ABSTRACT

Delphi is an established information gathering and forecasting approach that has proven to deliver valuable results in a variety of fields. Yet, Delphi studies have also continuously been subject to critique, particularly concerning their judgmental and forecasting accuracy. This can be attributed to the substantial discretion researchers have in their design and implementation. Awkwardly designed Delphi studies may lead to cognitive biases that adversely affect the research results. This paper takes a cognitive perspective by investigating how different cognitive biases take effect within Delphi studies and how their unfavorable impacts can be mitigated by thoroughly adapting specific Delphi design features.

KEYWORDS: Delphi, cognitive biases, framing, anchoring, desirability bias, bandwagon effect, belief perseverance, accuracy

INTRODUCTION

Delphi processes have been used for decades in a variety of fields and methodological variations. As a structured information gathering and forecasting approach it still enjoys unabated interest as indicated by recent applications (Wester & Borders 2014; Álvarez et al. 2014) and design considerations (e.g. Förster & von der Gracht 2014; Gallego & Bueno 2014). Delphi studies regularly deliver accurate and valuable results (e.g. Holmes et al. 2002; Parente & Anderson-Parente 2011) but continue to be criticized as well. The major concern of practitioners and academics is Delphi’s judgmental and forecasting accuracy (Shanshan et al. 2014; Parente & Anderson-Parente 2011; Fildes & Goodwin 2007). Researchers investigating the impact of different design features, e.g. statistical vs. argumentative feedback, on Delphi results’ accuracy found contradictory results (e.g. Rowe et al. 2005; Rowe & Wright 1996). However, these studies frequently do not apply a strong cognitive perspective on Delphi processes, i.e. they do not link the design choices to cognitive processes and biases they may cause or mitigate. Therefore, we argue that more conceptual and empirical work in this area is required as Delphi’s accuracy depends on i) how researchers use (or abuse) their high degree of discretion in terms of study design and execution (Rowe & Wright 1999; Story et al. 2001), and ii) to which extent several cognitive biases take effect at different stages of the process; the latter being to a large part dependent on the former.

As research in psychology and cognition sciences has identified hundreds of biases that could potentially take effect in some Delphi constellation as well, it is beyond the scope of a single study to provide a comprehensive overview on all cognitive biases without being overly superficial. We therefore decided to elaborate on the four cognitive biases encountered by Delphi participants that seem to be most frequent and most impactful in Delphi applications,
namely framing and anchoring, the desirability bias, the bandwagon effect, and belief perseverance. We believe that researchers controlling for these biases via specific Delphi design decisions could not increase accuracy much further by controlling for additional cognitive biases. Following our focus on participants' cognitive biases, we do not address other issues such as sampling biases, i.e. the selection of proper experts, that have been studied elsewhere (e.g. Okoli & Pawlowski 2004; Rowe & Wright 1999). Concerning the kind of Delphi studies, our analysis is focused on expert-based future-oriented studies looking at least five years ahead from today. For researchers applying other kinds of Delphi studies, for example laymen assessing contemporary almanac questions, other cognitive biases might be of relevance.

We structure our analysis along the typical process steps of a Delphi study, and indicate which biases may occur at which stage of the process. As illustrated in figure 1, framing and anchoring as well as the desirability bias impact experts' first estimates, while the bandwagon effect and belief perseverance come into effect during stage 4 which includes feedback and potential revisions of estimates. As participants are usually not involved in process steps 1, 2, and 5, cognitive biases on their side only occur in stages 3 and 4.

![Figure 1: Delphi process steps and biases.](image)

By bridging literature on Delphi research and the fields of cognition and psychology we hope to make a methodological contribution that is of value to both academics and practitioners applying Delphi studies in a variety of fields by a) discussing different cognitive biases and their modes of operation during Delphi applications, b) elaborating on the impact of certain design choices on the prevalence of cognitive biases in Delphi processes and c) developing design recommendations that aim to mitigate or avoid the negative effects of cognitive biases and work towards increasing Delphi accuracy.

**DELPHI**

The Delphi methodology is a structured, interactive group communication and judgmental forecasting process aiming at systematically exchanging informed opinion concerning an
uncertainty-bearing field of interest among a panel of selected experts and developing consensual understanding that reduces uncertainty and finally enhances decision quality (Hallowell & Gambatese 2010; Donohoe & Needham 2009; Dunn 2004; Linstone & Turoff 1975). In future-oriented Delphi studies the field of interest may concern issues lying as far as several decades ahead. Delphi rests on the assumption that structured group approaches provide more accurate judgments than a single expert (Donohoe & Needham 2009; Linstone & Turoff 1975) and are more appropriate than traditional meetings (Graefe & Armstrong 2011). Another underlying assumption is that, even in highly uncertain environments, some features of the future are predetermined and foreseeable (Walsh 2005).

