = .246. The regression coefficients of confidence and attitude strength were all positive, indicating that participants were most likely to correct attitudes which were both extreme and confidently held (see Table 1).

Effect of manipulation and correction on future ratings. We tested the effect of the false feedback during the two follow-up surveys (T2 and T3) in two regressions. In the first, we regressed attitude change on manipulated versus nonmanipulated trials together with an interaction with time. All variables were dummy coded taking T2, nonmanipulated trials as reference levels. In the

Table 1
All Estimated Regression Coefficients and Their Standard Error for Mixed-Model Analysis of Correction

Effect	Estimate	Standard error	Wald χ^2 (df = 1)	p value
Intercept	−.51	.26	—	—
Confidence	1.05	.33	3.48	.062
Attitude strength	.13	.27	1.28	.26
Condition	.38	.50	2.06	.151
Confidence × Attitude strength	.65	.22	8.09	.0044
Confidence × Condition	.11	.55	.034	.86
Attitude strength × Condition	.10	.54	.25	.61
Confidence × Attitude strength × Condition	.40	.40	1.02	.31

Note. For all predictors Wald chi-square and p values are also reported.

second, we regressed attitude change on accepted versus corrected *manipulated* trials, disregarding nonmanipulated trials, together with an interaction with time. All variables were dummy coded taking corrected trials at T2 as reference levels. We report each regression in turn.

The first regression tested whether attitude change differed on average between manipulated (M) and nonmanipulated (NM) trials. We found significant main effect of Manipulation, $\chi^2_{(1)} = 39.23$, $p = 3.7^{*}10^{-10}$, as well as a significant interaction between Time and Manipulation, $\chi^2_{(1)} = 31.64$, $p = 1.9^{*}10^{-8}$, but no main effect of Time, $\chi^2_{(1)} = 2.41$, $p = .12$, with model marginal $R^2 = .09$. Interpreting the coefficients, participants were highly accurate in restating their original attitude in T2 during nonmanipulated (NM) trials ($b_{intercept} = -1.1$ mm, $SE = 0.9$), and this changed little from T2 to T3 ($b_{T3} = -1.2$ mm, $SE = 1.2$). There was a large weakening of attitudes at T2 for manipulated (M) trials ($b_M = -12.8$ mm, $SE = 1.6$), which was attenuated at T3 ($b_{T3^{*}M} = 8.2$ mm, $SE = 1.7$).

We additionally examined whether initial confidence predicted later attitude shifts, by comparing the model fitted above, with one including an additional standardized confidence term and all interactions with Manipulation and Time. However, including the confidence term did not significantly improve fit, $\chi^2_{(3)} = 6.17$, $p = .09$, and the fitted coefficients of confidence indicated that any effects were negligibly small ($b_{Conf} = 0.5$ mm, $SE = 1.0$; $b_{Conf^{*}M} = 2.0$ mm, $SE = 1.4$; $b_{Conf^{*}T3} = 0.4$ mm, $SE = 1.3$; $b_{Conf^{*}T3^{*}M} = -2.5$ mm, $SE = 1.9$).

The second regression contrasted accepted (A) and corrected (C) manipulated trials, subsetting the data to only include manipulated trials. We found a significant main effect of Correction, $\chi^2_{(1)} = 98.52$, $p = 2.2^{*}10^{-16}$, and of Time, $\chi^2_{(1)} = 33.09$, $p = 8.79^{*}10^{-9}$, as well as a significant interaction between Time and Correction, $\chi^2_{(1)} = 11.21$, $p = .00082$, with model marginal $R^2 = .24$. Interpreting the coefficients, participants displayed virtually no directional change in attitudes at T2 during corrected trials ($b_{intercept} = -2.5$ mm, $SE = 1.3$), and this changed little from T2 to T3 ($b_{T3} = 2.8$ mm, $SE = 1.8$). Consistent with our hypotheses, we found a large weakening of attitudes in T2 for accepted (A) trials ($b_A = -21.6$ mm, $SE = 2.2$), an effect that was attenuated at T3 ($b_{T3^{*}A} = 8.3$ mm, $SE = 2.5$). To summarize: we found evidence of directional attitude change following from accepted but not for corrected false feedback trials. The effects where largest at T2 but remained robust at T3.

Qualitative shifts in position. Given the changes in ratings at T2 and T3, we examined the proportion of the trials that crossed the midline of the attitude scale, indicating a qualitative shift compared with the original T1 attitude. At T2, 73% of responses represented such a shift for Accepted trials, compared with 10% for Corrected trials and 11% for Non-Manipulated trials. At T3, where the attitudinal effects of the manipulation were attenuated, 41% of responses were still qualitatively shifted for Accepted trials compared with 10% for Corrected trials and 12% for Non-Manipulated trials.

Effect of confabulation on future ratings. We investigated the effect of Confabulation condition (dummy coded with the acknowledge condition as reference level), on subsequent attitude change. We first analyzed all trials, following the same analytical strategy as above, contrasting manipulated and nonmanipulated trials including interactions with Time and Confabulation condition. We found no main effect of Confabulation, $\chi^2_{(1)} = 0.0082$, $p = .93$; $b_{CONFAB} = -0.1$ mm, $SE = 1.9$, nor any interaction with Manipulation, $\chi^2_{(1)} = 1.42$, $p = .23$; $b_{M^{*}CONFAB} = -3.0$ mm, $SE = 3.1$, Time, $\chi^2_{(1)} = 0.83$, $p = .36$; $b_{T3^{*}CONFAB} = -1.7$ mm, $SE = 2.4$, or three-way interaction, $\chi^2_{(1)} = 0.01$, $p = .92$; $b_{M^{*}T3^{*}CONFAB} = -0.4$ mm, $SE = 3.4$ (see also Figure 2A and 2B). This shows that participants' attitude stability in general was not affected by the method of restating their attitudes. The remainder of the analysis yielded coefficients consistent with previous results (see Supplemental results).

Previously we showed that attitude change was only present for accepted manipulated trials. Therefore, we again subset the data on manipulated trials and contrasted corrected (C) and accepted (A) trials, including interactions with Time and Confabulation condition. We found that participants displayed no directional attitude change in T2, corrected trials in the acknowledge ($b_{intercept} = -2.9$ mm, $SE = 2.2$) or confabulation conditions ($b_{CONFAB} = -0.2$ mm, $SE = 3.1$; see Figure 2C), with similar results for T3 trials ($b_{T3} = -1.9$ mm, $SE = 2.8$; see Figure 2D). There was a large directional attitude change for the accepted trials ($b_A = -16.7$ mm, $SE = 2.8$; $\chi^2_{(1)} = 124.14$, $p < 2.2^{*}10^{-16}$). Importantly, in line with this we found main effects of Condition, $\chi^2_{(1)} = 4.81$, $p = .028$, and Time, $\chi^2_{(1)} = 30.74$, $p = 3.0^{*}10^{-8}$, and these were qualified by interactions between Correction and Condition, $\chi^2_{(1)} = 8.33$, $p = .0039$, and between Correction and Time, $\chi^2_{(1)} = 10.71$, $p = .0011$. Taken together, this means that the directional changes of accepted trials were, as hypothesized, enhanced in the Confabulation condition at T2, meaning a further weakening of the original attitude ($b_{A^{*}CONFAB} = -9.6$ mm, $SE = 4.0$). Attitude changes were attenuated at T3 ($b_{A^{*}T3} = 7.7$ mm, $SE = 3.6$). The interaction between Condition and Time, $\chi^2_{(1)} = 0.77$, $p = .38$; $b_{CONFAB^{*}T3} = -1.5$ mm, $SE = 3.7$, and the three-way interaction were not significant, $\chi^2_{(1)} = 0.078$, $p = .78$; $b_{A^{*}CONFAB^{*}T3} = -1.4$ mm, $SE = 5.1$. Model conditional $R^2 = .26$.

Summary of Experiment 1. We investigated whether false beliefs about one's own political attitudes, and confabulatory reasoning, could lead to lasting changes in these attitudes. We gave participants false feedback about some of their responses on a political survey, and asked half of them to merely acknowledge their responses, and the other half to also give verbal explanations to their responses. As expected, about half of the manipulations were accepted by the participants as being their own responses. Participants' future attitudes were strongly influenced by the false feedback, both directly following the manipulation and one week later. Additionally, we found that the attitude change was considerably larger if participants were asked to verbalize arguments, compared with only acknowledging its position.

Experiment 2

Experiment 2 was conducted with two aims in mind. The first was to run a high-powered direct replication of the findings in Experiment 1. The second was to investigate some possible factors that could moderate acceptance of the manipulation and the attitude change observed in Experiment 1. These factors are introduced below.

Our main finding in Experiment 1 was that attitude change is greater following confabulatory reasoning during the false feedback as compared with when only acknowledging the manipulated

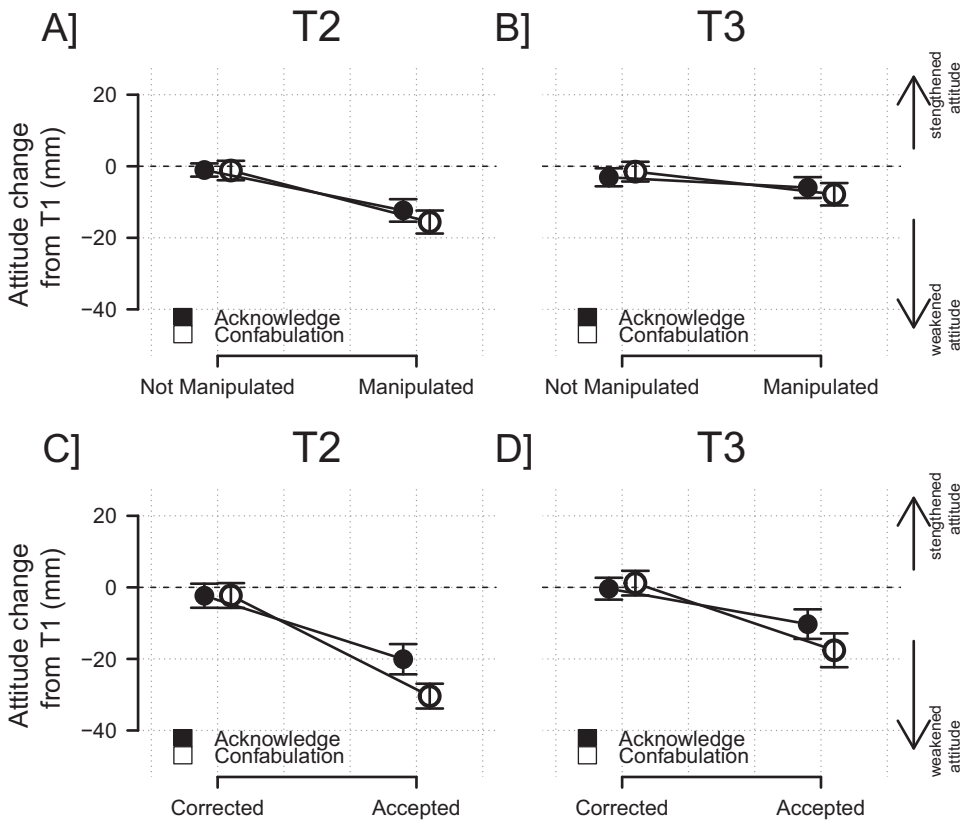

Figure 2. Attitude change. Average attitude change compared with original (T1) ratings. A negative difference indicates a weakening of the original attitude. For manipulated trials this always means a change in direction toward the rating indicated by the false feedback. (A–B) Attitude change in T2 (A) and T3 (B) for Non-Manipulated and Manipulated trials split by Confabulation condition. (C–D) Attitude change for Manipulated trials only. Difference shown in T2 (C) and T3 (D) for Corrected and Accepted trials, split by Confabulation condition. Error bars are 95% CI.

answer. One question that arises from this concerns what relation participants' confabulation stands to their later attitude change. One possibility is that merely engaging in the production of reasons gives an encoding advantage to the new attitude, leading to a greater shift in the participant's attitude. Alternatively, participants' attitude change might reflect a gradual depth of processing, as could be seen in the quantity of arguments given for the false feedback attitude. One simple unobtrusive measure is the amount of time participants spend engaging with the false feedback before answering the next question. If the magnitude of the participants' confabulatory argumentation is helping them cement their new attitude, we should expect the size of attitude change to be positively correlated with the length in time of their confabulatory engagement. To test this, we measured participants' talking time during the false feedback phase of the experiment.

A dominant view in much recent theorizing about information processing and reasoning, particularly in the political domain, has been that it is susceptible to the influence from strong motivational forces (Jost & Amodio, 2012; Kunda, 1990, 1987; Taber, Lodge, & Glathar, 2001). On this view, implicit motives, such as the need to be right about an issue, or to behave according to one's ideological values, can shape the interpretation of political information and the construction of reasons for having a belief (Jost & Amodio, 2012). This type of inferred justification strategy is supposedly used when there is a discrepancy between a belief and the external evidence contradicting the basis of that belief, and may help explain how people evaluate facts (Ditto & Lopez, 1992) and why some people label news as fake if they come from media houses with a political agenda opposite to their own (Flynn, Nyhan, & Reifler, 2017). In our study, participants faced a dilemma of sorts when viewed through a motivational lens. On the one hand, they should be motivated to defend their initial political attitudes which will, by definition, conflict with the false feedback. On the other, they should be motivated to defend their stated attitude, that is, whatever is presented to them as being their own attitude. To investigate the impact of global political beliefs on level of acceptance and attitude change, we therefore included a general measure of political involvement and a left- to right-wing ideology scale.

Recently, motivated cognition in politics has also been related to peoples' cognitive style. One common measure is the Cognitive Reflection Test (Frederick, 2005), which is hypothesized to capture individual differences in reflexivity and critical reasoning (Bialek & Pennycook, 2017; Pennycook & Ross, 2016). Kahan (2013) found that high Cognitive Reflection Test (CRT) scores

associated with greater propensity to engage in politically motivated reasoning. Similarly, higher CRT scores were also found to predict the ability to discern fake news (Pennycook & Rand, 2017). Although the false feedback presented to participants in our experiments is not exactly "fake news," it is counterfactual and runs against their prior attitudes. Hence, we can expect that higher CRT scores should correlate with correcting the false feedback.

In sum, we attempted a direct replication of our findings from Experiment 1, adding measures of confabulatory reasoning, political attitudes, and a CRT task.

Method

Participants. We recruited a total of 264 participants based on prior power calculations indicating that 240 participants would give high power to detect the crucial Correction and Confabulation condition interactions (>95%). Power was calculated based on the regression coefficients for the model estimated in Experiment 1 including Confabulation condition and Correction as factors analyzing attitude change for manipulated trials. We simulated data based on the estimated random and fixed effects, as well as the correction rates observed in Experiment 1 (Gelman & Hill, 2007). Thirty-two participants failed to show up for the T3 measurement or experienced equipment malfunction. The final sample therefore consisted of 232 participants (146 male, 85 female, one not identified), with an age range of 18–52 (M = 23.6, SD = 4.6).

Participants received two cinema vouchers in exchange for their participation in two experimental sessions, roughly one week apart (average 6.8 days [SD = 0.9]). Participant information and debriefing followed the procedures described for Experiment 1. The experiment was approved by the Lund University Ethics board, D.nr. 2016–1046.

Materials and design. The choice blindness and attitude change setups were identical to Experiment 1 (a combination of the STS and paper-pen surveys), including the political statements. CRT, political involvement, and left–right ideology were assessed on additional paper surveys. CRT consisted of the following questions, presented on separate pages: "(1) A bat and a ball costs $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?" [answer in cents] "(2) If it takes 5 machines 5 min to make 5 widgets, how long would it take 100 machines to make 100 widgets?" [answer in minutes] "(3) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?" [answer in days].[1] Political involvement was assessed with the following items: "(1) In your daily life, how engaged in political issues would you say that you are?" "(2) Are you engaged in any of the following: (a) political party, (b) environmental organization (such as Greenpeace; c) school organization (such as a teacher association)?" [yes/no]. Left–Right ideology was assessed using a scale with endpoints going from left to right. Further, participants stated their education level, education subject, age and gender. The political-, educational-, and demographical items were assessed on the final page. Just as in Experiment 1, a tape recorder was used to capture the verbal reports, and a timer was used to clock speaking time.

Procedure. Experiment 2 followed exactly the same procedure as Experiment 1, with two extensions. First, questionnaires measuring CRT, political involvement, and ideology were administered during the final session (T3), after the participant had completed the political attitude surveys, but before the debriefing. Second, the experimenter timed the participants' argumentation/confabulation using a timer on the computer. This way, the additional measures in Experiment 2 were unobtrusive and did not interfere with the direct replication of Experiment 1.

Analysis. We followed the same analytical strategy as for Experiment 1 with two additions. First, we also estimated random effects (intercept and slopes) grouped by stimulus ID to improve the generalizability of our estimates. Again, random effects were entered as maximal or the maximal that would converge. Second, to provide combined estimates of the effects from both experiment, we conducted an analysis of our main findings on the combined dataset using Bayesian estimation techniques of the maximal multilevel model using the *brms* package (Buerkner, 2016). For information about priors, see the online supplemental material.

Results

Correction of manipulated responses. Of the 464 manipulated (M) trials, 234 (50.4%) were corrected by the participants, meaning 49.6% were accepted. Average by participant correction rate was 1.0 trials (SD = 0.8). Sixty-eight (29%) participants made no corrections, 94 (41%) made one correction, and 70 (30%) made two corrections. All participants and trials were included in the analyses.

Predictors of correction. Participants corrected 55.6% of manipulations in the Confabulation condition, whereas participants in the Acknowledge condition corrected 45.7% of manipulations. Average attitude strength was M = 26.4, SD = 15, on a 0–50 scale where 0 represents the indifference point. Participants also rated how confident they felt about each response. Average confidence was high with an average of 68 of 100 (SD = 25). Confidence was higher for Corrected trials (M = 77, SD = 20) than for Accepted trials (M = 58, SD = 24; Welch t test t[307.22] = 8.66, p = $2.73^{*}10^{-16}$). Confidence was highly correlated with attitude strength, r = .71, 95% CI [.68, .74].

We analyzed the effects of nine possible predictors on the probability of correcting the manipulation, three were the same as analyzed in Experiment 1: Confidence, Attitude strength, and Confabulation condition. Six were added in Experiment 2: participant political involvement, membership in political party, environmental organization or school organization, left–right political attitude, and CRT score. Average political involvement was fairly high, 51 of 100 (SD = 21). Membership in organizations was low: 7.8% of participants were members of a political party, 8.7% of an environmental organization, and 5.2% of a school organization. Average political attitude on a left-right scale, where 0 is extreme left, 50 is neutral, and 100 is extreme right, was M = 35, SD = 22. For CRT we sampled an even distribution of scores; 32% of participants answered zero questions correct, 28% one question, 20% two questions, and 20% all three questions correct. The average score was M = 1.3.