A typical Delphi proceeds as follows. After designing a survey of questions or projections (Klenk & Hickey 2011) it is sent to a group of experts, each of whom provides individual evaluations, ratings or rankings (Chiravuri et al. 2011), e.g. concerning the probability or feasibility of the items under investigation (Klenk & Hickey 2011). Additionally, experts may be asked to provide qualitative arguments supporting their individual estimates (Graefe & Armstrong 2011). Having received all answers, the Delphi administrators consolidate and analyze the contributions (Klenk & Hickey 2011) and feed the results back to the experts, sometimes with a reworked questionnaire. Respondents are asked to review the estimations (and arguments, if any) of the other anonymous participants (Hallowell & Gambatese 2010), encouraged to reconsider their own contributions (Landeta & Barrutia 2011; Sharfman & Shaft 2011; Linstone & Turoff 1975) and given the opportunity to revise their estimates (Georgantzas & Acar 1995; Rowe & Wright 1999). This process can be repeated several times until a pre-determined criterion, e.g. a certain level of consensus, is met (Klenk & Hickey 2011).

Such a procedure comprising at least one round of reconsidering and possible adaption of prior estimates (Landeta & Barrutia 2011) constitutes the iterative character of Delphi that allows for accuracy-improving social learning (Dunn 2004; Hallowell & Gambatese 2010) and the reduction of noise (Strauss & Zeigler 1975) but also bears the risk of cognitive biases taking unfavorable effect as discussed below. Besides its iterative fashion the key characteristics of Delphi are controlled feedback, and the anonymity of participants (von der Gracht 2008; Elliott et al. 2010; Story et al. 2001; Rowe & Wright 1999; Georgantzas & Acar 1995; Yang et al. 2012). Controlled feedback means that the Delphi administrators decide on how feedback is provided and which aspects of the group’s responses are included (von der Gracht 2008).

Anonymity of participants is probably the most controversially discussed characteristic of Delphi as it brings along a number of advantages but also drawbacks. In general it is said that Delphi uses the positive attributes of structured group interaction while mitigating or avoiding the negative social, psychological, and power effects of direct confrontation (Kauko & Palmroos 2014; Graefe & Armstrong 2011; Klenk & Hickey 2011). To be more concrete, anonymity avoids experts’ statements to be biased by dominant personalities, panelists from higher hierarchy level or social status (“halo effect”), or such with strong oratorical abilities (Landeta & Barrutia 2011; Tersine & Riggs 1976). Furthermore, anonymity creates a free thinking space that reduces the unwillingness to give estimates on uncertain issues (Häder 2002), encourages to express and challenge unconventional opinions and alternative viewpoints (Roxburgh 2009; Donohoe & Needham 2009), and offers the opportunity to change a stand once taken without losing face (von der Gracht 2008; Rowe & Wright 1999; Okoli & Pawlowski 2004).

Delphi is best suited to fields and circumstances of application where objective factual data is scarce (Gray & Hovav 2008; Daft & Lengel 1986) and knowledge necessary to make profound decisions is incomplete (Skulmoski et al. 2007; Amos & Pearse 2008). Delphi is highly valuable
in situations of severe uncertainty stemming from rapidly unfolding, non-calculable dynamics, or uncertainty originating from large multidisciplinary problems in highly complex environments (Yang et al. 2012; Donohoe & Needham 2009; Ziglio 1996). In these situations precise analytical data processing techniques are not applicable (Melnyk et al. 2009; Donohoe & Needham 2009; Ziglio 1996) and trend extrapolation is mostly inadequate (Melnyk et al. 2009; Linstone & Turoff 1975). Instead, information collection and knowledge must be built on informed opinion and subjective expert judgments as well as experience-based interpretations (Yang et al. 2012; Melnyk et al. 2009; Linstone & Turoff 1975).

The major disadvantages attributed to the Delphi methodology comprise difficulties in assessing participants’ level of expertise, the potential of anonymity and iteration to lead to compromise rather than consensus, and limitations in assessing result accuracy and reliability – particularly when an issue in the long-term future is investigated (Story et al. 2001).