All variables were entered in a multilevel regression model together with the interaction between Confidence and Attitude strength. All continuous variables were standardized, except CRT

[1] The CRT problems were Swedish translations of the questions used in Frederick (2005).

score which was mean centered. Organization membership variables were also mean centered, with positive values indicating membership. Confabulation condition was coded (−.5 = Acknowledge, .5 = Confabulation). We found four significant predictors of correction. Participants' CRT scores, $\chi^2_{(1)} = 7.76$, $p = .0054$; $b = 0.41$, $SE = 0.15$, Confidence, $\chi^2_{(1)} = 5.97$, $p = .015$; $b = 0.69$, $SE = 0.28$, and Attitude strength, $\chi^2_{(1)} = 7.84$, $p = .0051$; $b = 0.79$, $SE = 0.28$, all positively predicted increasing probabilities of correcting the false feedback. Participants' left–right attitudes negatively predicted probability of correcting the false feedback, $\chi^2_{(1)} = 7.22$, $p = .0072$; $b = -0.45$, $SE = 0.17$, meaning that highly left-leaning participants made more corrections compared with other participants (see Figure S1). The remaining predictors were nonsignificant (see Table 2), and marginal Model $R^2 = .37$.

Effect of manipulation and correction on future ratings. We wanted to see whether accepted manipulated ratings would influence future ratings of the same issue. We repeated the analyses reported for Experiment 1 above. For brevity we only report the critical findings here and report the full analysis in the online supplemental materials. We replicated our findings from Experiment 1 and found once again a large weakening of original attitudes for T2 manipulated (M) trials, $\chi^2_{(1)} = 45.84$, $p = 1.29^{*}10^{-11}$; $b_M = -12.1$ mm, $SE = 1.6$, which decreased during T3, $\chi^2_{(1)} = 12.07$, $p = .00051$; $b_{T3^{*}M} = 4.8$ mm, $SE = 1.4$. Similarly, when comparing corrected and Accepted trials only, we found, consistent with our first main hypothesis and our findings in Experiment 1, a large weakening of original T2 attitudes for accepted (A) trials, $\chi^2_{(1)} = 41.45$, $p = 1.2^{*}10^{-10}$; $b_A = -20.9$ mm, $SE = 2.8$, which decreased somewhat at T3, $\chi^2_{(1)} = 14.01$, $p = .00018$; $b_{T3^{*}A} = 7.1$ mm, $SE = 1.9$.

Qualitative shifts in position. We examined the proportion of the trials that crossed the midline of the attitude spectrum, indicating a qualitative shift compared with the original T1 attitude. In T2, 67% of responses represented such a shift for Accepted trials, compared with 6% for Corrected trials and 13% for Non-Manipulated trials. In T3, where the attitudinal effects of the manipulation were attenuated, 47% of responses were still qualitatively shifted for Accepted trials compared with 8% for Corrected trials and 17% for Non-Manipulated trials. These findings mirrored those of Experiment 1.

Effect of confabulation on future ratings. Next, we analyzed the effect of confabulation condition (acknowledge or confabulation) on attitude change. We first contrasted manipulated and nonmanipulated trials (see also Figure 3A and 3B). Our findings were largely consistent with those of Experiment 1. We found no main effect of Confabulation, $\chi^2_{(1)} = 1.24$, $p = .27$; $b_{CONFAB} = 0.02$ mm, $SE = 2.3$, nor any interaction with Manipulation, $\chi^2_{(1)} = 2.97$, $p = .085$; $b_{M^{*}CONFAB} = -3.4$ mm, $SE = 3.7$, Time, $\chi^2_{(1)} = 0.00$, $p = .99$; $b_{T3^{*}\ CONFAB} = 1.1$ mm, $SE = 2.0$, or three-way interaction, $\chi^2_{(1)} = 0.50$, $p = .48$; $b_{M^{*}T3^{*}CONFAB} = -2.1$ mm, $SE = 2.9$. The remaining effects and coefficients were highly similar to those reported for Experiment 1 (see Table S1). Model marginal $R^2 = .09$.

Next, we conducted the crucial test of whether attitude change differed by Confabulation condition and Correction within the manipulated trials. Participants displayed small directional attitude change at T2, corrected trials in the acknowledge condition ($b_{intercept} = -3.6$ mm, $SE = 2.1$), and further shifted slightly more in the confabulation condition for T2, Corrected trials ($b_{CONFAB} = -1.4$ mm, $SE = 2.8$; see Figure 3C), with similar results for T3 trials ($b_{T3} = 0.6$ mm, $SE = 2.2$; see Figure 3D). For the accepted (A) trials, there was a large directional attitude change ($b_A = -16.3$ mm, $SE = 2.4$, $\chi^2_{(1)} = 63.3$, $p < 1.8^{*}10^{-15}$). The main effects of Condition, $\chi^2_{(1)} = 2.16$, $p = .14$, and Time, $\chi^2_{(1)} = 14.62$, $p = .00013$, were, again, qualified by interactions between Correction and Condition, $\chi^2_{(1)} = 4.78$, $p = .029$, and Correction and Time, $\chi^2_{(1)} = 5.04$, $p = .025$. As expected according to our second main hypothesis, and from Experiment 1, the directional changes of accepted trials were accentuated in the confabulation condition at T2, meaning a further weakening of the original attitude ($b_{A^{*}CONFAB} = -10.5$ mm, $SE = 5.6$). The attitude change was attenuated in T3 ($b_{A^{*}T3} = 8.3$ mm, $SE = 3.9$). The interaction between Condition and Time, $\chi^2_{(1)} = 0.00$, $p = .98$; $b_{CONFAB^{*}T3} = 0.8$ mm, $SE = 3.0$, and the three-way interaction, were not significant, $\chi^2_{(1)} = 0.17$, $p = .68$; $b_{A^{*}CONFAB^{*}T3} = -1.7$ mm, $SE = 4.1$. Model conditional $R^2 = .24$.

Effect of confabulation length on attitude change. In the Confabulation condition, we additionally measured how long participants took while stating reasons for the presented attitude. Confabulation Length ranged from 36 to 255 seconds, with an average of $M = 93$s, $SD = 39$s. To analyze the effects of Length on attitude change we subset the data from the Confabulation condition depending on whether the false feedback was corrected or accepted. The reason for doing so is that Length will have

Table 2
All Estimated Regression Coefficients and Their Standard Error for Mixed-Model Analysis of Correction From Experiment 2

Effect	Estimate	Standard error	Wald χ^2 ($df = 1$)	p value
Intercept	.05	.26	—	—
Political involvement	−.05	.27	.04	.84
Party member	1.32	.79	2.81	.094
Environmental org. member	1.55	1.40	1.22	.27
School org. member	.21	.80	.072	.79
Left–Right attitude	−.47	.18	7.16	.0075
CRT score	.41	.14	7.93	.0049
Confidence	.70	.27	6.84	.0089
Attitude strength	.73	.29	6.51	.011
Confabulation condition	.20	.36	.32	.57
Confidence × Attitude strength	−.10	.23	.20	.65

Note. For all predictors Wald chi-square and p values are also reported.

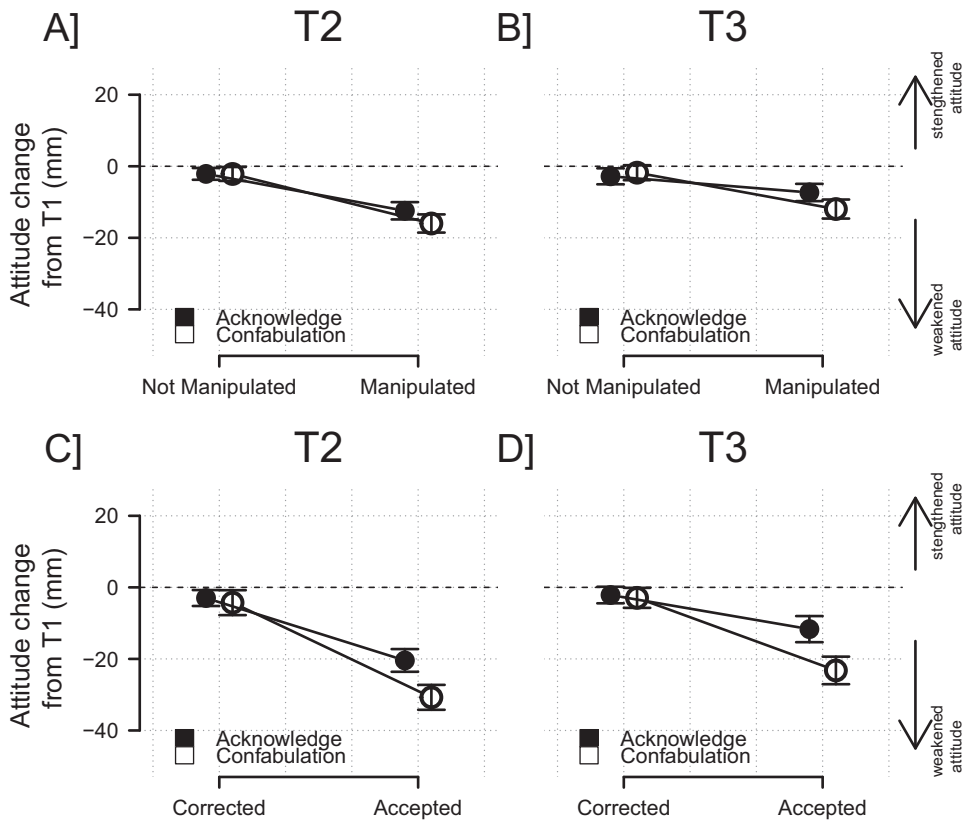

Figure 3. Attitude change. Average attitude change compared with original (T1) ratings in Experiment 2. A negative difference indicates a weakening of the original attitude. For manipulated trials this always means a change in direction toward the rating indicated by the false feedback. (A–B) Attitude change in T2 (A) and T3 (B) for Non-Manipulated and Manipulated trials split by Confabulation condition. (C–D) Attitude change for Manipulated trials only. Difference shown in T2 (C) and T3 (D) for Corrected and Accepted trials, split by Confabulation condition. Error bars are 95% CI.

slightly different meaning depending on whether the false feedback was accepted or not. For each subset we regressed Length, standardized, together with Time on Attitude Change.