Comparisons of Delphi and other techniques in terms of accuracy came to discordant results. Although there are several examples of Delphi studies delivering accurate results (Czinkota 1986; Czinkota & Ronkainen 1992, 1997, 2005; Gray & Hovav 2008; Holmes et al. 2002; Parente & Anderson-Parente 2011), and some researchers come to the conclusion that “Delphi’s effectiveness over comparative procedures, at least in terms of judgmental accuracy, has generally been demonstrated” (Rowe et al. 2005: 378), others state that “In comparison to other techniques aimed at enhancing judgmental accuracy, again Delphi’s worth has not been convincingly demonstrated” (Windschitl et al. 2010: 239). In their review of the literature, Rowe and Wright (1999) found that Delphi outperformed both the statistical average of participants’ estimates and face-to-face discussions in terms of accuracy. Similarly, recent studies report Delphi’s superiority compared with staticized groups, i.e., simple one-round surveys (Graefe and Armstrong 2011, Parente et al. 2005, Song et al. 2013). On the other hand there are several researchers doubting Delphi’s accuracy and reliability (e.g. Simoens 2006, Ayton et al. 1999). The equivocal nature of these results is indicative of Delphi’s high flexibility in terms of design features and application (Rowe et al. 1991). Hasson and Keeney (2011: 1701) speak of the “‘greyness’ of the technique, viewed by some as a key benefit allowing flexibility in its application but leading to serious repercussions for the technique’s scientific respectability”.

Several researchers argue that Delphi critique and findings shedding doubt on Delphi’s accuracy are not induced by the Delphi method itself but by examples of inappropriate application (Rowe & Wright 1999; Landeta 2006; Wakefield & Watson 2014). Among others, Rowe and Wright (1999) name features such as the use of non-experts like students, professionals from a single domain, or almanac questions (i.e. not addressing uncertain issues) as examples for misapplication (Rowe and Wright 1999). At this point of the dispute over Delphi accuracy we want to take Delphi’s flexibility and researchers’ discretion to the positive by recommending appropriate design choices that mitigate the negative impact of cognitive biases and work towards increased accuracy of future-oriented Delphi studies.

BIASES IN INITIAL ESTIMATES

We first elaborate on anchoring and framing as well the desirability bias because such cognitive biases predominantly take effect on experts’ initial estimates, i.e. in stage 3 of the standard Delphi process. We describe how they take effect, discuss in how far they are mitigated by usual Delphi design features like feedback processes, and, in particular, how they can be further controlled by specific design choices. We conceive anchoring and framing to be related as they
share a mode of operation where pre-given information known to the experts influences their judgment of some contextually, temporally or otherwise related issue.

**Framing and Anchoring**

**Framing**

Framing refers to the phenomenon that a modification in the presentation of an issue under consideration changes people’s assessment of this issue. That means that – without changing the issue itself – a different depiction of it causes the characteristics, likelihood of appearance, favorability (Cheng & Wu 2010) or degree and nature of impact etc. to be estimated differently than by applying a different depiction (frame) (Yaniv 2011). Cheng and Wu (2010) distinguish three different kinds of framing: Attribute framing whereby a single attribute of an issue or object is framed positively or negatively; goal framing which refers to the – positively or negatively framed – consequences of an issue; and risky choice framing which occurs when the choice between a risky and a riskless option of equal expected value is influenced by the applied frame (Cheng & Wu 2010). Framing violates the normative principle of description invariance which requires rational decisions to be invulnerable to superficial changes in an issue’s description, i.e. its depiction in terms of gains or losses (Kahneman & Tversky 1984).

In the context of Delphi studies – and other group judgment processes – framing may be particularly disturbing as frames shared by group members tend to be magnified, i.e. individuals’ pre-discussion attitudes are amplified by group interaction and lead to polarized group judgments (Whyte 1989; Paese et al. 1993). Judgments made by groups composed of like-minded individuals tend to be more extreme than the average of individuals’ judgments (Isenberg 1986; Myers & Lamm 1976). Kerr and Tindale (2011) speak of dysfunctional shared representations that bias group judgment by amplifying the framing effect. For example, a group composed of individuals disposed toward risk is likely to reach a consensus that is even riskier than the average of the individual pre-discussion dispositions (Whyte 1989).

Concerning the effect at work, informational influence theories suggest that in groups of similarly framed members there is a preponderance of arguments supporting the dominant initial position which will therefore be amplified (Dequech 2006). The mode of operation lying behind is similar to that of ‘confirmatory search’ (Chapman and Johnson, 1994) and ‘selective accessibility’ (Strack and Mussweiler, 1997) as also described for anchoring and belief-perseverance below. Groups of similarly framed members only have access to others sharing their perspective and are only confronted with confirmatory arguments, and they typically do not search for and access contradictory accounts from outside the group. These effects will increase both the judgment itself to a more extreme level and also group members’ (potentially unjustified) confidence in the accuracy of their own judgments (Bolger & Wright 2011).

First empirical proofs of the framing bias were provided by Tversky and Kahneman (1981) who described measures against a disease either in terms of the likelihood of lives saved or the likelihood of lives lost. Depending on the frame used the relative attractiveness of options was evaluated significantly different by participants of their experiment (Tversky and Kahneman 1981).

Empirical research also provides evidence for individuals preferring the risk-free option significantly more often when the options are positively framed in terms of gains as opposed to a negative framing in terms of losses (cf. Yaniv 2011). Concerning the polarization effect in
group judgments studies yielded diverging results. While Neale et al. (1986) found that group interaction mitigated the individual level framing effect, Milch et al. (2009) reported that individual framing effects were neither amplified nor hampered in groups. Paese et al. (1993) found that individual frames were amplified in groups that were presented the same frame, and reduced when the group was presented the opposite frame of the individual. Recently, Yaniv (2011) reported similar effects by distinguishing between homogeneous groups composed of similarly framed members and heterogeneous groups with divergently framed individuals. “[H]omogeneous groups’ preferences were polarized, and thus the framing effect was amplified; in contrast, the heterogeneous groups’ preferences converged, and thus the framing effect was reduced to zero” (Yaniv 2011: 41).

Empirical research identified some variables such as the level of involvement (Maheswaran & Meyers-Levy; Rothman et al. 1993) and attention (Sieck & Yates 1997; Smith & Levin 1996) that influence the magnitude of framing effects. Involvement – defined as “a person’s perceived relevance of the object based on inherent needs, values, and interests” (Zaichkowsky 1985: 342) – seems to have a positive effect, i.e. more involved Delphi participants are less susceptible to framing effects (Cheng & Wu 2010). These findings are supported by Wright & Goodwin (2002) who reported significant framing effects for undergraduates (low involvement) and no framing effect for experienced respondents (high involvement).

The most important and most obvious cure against the polarization of framing effects in Delphi panels is group heterogeneity. As group judgments were shown to be significantly better or worse than individual judgments depending on the diversity of group members’ frames (Yaniv 2011), Delphi participants should be selected to be as heterogeneous as possible to ensure that frames brought into the group interaction process by individual panelists are confronted by diverging frames eventually resulting in the neutralization of any framing bias (Yaniv 2011; Paese et al. 1993). Otherwise shared individual frames might even be amplified during group interaction (Paese et al. 1993).

We recommend ensuring a high degree of heterogeneity within Delphi panels and the inclusion of participants known to have a maverick perspective on issues under investigation (Bolger & Wright 2011) as they add fruitful controversy to the group interaction which works against a framing bias. We even propose to sort arguments in a way that such stemming from known mavericks are listed on top. This way even busy participants that only read the first few of their fellow panelists’ arguments are confronted with controversial input and incentivized to rethink their – potentially framed – views. For real-time Delphi studies (Winkler et al., 2015) mavericks could be invited first so that each following participant receives their fruitfully provoking inputs. We further suggest not applying pyramid search of experts, i.e. a search method where the already identified experts recommend other experts (Baur et al., 2012). Friends, colleagues and other contacts might be similarly framed with regard to several aspects, leading to a “cozy group of like-thinking individuals which excludes mavericks and becomes a vehicle for inbreeding” (Linstone 1975: 568).

Cheng and Wu (2010) provide interesting insights on how warning works against framing. They found that framing effects were attenuated with weak and completely eliminated with strong warning conditions (Cheng & Wu 2010), i.e. warnings indicating that a decision or judgment might be subject to a framing effect. Referring to the impact of involvement they found that weak warning messages were sufficient to eliminate framing effects for high involvement participants but stronger warning messages were needed to eliminate the framing effect for low involvement participants (Cheng & Wu 2010). Similarly Wright and Goodwin (2002) show that the framing
effect (which they only found for non-expert participants) can be eliminated by “think harder
manipulations” (p. 1059). The warning about the possibility of a bias – potentially including a
description of its direction (Cheng & Wu 2010) – makes people think more thoroughly about
their judgments (Sieck & Yates 1997; Smith & Levin 1996). Participants suspected to have fallen
victim to framing effects may also be prompted to provide or consider arguments countering
their presumably biased view (Baron 2003). Daft and Lengel (1986) found that “the failure to
think of arguments on the other side is typically not the result of not knowing them” (Hallowell &
Gambatese 2010: 105). Further, Delphi panelists should have a high level of expertise as this
provides a higher level of involvement and attention, conditions that were shown to effectively
mitigate framing effects.

---

**Anchoring**

The anchoring bias, or anchoring-and-adjustment as it was termed by Tversky and Kahneman
(1974), refers to a situation where the value of a variable is estimated or forecasted by referring
to a known previous value of that same variable (Harvey 2007). People making estimates often
start off with an anchor, i.e. an initial value – available from their own knowledge or given to
them in the context of the judgment task – and then adjust it in order to yield the final answer
(Tversky & Kahneman, 1974; Harvey 2007). While this heuristic can be very effective and
efficient, the result is often biased because the anchor value is adjusted insufficiently (Harvey
2007; Furnham & Boo 2011) as people overweight the anchor and underweight other
information (Campbell & Sharpe 2009). Accordingly, “predictions by individuals systematically
deviate too little from seemingly arbitrary reference points” (Campbell & Sharpe 2009: 371).