For accepted trials, Length captures the amount of time participants spend giving confabulatory reasoning for their presented attitude. For these trials, although we found that the estimates were in the expected direction, that is, longer Length increases attitude change, the magnitude of the estimates was both small and nonsignificant ($b_{LENGTH} = -0.2$ mm, $SE = 2.9$; $\chi^2_{(1)} = 0.06$, $p = .81$; $b_{LENGTH*T3} = -1.2$ mm, $SE = 2.4$; $\chi^2_{(1)} = 0.23$, $p = .63$).

For corrected trials, however, Length captures both confabulatory reasoning as well as the time it takes for them to correct the presented attitude and enter a new one onto the tablet. Here we found a main effect of Length ($b_{LENGTH} = -4.4$ mm, $SE = 1.6$; $\chi^2_{(1)} = 9.04$, $p = .0026$), such that participants shifted their attitudes more in the directions of the manipulation the longer time they spent engaging with the false feedback, even if they ultimately corrected the presented attitude. There was no interaction effect of Length and Time ($b_{LENGTH*T3} = 0.1$ mm, $SE = 1.5$; $\chi^2_{(1)} = .006$, $p = .94$), nor any significant effect of time ($b_{T3} = 1.6$, $SE = 1.5$; $\chi^2_{(1)} = 1.22$, $p = .27$). The intercept, reflecting attitude change at T2 at average Length, was estimated as ($b_{intercept} = -5.1$ mm, $SE = 2.5$).

Possible moderators of attitude change. We examined three additional potential moderators of the attitude change observed: participants' CRT score, political involvement, and left–right attitude. All measures were entered into separate regressions together with Correction, Condition, and Time. No effects involving any of the candidate variables reached significance (all $ps > .066$). We report all coefficients and p values from all three models in Tables S2–S4.

Bayesian estimation of effects from both experiments. Finally, we combined the data from Experiment 1 and Experiment 2 and analyzed them using Bayesian multilevel regression estimating attitude change for Corrected and Accepted trials together with Time and Confabulation condition. This provides our best estimates of the effects of our main findings and of the posterior uncertainty surrounding our estimates. The model was fit using the full random effects structure grouped by both participant and question ID. Figure 4 shows the combined data from both experiments. Figure 5 shows the results from the Bayesian regression, with panel A displaying the regression coefficients mirroring the reporting from the separate analyses provided above. In panel B, posterior predictions of the average attitude changes are displayed for Corrected and Accepted trials.

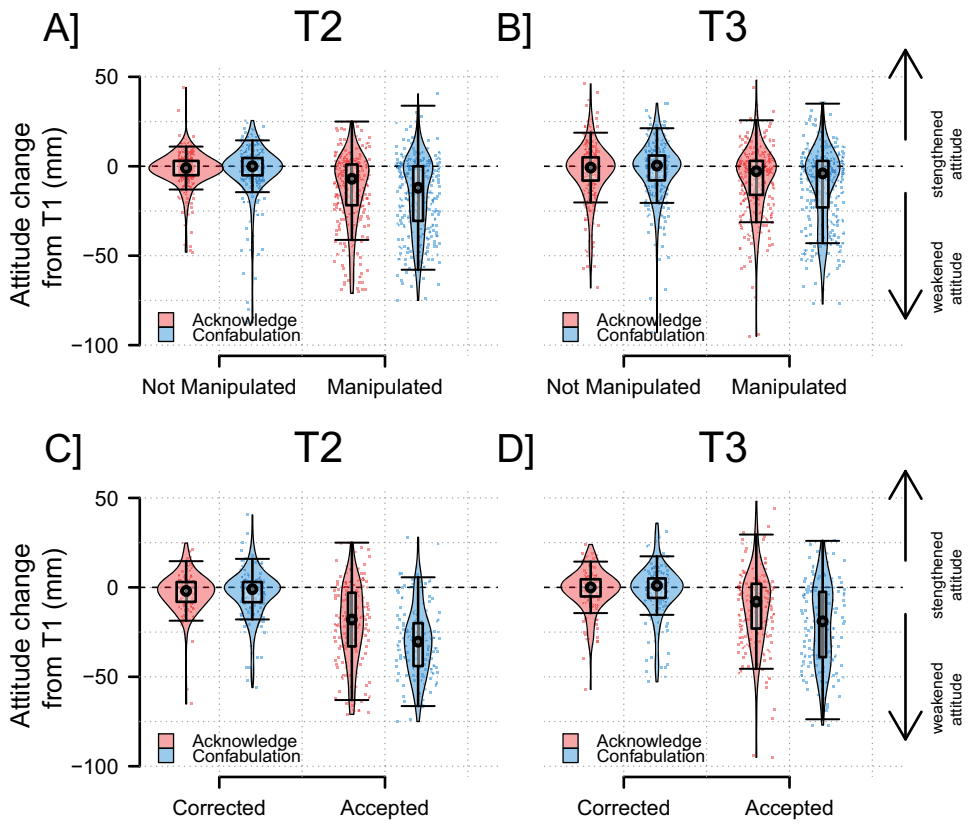

Figure 4. Data from both experiments. Attitude change compared with original (T1) ratings. A negative difference indicates a weakening of the original attitude. For manipulated trials this always means a change in direction toward the rating indicated by the false feedback. (A–B) Attitude change in T2 (A) and T3 (B) for Non-Manipulated and Manipulated trials split by Confabulation condition. (C–D) Attitude change for Manipulated trials only. Difference shown in T2 (C) and T3 (D) for Corrected and Accepted trials, split by Confabulation condition. Points represent individual trials. Boxplots depict median (large circle), 25th and 75th quantile (box edges) values, as well as 1.5*interquartile range (hinges). See the online article for the color version of this figure.

Discussion

In two experiments, we investigated whether false feedback concerning specific responses to political statements on a survey would influence later attitudes toward these issues. We found that half of the manipulations were accepted by the participants as being their own responses. Participants' responses were strongly affected by the false feedback, both in a session directly following the manipulation and one week later. In both experiments, we found that attitude change was much larger if participants were asked to reason about why they had stated the attitude falsely presented as their own compared with when only acknowledging its position.

Correction of the False Feedback

An important part of any experiment involving the CBP concerns the correction or acceptance of the false feedback. In this study we found that about half of manipulated responses were corrected by the participants, which is in line with our previous results in the moral and political domains (Hall et al., 2012, 2013). Naturally, participants were more likely to correct a manipulated rating if their original response was extreme, and if the confidence rating regarding the attitude was high, however this was not predictive of the size of the ensuing attitude change. To get a better understanding of what increases the likelihood of a manipulation to be accepted or corrected, we added several related individual difference measures. In Experiment 2, participants reported their degree of political involvement, and where they would place themselves on the left–right spectrum. They also completed the CRT (Frederick, 2005), which is a short measure of reflexivity and critical reasoning.

We found no correlation between level of correction and self-rated political involvement. This is noteworthy, given the common assumption that increased political involvement also entails increased political awareness and more stable attitudes (Converse, 1964; Zaller, 1992), and how the result contrasts with previous findings from our own lab (Hall et al., 2013; Strandberg, Björklund, Pärnamets, Hall, & Johansson, 2018). However, political orientation on a left-right political ideology scale predicted correction, such that more left-leaning participants had higher rate of correction. However, this effect is probably best explained by the fact that more participants rated