---

> Theory

Insufficient adjustment can be attributed to several mechanisms. Proponents of the scale
distortion theory of anchoring (Frederick & Mochon 2012) suggest that anchors do not affect
one’s beliefs about an issue under judgment but rather the response scale on which judgments
are considered. They provide the example of a 70-mile anchor provided to people asked to
guess the length of the Mississippi river (2,320 miles). Provided with an anchor as small as this,
people’s guesses tend to be on a much lower scale (i.e. only some hundred miles) compared to
non-anchored people who, on average, provide significantly larger, i.e. more accurate,
estimates (Mochon & Frederick 2012).

Some researchers suggest anchoring processes to be akin to satisficing (Simon 1979), i.e. one
moves from the anchor in the direction of the correct value and stops at the first value which
seems to be a plausible answer (Chapman & Johnson 2002; Epley & Gilovich 2005).
Consequently, persons provided with an anchor above (below) the true value systematically
come to an estimate higher (lower) than the true value. That explains why participants in
Tversky and Kahneman’s (1974) experiment that were asked for the number of African
countries in the UN estimated it to be on average 25 (when provided with an anchor of 10), and
on average 45 (when provided with an anchor of 65). The answers of both groups are biased
towards the respective anchor (George et al. 2000).

An alternative, or complementary, explanation is provided by confirmatory search and
selective accessibility arguments (Chapman & Johnson 1994, 1999; Mussweiler and Strack,
1999; Strack and Mussweiler, 1997). They suggest that people only access and consider
information that is consistent with the anchor that they started off with (Furnham & Boo 2011).
Hence, they disregard information that could serve as an incentive to move further away from the anchor.

Mussweiler and Strack (2001) also conceptualize anchoring as a form of **availability or accessibility** bias and provide multiple arguments. “[A]nchoring effects are indeed knowledge accessibility effects in essence” (Mussweiler and Strack 2001: 238). Epley and Gilovich (2001, 2005) find that selective accessibility is particularly relevant when anchors are not self-induced but externally provided by the experimenter or some other external source. Mussweiler and Strack (2001) demonstrated that anchoring effects may also occur for non-numeric stimuli and that this semantic anchoring can even be more potent than purely numeric effects. This view is supported by Oppenheimer et al. (2007).

**Empirics**

Empirical research reported anchoring biases in performance judgments (Thorsteinson et al. 2008), time estimation (Thomas and Handley 2008), and probability estimates (Plous 1999); both within and outside the laboratory. Several studies showed that even irrelevant, contextually independent anchors may cause an anchoring bias (Tversky and Kahneman 1974; Englich and Mussweiler 2001; Englich et al. 2006; Critcher and Gilovich 2008). Empirical findings concerning the impact of implausible and extreme anchors are controversial. Some researchers argue that these anchors lead to a larger anchoring effect than plausible anchors (Strack and Mussweiler 1997; Wegener et al. 2010). In contrast, Mussweiler and Strack (2001) and Wegener et al. (2001) found no increase in anchoring bias when anchors were manipulated to be extreme and beyond the range of plausible values. They argue that people generate counterarguments or ignore the values if they are too extreme (Wegener et al. 2001).

Mussweiler and Strack (2001) found that the provision of a **very implausible anchor** may lead to a contrast effect which is the opposite effect to anchoring, i.e. an over-adjustment of the anchor. They provide the example of experiment participants asked for the mean winter temperature in Hawaii. Estimates of participants provided with an anchor of -50°C were higher than those provided with a substantially higher anchor. The applicability of the anchor to the judgment task seems to make the difference (Mussweiler and Strack 2001).

**Influence variables**

Several situation-specific and person-specific characteristics were found to influence the level of anchoring. Anchoring effects were found to be more pronounced in case of high ambiguity, low familiarity with the issue under judgment, and a more trustworthy source of the anchor (van Exel et al. 2006). People being in a sad mood (Bodenhausen et al. 2000; Englich & Soder 2009), having high conscientiousness, high agreeableness, or low extraversion (Eroglu & Croxton 2010), or having high openness-to-experience (McElroy & Dowd 2007) were found to be more susceptible to an anchoring bias. Participants with high expertise (Wilson et al. 1996) and high cognitive abilities (Bergman et al. 2010, Stanovich and West 2008) were found to be less susceptible to anchoring. However, the anchoring bias cannot be eliminated; there are still significant anchoring effects for experts with high degrees of knowledge (Englich & Mussweiler 2001; Englich et al. 2006) and high cognitive abilities (Bergman et al. 2010; Oechssler et al. 2009).

**Cure**

How to then design a Delphi study to mitigate anchoring effects as much as possible? Just as with framing, a very effective cure for anchoring in Delphi studies is ensuring sufficient
heterogeneity among the panelists, so that very different anchors are in play, and group polarization is avoided (Belsky and Gilovich 1999).

Several studies found warnings to be effective in mitigating anchoring effects. George et al. (2000) reported that experiment participants who received warnings provided final values further away from the anchor than those who did not receive any warning (George et al. 2000). Although their findings were not statistically significant they were supported by other studies that found significant reductions in anchoring effects following warnings (LeBoeuf & Shafir 2009; Block & Harper 1991).

However, anchoring could not be eliminated completely through warnings (Tversky and Kahneman 1974; Wilson et al. 1996; Block & Harper 1991; George et al. 2000). George et al. (2000) showed the anchoring bias to be robust within decision support systems. Epley and Gilovich (2005) find that forewarning is only effective with self-generated anchors; they argue that those are known to be inaccurate by the persons who apply them.

Consequently, we recommend the use of warnings making Delphi participants aware of their own potential anchors. Furthermore, we suggest avoiding the provision of any kind of anchor in Delphi studies. Particularly, when experts are asked for future developments, states or characteristics under investigation (e.g. for the year 2025), Delphi administrators at times provide some additional information concerning the current manifestation of the asked future variable (i.e. the according 2014 value). Although this might be a very helpful starting point for the participant’s estimate – particularly in case of lower expertise – it provides a fruitful ground for a strong anchoring bias. Moreover, according to Epley and Gilovich (2005) this kind of anchoring effect can hardly be tackled by warning as the anchor is externally provided by the Delphi itself. Admittedly, “One may argue that it is “better” to provide a reasonable anchor and allow people with strong preferences to deviate, than to loose many respondents and have the remaining respondents using their own (uncontrolled and perhaps irrelevant) anchors” (Van Exel et al. 2006: 849).

In this context we further recommend to use the level of expertise as a strong filter for participation in any Delphi study as the above-mentioned empirical findings show that panelists with high levels of knowledge and cognitive abilities are less susceptible to anchoring effects.

We also propose to present Delphi questions/projections in an order that assures that each question/projection is as unrelated to the preceding one as possible. In this way, the Delphi questionnaire does not provide any anchor or other kind of context-induced disturbance (Hallowell & Gambatese 2010).

Finally, it might be helpful to ask participants for a proper argumentation of their estimates as an indication of whether the provided answer is thoroughly reasoned, which might not be possible for a sloppily adjusted anchor value (Shanshan et al. 2014). Mussweiler et al. (2000) who suggested a consider-the-opposite strategy where respondents are asked to provide an anchor-inconsistent argument in order to debias their responses, found it to be ineffective. However, in some cases it might be helpful as the consider-the-opposite approach works against the mechanisms of confirmatory search and selective accessibility.
Desirability Bias

> Definition
Another bias potentially impacting Delphi results is the desirability bias. It occurs when “participants systematically estimate the probability of occurrence for desirable (undesirable) future projections higher (lower) than the probability for projections with neutral desirability” (Ecken et al. 2011: 1654). Thus, the individual desirability of an event positively influences a person’s likelihood judgment (Krizan & Windschitl 2007). Accordingly, the desirability bias can be seen as a special kind of motivated reasoning (Kunda 1990; Krizan & Windschitl 2007). Delphi results that are affected by desirability could lead to biased decision making and might be inadequate to cope with the developments under investigation (Ecken et al. 2011).

> Mechanisms
Several theoretical concepts have been used to explain the occurrence of a desirability bias. We want to briefly introduce the three concepts that seem most applicable to Delphi studies addressing highly uncertain future developments. First, researchers suggest that confirmatory search and selective accessibility may cause Delphi participants to predominantly select and consider information that is consistent with their personal desires (Krizan & Windschitl 2007). Recently, Lench et al. (2014) raised a similar argument. They suggest that people perceive available information to be more open to interpretation when they are motivated and are therefore more likely to make judgments consistent with their preferences (Lench et al. 2014).