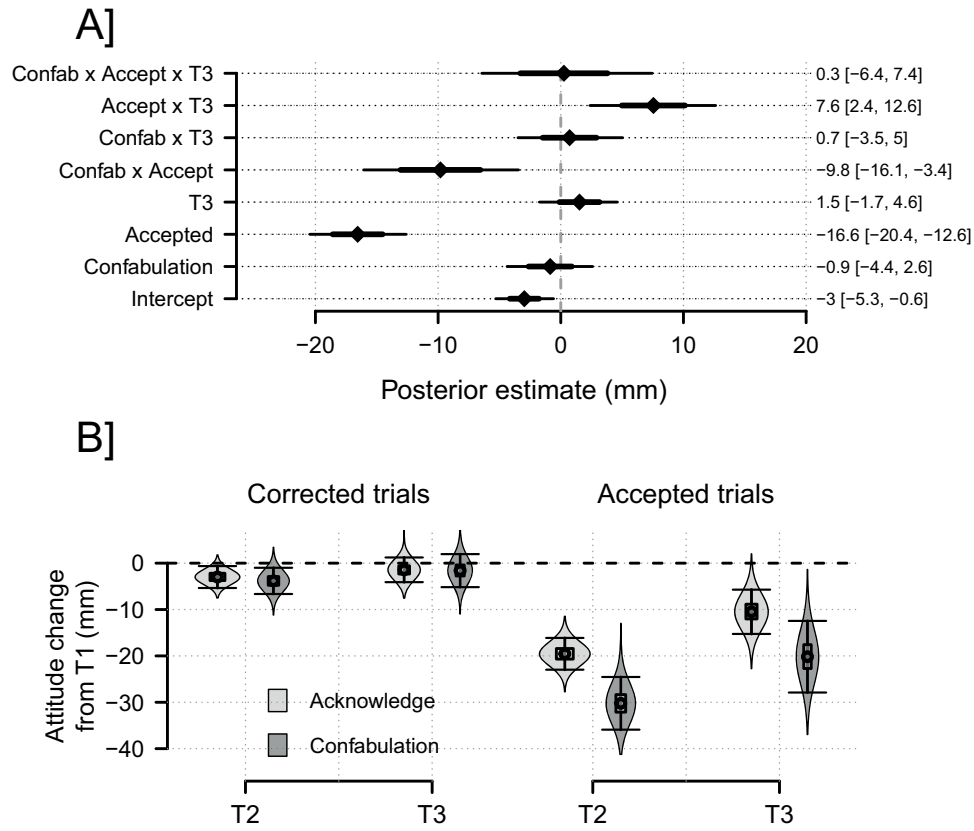

Figure 5. Results from Bayesian regression. (A) Posterior estimates from Bayesian regression combining data from Experiment 1 and Experiment 2 from manipulated trials only. Estimates reflect the coefficients contribution to attitude change measured as a difference from the original (T1) ratings. A negative difference indicates a weakening of the original attitude (in the direction of the false feedback). The reference level captured by the intercept reflects attitude change for Corrected trials in the Acknowledge condition at T2. All regressors were dummy coded. Points represent the mean posterior estimate; thick bars represent the standard deviation of the posterior and thin bars the 95% credible interval. The numerical column displays the mean of the posterior and 95% credible intervals. (B) Violin plots depicting distribution of posterior predictions from a Bayesian regression model combining data from Experiment 1 and Experiment 2. Estimates reflect predicted attitude change compared with original (T1). A negative difference indicates a weakening of the original attitude (in the direction of the false feedback). Left panel depicts Corrected trials and right panel depicts Accepted trials. Points represent the mean posterior prediction. Boxes show the interquartile range (IQR) and hinges 1.5*IQR.

themselves to be strongly left compared than participants being strongly right (see distribution in Figure S1).

It has recently been found that there is a positive correlation between CRT score and ability to differentiate between real and fake news (Pennycook & Rand, 2017), as well as between CRT and measures of politically motivated cognition (Kahan, 2013). Considering this research, and the basic assumption that CRT captures analytic skill, we hypothesized that it would correlate with level of correction. This is also what we found, with participants scoring higher on CRT also having a higher likelihood of correcting the false feedback. Few individual difference predictors of correction have been found in previous research using the CBP (McLaughlin & Somerville, 2013; Sagana et al., 2016; Strandberg et al., 2018, but see Aardema et al., 2014), making this result of general interest. More research is needed to establish which mechanism is captured by CRT in this context—whether it is memory of prior answers, or more elaborate belief structures, or some other factor.

Influence of False Feedback on Future Attitudes

As a backdrop to the false feedback manipulations in our study, and given the debate we outlined in the Introduction between stable and flexible attitudes (e.g., Alwin, 1994; Bishop, 2005; Converse, 1975; Gerber et al., 2011; Haidt, 2001; Hall et al., 2013; Hatemi et al., 2009; Hooghe & Wilkenfeld, 2008; Sears & Funk, 1990; Zaller, 1992), it is important to note that our participants generally displayed stability in their attitudes. For the nonmanipulated trials there were no attitude shifts during the first follow-up, and one week later, during the second follow-up, these responses remained at their original positions. Generally, this was the case also for the trials where the participants corrected the false feedback.

In contrast, for the manipulated trials in both experiments, we found that participants' attitudes following the first session, as well as one week later, were shifted in the direction of the false feedback. The observed changes are consistent with previous work demonstrating

preference change through choice using various false feedback procedures (Izuma et al., 2015; Janis & King, 1954; Johansson et al., 2014; Luo & Yu, 2016; Sharot, Fleming, Yu, Koster, & Dolan, 2012). However, our findings are noteworthy given the prior mixed evidence for more enduring changes in these paradigms (Sharot et al., 2012; Taya et al., 2014). In addition, prior studies have concerned preferential binary choices between pairs of faces and abstract images, or ratings of near equally preferred holiday destination, or hypothetical estimations of future events. To avoid these problems, we employed a more ecological procedure in the form of a political attitude survey focusing on specific, current political issues. This is not only a domain of great general importance, but one where preferences are supposed to be more resilient to change (Bartels, 2002; Gerber et al., 2011; Hatemi et al., 2009; Hooghe & Wilkenfeld, 2008; Sears & Funk, 1990), as we also saw with the nonmanipulated trials in our experiments. The specificity of the political questions, together with our confrontation procedure which required participants to both read the statement and the presented rating, suggests that the changes observed cannot be explained as being due to any vagueness in the targeted preference statements or a change in abstract values rather than specific attitudes as in some of the past research (e.g., Rokeach, 1971).

In both of our experiments, the average observed changes were large. The differences in ratings between Session 1 compared with Session 2 reached almost a full quarter of the length of the rating scale, and in most of Accepted trials these shifts crossed the midline (i.e., clearly defining the position as different from the original attitude). A week after the manipulation, the combined estimates from both experiments indicate that the attitude changes linger between about 10 mm and 20 mm for the accepted trials (Acknowledge and Confabulation conditions, respectively; see Figure 5). These effect sizes are notable when, for example, compared with those of around 10 points (of 100) found by Broockman and Kalla (2016) using a considerably longer and more involved intervention. The attitude changes were obtained absent of any reinforcement following the false feedback manipulation; the participants only viewed the manipulation once, and then immersed themselves in their ordinary life for a full week, with their usual sources of information and personal political biases. Even in the confabulation condition, the experimenter only asked the participants to explain the reasons behind their (manipulated) attitudes, and avoided further engagement in the argumentation. Considering this, the findings here present a strong demonstration of the power of even brief false feedback to engender attitude changes.

Confabulating About False Feedback Influences Future Responses

To investigate confabulation as a possible vehicle of attitude change, we varied the amount of confabulation participants gave in response to the manipulated ratings. In both experiments, we found that participants who had been asked to explain their responses, compared with those who merely acknowledged their (manipulated) attitude, showed larger attitude changes, both shortly after the manipulation and one week later. The average increase in rating difference was around 50% in the confabulation condition compared with the acknowledge condition at T2 and almost twice as large at T3, representing a considerable increase in relative effect size. This shows how the perception and verbalization of one's own reasoning can influence one's attitudes (cf. Barden & Tormala, 2014; Tormala & Petty, 2002), but as far as we know, the effect of confabulatory reasoning in facilitating attitude change is previously unstudied.

In the analysis of the confabulation condition in Experiment 2, we also looked at trial-based speaking time as an estimate of confabulation length. Using this more fine-grained measure, we found no correlation between confabulation length and the magnitude of attitude change in the accepted manipulated trials. This indicates that the exploratory measure of time taken during confabulation is not sufficient to capture what it is about confabulation that engenders attitude change. This is notable given previous research showing that differences in time spent merely thinking about an object can have varying influence on the attitudes toward that object (Clarkson et al., 2011). Testing a greater span of measures, including various forms of content and semantic analysis, will be necessary to fully explain the details of the effect confabulation have on attitude change. In the corrected trials, however, we found a correlation between confabulation length and attitude change, such that the longer time the participants spent engaging with the false feedback the more they shifted in the manipulated direction. Our interpretation, based on informal observations, is that these participants often start constructing arguments for the manipulated position before instead backtracking to correct the presented attitude. This indicates that, under some circumstances, even small amounts of confabulation can influence a person's beliefs.

Although it is important to acknowledge that similar findings have been reported in the literature on self-persuasion using other methods, such as perspective taking (Broockman & Kalla, 2016), imagination (Carroll, 1978; Gregory, Cialdini, & Carpenter, 1982; Watts, 1967), or counterattitudinal argumentation (Lord, Lepper, & Preston, 1984; Mussweiler, Strack, & Pfeiffer, 2000; Watts, 1967), these approaches all suffer from different limitations. In the traditional self-persuasion experiments, participants' attitudes are often compared with control groups (Watts, 1967), the original attitude is established several months prior to the experiment (King & Janis, 1956), or they are asked to assess their own attitudinal change (Lord et al., 1984), resulting in uncertainty about what the participants' original attitudes were and whether any change has taken place. Crucially, in those experiments, participants are also fully aware that the attitude they are asked to express is not their own, and that the arguments they produce are hypothetical (e.g., Janis & King, 1954; Lord et al., 1984), whereas in a CBP experiment participants believe the manipulated response to reflect their own true attitude. In the Confabulation condition, the participants produce arguments in favor of that attitude, just like they would have in an everyday interaction. This means that the present study removes the pressing problems of demand effects as an explanation for the observed attitude change, a concern present in most prior studies. Thus, a key contribution of the present study is that it provides clearer and firmer support for the hypothesis that processes of self-perception can be involved in attitude change.

Implications for Attitudes and Preferences

How do these findings relate to theories of attitudes and preferences more broadly? One lesson to learn from this study, in relation to the overarching tension between views of political attitudes as stable or flexible, is that both perspectives may capture

important aspects of how such attitudes function. On the one hand, absent any manipulation, participants gave the same responses throughout the experiment, clearly indicating they had a stable set of political attitudes. On the other hand, the same participants exhibited large lasting attitude shifts after having accepted the false feedback.

We have previously shown that participants often accept false feedback about their political attitudes, thus revealing a previously undiscovered flexibility to reason beyond ideological labels (Hall et al., 2013). However, these attitude shifts were only measured at the moment of the feedback in terms of accepting the manipulation, but no subsequent follow-up attitude measurements were performed. Here we have extended that work, by showing lasting attitude changes measured during two follow-up elicitations, demonstrating that participants' initial attitudinal flexibility extends far beyond that of the immediate confrontation with the false feedback. The attitude shifts at the latter stages of the study were not as large as those implied by the false feedback and accepted by the participants. This might signal an upper bound on attitude flexibility when translated into future behavior but might also be due to some form of gravitational pull from interlocking opposing attitudes, or counter pushing from everyday influences in the life of the participants (family and friends, selective news circles, etc.), or just simply noise induced by memory decay. If so, reinforcing the shifted attitudes, by for example exposing participants to extra arguments supporting their new position, would likely lead participants to coalesce their position closer to the one implied by the false feedback.

Another way of approaching the stable/flexible dichotomy is through the lens of inferential and constructivist accounts of preference and attitude formation (Ariely & Norton, 2008; Slovic, 1995; Warren, McGraw, & Van Boven, 2011). On strong versions of such accounts, the act of choosing has a constitutive role in the genesis of a persons' preference set (Ariely & Norton, 2008; Slovic, 1995), to the point that some choices might reflect purely arbitrary influences on the preference (Ariely, Loewenstein, & Prelec, 2003; Chater, Johansson, & Hall, 2011). A more balanced view instead holds that preferences and attitudes are calculated to some degree at the time of choice (Warren et al., 2011), recasting the question of stable versus flexible attitudes from a categorical one into a continuum. Instead it becomes key to discover what factors influence the degree of calculation and how that process is supported. In this vein we have previously argued, based on preference changes for faces induced using the CBP (Johansson et al., 2014), that preference or attitude change in the CBP taps into a specific aspect of preference calculation, namely that preference calculation is supported by a process of self-perception. Inferences about one's own attitudes or preferences go via observations of the outcomes of past behavior. In other words, we often infer our own preferences much like we infer other peoples' preferences, by observing and interpreting our own overt behavior (Bem, 1967; Johansson et al., 2014). Once we believe we have stated some attitude, it follows that we should infer that we also hold that attitude. For example, recent work has demonstrated that once beliefs change, recollections of past beliefs become biased to match the current belief (Wolfe & Williams, 2017).

The proposition that participants rely on their beliefs about their past attitude ratings to inform their new ratings bears structural similarities with "options-as-information" theory, developed to account for some challenges to classical decision theory arising from observed preference reversals in multiattribute choice (Müller-Trede, Sher, & McKenzie, 2015; Sher & McKenzie, 2014). The theory takes the form of a rational analysis (Oaksford & Chater, 1994), positing that by accounting for participants' prior beliefs going into a decision task, seemingly inconsistent patterns of preferences can be accommodated using a normative framework based on Bayesian updating. The decisions analyzed differ from the conditions of the present study, but nevertheless the question arises to what extent a framework such as "options-as-information," or broadly, a conception of decision makers as performing updating of their attitudes according to Bayesian normative theory, can be useful in explaining the observed attitude changes reported here.

One way of understanding participants' behavior at T2, in the accepted manipulated trials, is that they must reconcile two conflicting representations of their past attitudes. One being the trace of their original attitude, the second being the one presented during the false feedback confrontation. Depending on the weighting between these representations the participants' new attitudes should fall within that interval. If the weighting is equal the average attitude change should be half the average manipulation length, which is consistent with the data presented here, at least for T2. This suggests at least a tentative compatibility of the predictions of a theory like "options-as-information" and our findings, though more formal analysis and experiments specifically designed to test this would be required. Regardless, some rationalization of participants' behavior should be forthcoming. It is important for us to stress that although findings of choice blindness are counterintuitive by folk psychological reasoning, and perhaps the ensuing attitude changes reported here even more so, we do not take the findings presented here to demonstrate some fundamental irrationality on part of the participants. Rather, our findings highlight the continuous and dynamic evolution of attitudes with respect to new information about oneself and one's beliefs.

That beliefs play a role aligns with a growing consensus across the decision sciences regarding the importance of memory processes for understanding value-based choice, where much recent work has focused on the influence of past episodes for the calculation of preferences (Bornstein, Khaw, Shohamy, & Daw, 2017; Murty, FeldmanHall, Hunter, Phelps, & Davachi, 2016; Shadlen & Shohamy, 2016). Using the CBP, we have previously shown that false feedback about choices leads to systematic distortions of participants' source memory, thus demonstrating that beliefs are formed resulting from acceptance of the false feedback (Pärnamets et al., 2015). This is consistent with other work showing source memory distortions when reasoning about past choices (Mather, Shafir, & Johnson, 2000). Understood in the light of the present study, observations of our own past political survey responses lead to the inference that we hold those attitudes, this belief then influences later attitude construction when queried in the future.

Strengths, Limitations, and Future Studies

Future work should address questions arising both from the findings reported here and from limitations in the study design. We have demonstrated lasting attitude change following a simple false feedback manipulation. One route toward deepening our understanding of this finding is to investigate how far attitudes can be

shifted. This would include follow-up sessions over longer periods of time as well as adopting a procedure where participants' false beliefs about their past attitudes were reinforced, perhaps by supplanting participants with additional arguments to buttress their new-found positions. Together this would allow us to better understand the interplay between original and implanted attitudes, and perhaps better model attitude shifts arising from malicious information sources in the world outside the lab. We have also argued that our attitude shifts are dependent on participants gaining false beliefs about their past attitudes. Hence, a key area to look at in future studies would be how false beliefs about past attitudes are integrated into participants' broader belief structure and how resulting changes in participants' memories about their own attitudes are maintained.

There is also the possibility to use CBP to explore other domains than politics, such as personal values, personality traits, or character attributes. The case of values is particularly relevant to the present study as values are thought to underpin many political attitudes (Schwartz et al., 2012). Although previous work applying CBP to moral questions (Hall et al., 2012), including moral principles, indicates that also values should be susceptible to false feedback manipulations, little is known how these effects translate back into attitudes or behavior. Studies have shown that values and value-relevant behavior can be susceptible to influence—for example by priming reasons or making the reasons more salient (Maio, Hahn, Frost, & Cheung, 2009), and it is possible that accepting false feedback about values might recruit similar processes on downstream behavior. Nevertheless, other value changes appear to occur on longer time-scales in relation to significant life events (Bardi, Buchanan, Goodwin, Slabu, & Robinson, 2014) or not at all (Manfredo et al., 2017). This leaves an important avenue for exploring whether people can become, for example, more altruistic, fair, or patriotic, by making them adopt and argue for false beliefs about their values.

To increase the generalizability of our study, replicating it on a sample representative of the general population would be desirable. In a similar vein, assessing whether the findings are limited to a Western, educated, industrialized, rich, and democratic (WEIRD) population is of importance (Henrich, Heine, & Norenzayan, 2010). In this study we targeted political attitudes from two salient domains, education and environmental issues. Of course, this does not exhaust the spectrum of political topics, and it is important to assess whether political attitudes behave the same across varying topics and questions, with various levels of polarization and acrimony. Nevertheless, unpublished data from studies conducted during the 2016 U.S. election indicate that at least some of these effects are transferable to domains involving political leaders and generalize to a broader U.S. population (Strandberg, Olson, Hall, Raz, & Johansson, 2018).

As we see it, one of the clearest theoretical contributions of the current study is that we create a self-perception situation where the participants truly believe the manipulated attitudes to be their own, thus creating much stronger grounds for consequential self-inferences. As we detail below, this ought not be interpreted as an irrational, or worse, even pathological, process, but instead as a reasonable inferential response to a peculiar array of evidence. However, more speculatively, some self-perception theories have suggested that there might be a special relationship between attitudes and first-person authority, such that attitudes we endorse (either by acknowledgment or confabulation in the current study), also creates a special sense of agency or ownership of that attitude (see Carruthers, 2011; Martin & Pacherie, 2013; Moran, 2001). This phenomenological emotional component might then feed into or enhance the self-inferences seen in the CBP compared with previous paradigms. Unfortunately, there is nothing in the current design that allow us to disentangle these possibilities, so this remains as an exciting avenue for future research.

As detailed above, our preferable way of framing the self-inferential process would be in terms of Bayesian updating of beliefs. From this standpoint, the difference between the Acknowledge and the Confabulation condition is one of degree, where confabulation simply adds another layer of evidence to the self-inferences. Similarly, other theoretical frameworks of attitude change, such as the Elaboration Likelihood Model (ELM; Petty & Cacioppo, 1986; Petty, Haugtvedt, & Smith, 1995), could potentially help to explain the differences in change found between the Acknowledge and the Confabulation conditions. According to ELM, in the Confabulation condition, participants can be expected to make thoughtful and deliberate considerations of the arguments they generate. This would allow them to engage in deeper information processing compared with participants that simply acknowledge the stated attitude as their own, and this difference in information processing could be used to explain the different T2 and T3 effects between conditions.

Potentially, the matrix of evidence in CPB might also include our beliefs and expectations about *other* people, and their reactions to our opinions—that is, part of the difference between the two conditions might reside in the confabulations functioning as a public *commitment* (as has been explored in the literature on conversational implicature (Brandom, 1994; Grice, 1975). In future studies, this would be an interesting dimension to explore, by creating contexts with potentially more or less social commitment, e.g., by comparing the role of a politician to an entertainer, or a teacher to a student.