Second, Delphi participants might apply strategic optimism or pessimism. Strategic optimism was found to support people in developing action plans, retain an appropriate level of persistence and deal with negative outcomes (Armor & Taylor 1998), i.e. it is particularly applicable when respondents have some control over the development (Krizan & Windschitl 2007). Others argued that people might sometimes be overpessimistic in their likelihood judgments (Shepperd et al. 2000; Golub et al. 2009). This can be attributed to strategic pessimism that serves to protect them against disappointment when unfavorable outcomes materialize (Krizan & Windschitl 2007; Shepperd et al. 2000). Shepperd et al. (2000) provide the example of financially needy students being more pessimistic about their chances of receiving a scholarship than less needy students. As experts participating in Delphi studies commonly have some stake in the developments that they are to judge within the Delphi (Vosgerau 2010) these strategic optimism or pessimism accounts might well be in play.

Third, humans tend to unconsciously approach “good” events and avoid “bad” events (Ecken et al. 2011; Lench 2009). Some argue this tendency can be ascribed to affective reactions to good (desirable) or bad (undesirable) events (Ecken et al. 2011). Accordingly, judgmental processes like Delphi studies can be affected in terms of estimates biased by the desirability of the event addressed by a question/projection (Ecken et al. 2011). Vosgerau (2010) applies a similar argumentation in the context of his stake-likelihood-hypothesis. He postulates that people having a stake in the outcome of an event misattribute arousal regarding the stake itself to the likelihood of the outcome (Windschitl et al. 2013). He assumes that “w]henever a future outcome is desirable (i.e., denotes a success) or undesirable (i.e., denotes a failure), a decision maker has a stake in the outcome that causes arousal. Arousal is hypothesized to be misattributed to likelihoods, thereby making people more optimistic and more pessimistic depending on what outcome, success or failure, they focus on” (Vosgerau 2010: 34).
Drivers

Desirability bias is suggested to increase with the level of uncertainty and ambiguity as the amount of guessing increases; this is true no matter whether uncertainty is due to dynamism or complexity (Ecken et al. 2011; Armor & Taylor 1998). Another driver promoting the desirability bias is the prospect horizon. Literature suggests that the prevalence of desirability bias increases the farther one looks ahead in the context of the judgmental tasks (Trope and Liberman 2003; Armor & Taylor 1998). In light of shorter prospect horizons people are less likely to fall victim to desirability bias as they are aware that the moment of truth when their estimates are judged against reality is closer (Krizan & Windschitl 2007; Windschitl et al. 2013). Therefore, they will be more careful to select a balanced set of information including information that both support and challenge their desired outcome (Tyler & Rosier 2009; Sweeny & Krizan 2013). Hence, they will provide a more humble and moderate answer.

Empirics

Studies examining whether likelihood judgments are affected by desirability have produced mixed results (Bar-Hillel and Budescu 1995; Krizan & Windschitl 2007; Vosgerau 2010) and the strongest evidence comes from one specific paradigm – the marked card paradigm (Windschitl et al. 2010). Participants predict a marked card to be drawn significantly more often when it is related to some kind of incentive, e.g. a monetary gain (Windschitl et al. 2010).

However, the generalizability of these laboratory results and their practical implications are limited (Ecken et al. 2011) as it was found that optimistic predictions are more common when stochastic outcomes are concerned compared to situations when non-stochastic, epistemic uncertainty is prevalent. The latter is the case in most decision-making situations outside the laboratory (Windschitl et al. 2010). Likewise, the desirability biases found by Windschitl et al. (2010) vanished when they switched from marked-card experiments to scaled likelihood judgments in otherwise unchanged setups.

Apart from the marked-card paradigm there is no strong evidence for desirability bias in Delphi-typical judgment tasks (Bar-Hillel and Budescu 1995; Krizan & Windschitl 2007; Vosgerau 2010; Windschitl et al. 2010, 2013). Krizan and Windschitl (2007) conclude that there is a dearth of studies manipulating the desirability of non-stochastic outcomes, people asked for likelihood estimates very limitedly show a desirability bias, and the average desirability effect in their meta-analysis is significant but small.

Windschitl et al. (2013) provide empirical evidence for the selective accessibility and confirmatory search arguments by showing that people tend to select information that provides additional arguments for the occurrence of their desired outcomes. They suppose that other studies have failed to find a substantial desirability bias because they did not include the selection of additional information in the experimental design (Windschitl et al. 2013).

The validity of the arguments for a relationship between affective reactions and the desirability bias has been shown by Lench (2009). Her empirical findings support the notion that people approach positive and avoid negative outcomes, and that this tendency also affects judgments of the likelihood of future events (Lench 2009). She further show that the desirability bias can be reduced when participants misattribute affective reactions to an object other than the one they make estimations about (Lench 2009).
> Cure
As in the case of framing and anchoring, bias in the results of a Delphi study is supposed to be particularly prevalent when panelists are too similar to each other – in terms of profession, origin, education and other attributes depending on the researched topic. Regarding the desirability bias “the quality of decisions based on Delphi results may be adversely affected if experts share a pronounced and common desirability for a future projection” (Ecken et al. 2011: 1654). Hence, our recommendations of Delphi design features mitigating the desirability bias start with the advice to ensure a proper level of heterogeneity among participants (Ecken et al. 2011).