Conclusions

In summary, the results presented here demonstrate attitude flexibility in the face of accepted false feedback about previously held positions and how confabulatory reasoning facilitates shifts away from the original position. These results were obtained studying political attitudes; a domain of central importance to public life. On the face of it, this might seem like a troubling result, showcasing the shallowness of our political attitudes (Converse, 1975, 1964; Zaller, 1992), and potentially exposing us to manipulation by malicious opponents. Even though our study was not an attempt at a practical canvassing effort, like Broockman and Kalla (2016), this possibility should not be downplayed. Although scientific methods can sometimes be misused by unscrupulous individuals, we take issue with the interpretation that the current findings reveal inherent flaws in our attitudes. Indeed, why should it be considered an ideal to have attitudes so firmly chiseled and bounded that one would consistently notice all CB manipulations? This position is only intelligible against a backdrop of a society where particularly firm opinions are held in reverie, and where undecideds and moderates are derided as "wishy-washers" and "flip-floppers." But this might be a harmful standard (cf. Hall et al., 2012, 2013). As we see it, the current run of hyper-polarization

in politics is not only simple aggregation of individual attitudes but also a result of our larger views of what it is to hold an attitude. In times of information bubbles, fake news, political acrimony, and gridlock, we find it encouraging that a brief CBP intervention can nudge people to find support for positions other than those originally held. This opens up new perspectives for understanding across the political divide and serves as a reminder that people can demonstrate flexibility when they are induced to reason about complex political issues.

Context of Research

The research reported in this article originated in our earlier work observing choice blindness for political attitudes as well as effects of choice blindness on later choices and memories for simpler preferential decisions. We were interested in testing whether political attitudes could be changed by giving false feedback to participants about their own prior responses. Additionally, this allowed us to visit an underexplored aspect of the choice blindness paradigm: the role of the confabulatory statements participants make in support of the false feedback response. We hypothesized that if participants have formed a false belief about their past attitude, then confabulating reasons for that attitude should increase the change observed in their later responses. Key ideas for future work will be to compare similarities and differences in argument content and paralinguisic markers when defending manipulated versus nonmanipulated responses. We will also investigate how the memory of past attitudes is influenced when false beliefs about one's attitudes are adopted. By implementing a self-inferential, constructivist approach to the study of political attitudes, we believe that this research can contribute to the understanding of mass opinion.

References

Aardema, F., Johansson, P., Hall, L., Paradisis, S.-M., Zidani, M., & Roberts, S. (2014). Choice blindness, confabulatory introspection, and obsessive-compulsive symptoms: A new area of investigation. *International Journal of Cognitive Therapy, 7,* 83–102. http://dx.doi.org/10.1521/ijct.2014.7.1.83

Alwin, D. (1994). Aging, personality, and social change: The stability of individual differences over the adult life span. In D. L. Featherman, R. M. Lerner, & M. Perlmutter (Eds.), *Life-span development and behavior* (pp. 135–185). Hillsdale, NJ: Erlbaum.

Anand, S., & Krosnick, J. A. (2003). The impact of attitudes towards foreign policy goals on public preferences among presidential candidates: A study of issue publics and the attentive public in the 2000 U. S presidential election. *Presidential Studies Quarterly, 33,* 31–71. http://dx.doi.org/10.1177/0360491802250541

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics, 118,* 73–106. http://dx.doi.org/10.1162/00335530360535153

Ariely, D., & Norton, M. I. (2008). How actions create—not just reveal—preferences. *Trends in Cognitive Sciences, 12,* 13–16. http://dx.doi.org/10.1016/j.tics.2007.10.008

Barden, J., & Tormala, Z. L. (2014). Elaboration and attitude strength: The new meta-cognitive perspective. *Social and Personality Psychology Compass, 8,* 17–29. http://dx.doi.org/10.1111/spc3.12078

Bardi, A., Buchanan, K. E., Goodwin, R., Slabu, L., & Robinson, M. (2014). Value stability and change during self-chosen life transitions: Self-selection versus socialization effects. *Journal of Personality and Social Psychology, 106,* 131–147. http://dx.doi.org/10.1037/a0034818

Bartels, L. M. (2002). Beyond the running tally: Partisan bias in political perceptions. *Political Behavior, 24,* 117–150. http://dx.doi.org/10.1023/A:1021226224601

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review, 74,* 183–200. http://dx.doi.org/10.1037/h0024835

Bernstein, D. M., Laney, C., Morris, E. K., & Loftus, E. (2005). False memories about food can lead to food avoidance. *Social Cognition, 23,* 11–34. http://dx.doi.org/10.1521/soco.23.1.11.59195

Bialek, M., & Pennycook, G. (2017). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods*. Advance online publication. http://dx.doi.org/10.3758/s13428-017-0963-x

Bishop, G. F. (2005). *The illusion of public opinion: Fact and artifact in American public opinion polls*. Lanham, MD: Rowman and Littlefield.

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications, 8,* 15958. http://dx.doi.org/10.1038/ncomms15958

Brandom, R. (1994). *Making it explicit*. Cambridge, MA: Harvard university press.

Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science, 352,* 220–224. http://dx.doi.org/10.1126/science.aad9713

Buerkner, P. C. (2016). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80,* 1–28.

Bullock, J. G. (2011). Elite influence on public opinion in an informed electorate. *The American Political Science Review, 105,* 496–515. http://dx.doi.org/10.1017/S0003055411000165

Burke, E. (1774). *On American taxation*. Indianapolis, IN: Liberty Fund.

Carroll, J. S. (1978). The effect of imagining an event on expectations for the event: An interpretation in terms of the availability heuristic. *Journal of Experimental Social Psychology, 14,* 88–96. http://dx.doi.org/10.1016/0022-1031(78)90062-8

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199596195.001.0001

Chater, N., Johansson, P., & Hall, L. (2011). The non-existence of risk attitude. *Frontiers in Psychology, 2,* 303. http://dx.doi.org/10.3389/fpsyg.2011.00303

Clarkson, J. J., Tormala, Z. L., & Leone, C. (2011). A self-validation perspective on the mere thought effect. *Journal of Experimental Social Psychology, 47,* 449–454. http://dx.doi.org/10.1016/j.jesp.2010.12.003

Converse, P. (1964). The nature of belief systems in mass publics. In D. E. Apter (Ed.), *Ideology and discontent* (pp. 206–261). New York, NY: The Free Press.

Converse, P. (1975). Public opinion and voting behavior. In F. Greenstein & N. Polsby (Eds.), *Handbook of political science 4* (pp. 75–171). Reading, UK: Addison Wesley.

Dennett, D. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 63,* 568–584. http://dx.doi.org/10.1037/0022-3514.63.4.568

Druckman, J. N. (2004). Political preference formation: Competition, deliberation, and the (ir)relevance of framing effects. *The American Political Science Review, 98,* 671–686. http://dx.doi.org/10.1017/S0003055404041413

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about pol-

itics. *Political Psychology, 38,* 127–150. http://dx.doi.org/10.1111/pops.12394

Fotopoulou, A., Conway, M. A., & Solms, M. (2007). Confabulation: Motivated reality monitoring. *Neuropsychologia, 45,* 2180–2190. http://dx.doi.org/10.1016/j.neuropsychologia.2007.03.003

Fox, J., & Weisberg, S. (2011). *An {R} companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives, 19,* 25–42. http://dx.doi.org/10.1257/089533005775196732

French, L., Garry, M., & Loftus, E. (2009). False memories: A kind of confabulation in non-clinical subjects. In W. Hirstein (Ed.), *Confabulation: Views from neuroscience, psychiatry, psychology, and philosophy* (pp. 33–66). Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199208913.003.02

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2011). The big five personality traits in the political arena. *Annual Review of Political Science, 14,* 265–287. http://dx.doi.org/10.1146/annurev-polisci-051010-111659

Gregory, W. L., Cialdini, R. B., & Carpenter, K. M. (1982). Self-relevant scenarios as mediators of likelihood estimates and compliance: Does imagining make it so? *Journal of Personality and Social Psychology, 43,* 89–99. http://dx.doi.org/10.1037/0022-3514.43.1.89

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Studies in syntax and semantics III: Speech acts* (pp. 41–58) New York, NY: Academic Press.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108,* 814–834. http://dx.doi.org/10.1037/0033-295X.108.4.814

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PLoS ONE, 7,* e45457. http://dx.doi.org/10.1371/journal.pone.0045457

Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010). Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition, 117,* 54–61. http://dx.doi.org/10.1016/j.cognition.2010.06.010

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PLoS ONE, 8,* e60554. http://dx.doi.org/10.1371/journal.pone.0060554

Hatemi, P. K., Funk, C. L., Medland, S. E., Maes, H. M., Silberg, J. L., Martin, N. G., & Eaves, L. J. (2009). Genetic and environmental transmission of political attitudes over a life time. *The Journal of Politics, 71,* 1141–1156. http://dx.doi.org/10.1017/S0022381609090938

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33,* 61–83. http://dx.doi.org/10.1017/S0140525X0999152X

Hirstein, W. (2009). *Confabulation: Views from neuroscience, psychiatry, psychology, and philosophy*. Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199208913.001.0001

Hooghe, M., & Wilkenfeld, B. (2008). The stability of political attitudes and behaviors across adolescence and early adulthood: A comparison of survey data on adolescents and young adults in eight countries. *Journal of Youth and Adolescence, 37,* 155–167. http://dx.doi.org/10.1007/s10964-007-9199-x

Isenberg, I. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology, 50,* 1141–1151. http://dx.doi.org/10.1037/0022-3514.50.6.1141

Izuma, K., Akula, S., Murayama, K., Wu, D. A., Iacoboni, M., & Adolphs, R. (2015). A causal role for posterior medial frontal cortex in choice-induced preference change. *The Journal of Neuroscience, 35,* 3598–3606. http://dx.doi.org/10.1523/JNEUROSCI.4591-14.2015

Janis, I. L., & King, B. T. (1954). The influence of role playing on opinion change. *Journal of Abnormal Psychology, 49,* 211–218. http://dx.doi.org/10.1037/h0056957

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science, 310,* 116–119. http://dx.doi.org/10.1126/science.1111709

Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition, 15,* 673–692. http://dx.doi.org/10.1016/j.concog.2006.09.004

Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2014). Choice blindness and preference change. *Journal of Behavioral Decision Making, 27,* 281–289. http://dx.doi.org/10.1002/bdm.1807

Jost, J. T., & Amodio, D. M. (2012). Political ideology as motivated social cognition: Behavioral and neuroscientific evidence. *Motivation and Emotion, 36,* 55–64. http://dx.doi.org/10.1007/s11031-011-9260-7

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making, 8,* 407–424.

King, B. T., & Janis, I. L. (1956). Comparison of the effectiveness of improvised versus non-improvised role-playing in producing opinion changes. *Human Relations, 9,* 177–186. http://dx.doi.org/10.1177/001872675600900202

Knowles, E. S., & Linn, J. A. (Eds.). (2004). The importance of resistance to persuasion. *Resistance and persuasion* (pp. 3–9). Mahwah, NJ: Erlbaum.

Kogan, N., & Wallach, M. A. (1967). Group risk taking as a function of members' anxiety and defensiveness levels. *Journal of Personality, 35,* 50–63. http://dx.doi.org/10.1111/j.1467-6494.1967.tb01415.x

Kunda, Z. (1987). Motivation and inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology, 53,* 636–647. http://dx.doi.org/10.1037/0022-3514.53.4.636

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480–498. http://dx.doi.org/10.1037/0033-2909.108.3.480

Ledford, H. (2016, April 7). Door-to-door canvassing reduces transphobia. *NATNews*. Advance online publication. http://dx.doi.org/10.1038/nature.2016.19713

Lefcheck, J. S. (2016). piecewiseSEM: Piecewise structural equation modeling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution, 7,* 573–579. http://dx.doi.org/10.1111/2041-210X.12512

Lewis, G. J. (2018). Early-childhood conduct problems predict Economic and political discontent in adulthood: Evidence from two large, longitudinal U. K. Cohorts. *Psychological Science, 29,* 711–722. http://dx.doi.org/10.1177/0956797617742159

Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014). Speakers' acceptance of real-time speech exchange indicates that we use auditory feedback to specify the meaning of what we say. *Psychological Science, 25,* 1198–1205. http://dx.doi.org/10.1177/0956797614529797

Loftus, E., & Zanni, G. (1975). Eyewitness testimony: The influence of the wording of a questions. *Bulletin of the Psychonomic Society, 5,* 86–88. http://dx.doi.org/10.3758/BF03336715

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47,* 1231–1243. http://dx.doi.org/10.1037/0022-3514.47.6.1231

Luo, J., & Yu, R. (2016). The Spreading of alternatives: Is it the perceived choice or actual choice that changes our preference? *Journal of Behavioral Decision Making*. Advance online publication. http://dx.doi.org/10.1002/bdm.1967

Maio, G. R., Hahn, U., Frost, J. M., & Cheung, W. Y. (2009). Applying the value of equality unequally: Effects of value instantiations that vary in typicality. *Journal of Personality and Social Psychology, 97,* 598–614. http://dx.doi.org/10.1037/a0016683

Manfredo, M. J., Bruskotter, J. T., Teel, T. L., Fulton, D., Schwartz, S. H., Arlinghaus, R., . . . Sullivan, L. (2017). Why social values cannot be changed for the sake of conservation. *Conservation Biology, 31,* 772–780. http://dx.doi.org/10.1111/cobi.12855

Martin, J. R., & Pacherie, E. (2013). Out of nowhere: Thought insertion, ownership and context-integration. *Consciousness and Cognition, 22,* 111–122. http://dx.doi.org/10.1016/j.concog.2012.11.012

Mather, M., Shafir, E., & Johnson, M. K. (2000). Misrememberance of options past: Source monitoring and choice. *Psychological Science, 11,* 132–138.

McLaughlin, O., & Somerville, J. (2013). Choice blindness in financial decision making. *Judgment and Decision Making, 8,* 577.

McNair, B. (2017). *Fake news: Falsehood, fabrication and fantasy in journalism*. London, UK: Routledge.

Moran, R. (2001). *Authority and estrangement: An essay on self-knowledge*. Princeton, NJ: Princeton University Press.

Müller-Trede, J., Sher, S., & McKenzie, C. R. (2015). Transitivity in context: A rational analysis of intransitive choice and context-sensitive preference. *Decision, 2,* 280–305. http://dx.doi.org/10.1037/dec0000037

Murty, V. P., FeldmanHall, O., Hunter, L. E., Phelps, E. A., & Davachi, L. (2016). Episodic memories predict adaptive value-based decision-making. *Journal of Experimental Psychology: General, 145,* 548–558. http://dx.doi.org/10.1037/xge0000158

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin, 26,* 1142–1150. http://dx.doi.org/10.1177/01461672002611010

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101,* 608–631. http://dx.doi.org/10.1037/0033-295X.101.4.608

Pärnamets, P., Hall, L., & Johansson, P. (2015). Memory distortions resulting from a choice blindness task. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1823–1828). Austin, TX: Cognitive Science Society.

Pennycook, G., & Rand, D. G. (2017). *Who falls for fake news? The roles of analytic thinking, motivated reasoning, political ideology, and bullshit receptivity* (September 12, 2017). Retrieved from https://ssrn.com/abstract=3023545

Pennycook, G., & Ross, R. M. (2016). Commentary: Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology, 7,* 9. http://dx.doi.org/10.3389/fpsyg.2016.00009

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology, 19,* 123–205. http://dx.doi.org/10.1016/S0065-2601(08)60214-2

Petty, R. E., Haugtvedt, C. P., & Smith, S. M. (1995). Elaboration as a determinant of attitude strength. *Attitude strength, 4,* 93–130.

Rokeach, M. (1971). Long-range experimental modification of values, attitudes, and behavior. *American Psychologist, 26,* 453–459. http://dx.doi.org/10.1037/h0031450

Sagana, A., Sauerland, M., & Merckelbach, H. (2016). The effect of choice reversals on blindness for identification decisions. *Psychology, Crime & Law, 22,* 303–314. http://dx.doi.org/10.1080/1068316X.2015.1085984

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., . . . Konty, M. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology, 103,* 663–688. http://dx.doi.org/10.1037/a0029393

Sears, D. O. (1983). The persistence of early political predispositions: The roles of attitude object and life stage. *Review of Personality and Social Psychology, 4,* 79–116.

Sears, D. O., & Funk, D. L. (1990). Evidence of long-term persistence of adults' political predispositions. *The Journal of Politics, 61,* 1–28. http://dx.doi.org/10.2307/2647773

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron, 90,* 927–939. http://dx.doi.org/10.1016/j.neuron.2016.04.036

Sharot, T., Fleming, S. M., Yu, X., Koster, R., & Dolan, R. J. (2012). Is choice-induced preference change long lasting? *Psychological Science, 23,* 1123–1129. http://dx.doi.org/10.1177/0956797612438733

Sher, S., & McKenzie, C. R. (2014). Options as information: Rational reversals of evaluation and preference. *Journal of Experimental Psychology: General, 143,* 1127–1143. http://dx.doi.org/10.1037/a0035128

Slovic, P. (1995). The construction of preference. *American Psychologist, 50,* 364–371. http://dx.doi.org/10.1037/0003-066X.50.5.364

Steenfeldt-Kristensen, C., & Thornton, I. M. (2013). Haptic choice blindness. *i-Perception, 4,* 207–210. http://dx.doi.org/10.1068/i0581sas

Strandberg, T., Björklund, F., Pärnamets, P., Hall, L., & Johansson, P. (2018). *The self-transforming survey*. Manuscript in preparation.

Strandberg, T., Olson, J. A., Hall, L., Raz, A., & Johansson, P. (2018). *Opening American minds: False beliefs can induce open-minded evaluations of presidential candidates*. Manuscript submitted for publication.

Taber, M., & Lodge, C. S. (2013). *The rationalizing voter*. New York, NY: Cambridge University Press.

Taber, M., Lodge, C. S., & Glathar, J. (2001). The motivated construction of political judgments. In J. Kuklinski (Ed.), *Citizens and politics: Perspectives from political psychology* (pp. 198–226). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511896941.010

Taya, F., Gupta, S., Farber, I., & Mullette-Gillman, O. A. (2014). Manipulation detection and preference alterations in a choice blindness paradigm. *PLoS ONE, 9,* e108515. http://dx.doi.org/10.1371/journal.pone.0108515

Tormala, Z. L., & Petty, R. E. (2002). What doesn't kill me makes me stronger: The effects of resisting persuasion on attitude certainty. *Journal of Personality and Social Psychology, 83,* 1298–1313. http://dx.doi.org/10.1037/0022-3514.83.6.1298

Warren, C., McGraw, A. P., & Van Boven, L. (2011). Values and preferences: Defining preference construction. *WIREs Cognitive Science, 2,* 193–205. http://dx.doi.org/10.1002/wcs.98

Watts, W. A. (1967). Relative persistence of opinion change induced by active compared to passive participation. *Journal of Personality and Social Psychology, 5,* 4–15. http://dx.doi.org/10.1037/h0021196

Wolfe, M. B., & Williams, T. J. (2017). Effects of text content and beliefs on informal argument evaluation. *Discourse Processes, 54,* 446–462. http://dx.doi.org/10.1080/0163853X.2017.1319654

Zaller, J. R. (1992). *The nature and origins of mass opinion*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511818691

Received May 9, 2017
Revision received June 27, 2018
Accepted July 3, 2018 ■