Given considerable panel heterogeneity, the Delphi-typical feedback and iteration process itself already provides some cure to the desirability bias. Although, Delphi is not generally capable of eliminating it completely, the built-in reconsideration feature causes second and further estimates to be less infected by desirability than first-round estimates (Ecken et al. 2011; Rowe & Wright 1996).

Concerning expert selection it was shown that a high level of expertise does not prevent respondents from the influence of desirability bias (Massey et al. 2011). Yet, Vosgerau (2010) found that arousal is less likely to be misattributed to judgments when participants have the cognitive abilities and motivation to engage in effortful deliberation. We expect this to be more frequently provided by experts in a field of investigation than by non-expert panelists.

Babad et al. (1992) showed that making panelists aware of biases and instructing them to be objective is not effective in reducing the desirability bias. Although research on the effect of warning messages on the desirability bias in predictions is very limited, we hypothesize warnings explicitly pointing to the potential impact of preferences to have some positive effect, particularly in Delphi studies asking participants for less arousal-bearing topics than the prediction of political election outcomes, as was the case in Babad et al.’s (1992) study.

As a means beyond classical warning we suggest to ensure profound understanding of the process and purpose of future-oriented Delphi studies on the part of the panelists. Quantitative inputs are necessary but thoroughly considering other panelists’ inputs and backing the own estimates with rich arguments is the key to gaining valuable insights – for both the Delphi administrators and the participants. Experts being aware of this are assumed to be less susceptible to cognitive biases such as the desirability bias as they are less driven by their desires than by neutral curiosity.

Ecken et al. (2011) suggest asking Delphi participants for their desirability of each projection along with their probability estimates. They suppose a post-hoc procedure to quantify the desirability bias in Delphi studies and apply a statistical approach to adapt panelists’ probability estimates in accordance to the identified level of desirability bias (Ecken et al. 2011). In light of the above-mentioned mixed empirical results for the desirability bias we share the recommendation to quantify the desirability bias in Delphi studies and suggest to apply the post-hoc procedure in case of an unacceptable degree of desirability bias. As the desirability bias was shown to increase with the time left to the occurrence of the specific event (e.g. Trope & Liberman 2003), this will be particularly advisable in case of long prospect horizons. In this regard we also suggest not to extent the prospect horizon farther than necessary for the issue under investigation as this will arguably increase uncertainty and promote the biasing effect of desirability.
BIASES IN OPINION CHANGE

Any improvement in Delphi accuracy occurring after the initial estimates must necessarily stem from opinion change during subsequent iterations. Accuracy improvement through the revision of initial estimates requires that not all panelists perform an equally large change towards the average of first-round estimates (Bolger & Wright 2011). Unless the average of first round estimates and the accurate value are very close this would only increase (mistaken) consensus but not accuracy (Bolger & Wright 2011). Accuracy improvement during second and later iterations requires some participants to change their initial judgment more than others (Bolger & Wright 2011).

The theory of errors – which constitutes a basic principle underlying Delphi – postulates that more knowledgeable participants that provided more accurate estimates during the first round change their opinion less than panelists that submitted more erroneous inputs (Dalkey 1975). It is supposed that respondents who are aware of the fact that they have limited expertise or applied little consideration are more willing to modify their response (Kauko & Palmroos 2014). In parallel, panelists that provided the most accurate inputs ideally do not change anything (Kauko & Palmroos 2014). If this logic applies the iterative process leads to stepwise accuracy improvement of the group estimate. Provided that there are sufficient rounds allowing the less expert panelists to change their opinions as much as they are prepared to, Delphi delivers accurate results (Bolger & Wright 2011). However, there are some cognitive biases at play that might prevent the mechanism from working properly. The question arising is formulated in Rowe and Wright (1999: 139): “Which individuals change their judgments in response to Delphi feedback – the least confident, the most dogmatic, the least expert?”.

Empirical research provides evidence for the theory of errors. Hussler et al.’s (2011) empirical study found that experts’ estimates remained rather stable even when confronted with contradictory judgments. Only 3% of opinions given by experts were changed during the second round, as opposed to 21% of the laypersons’ responses (Hussler et al. 2011). Rowe and Wright (1996) found a significant negative correlation between the level of objective expertise and the propensity to change initial estimates. Rowe et al. (2005) report that participants who provided more accurate probability forecasts in the first round exhibited least opinion change afterwards. Similar results are reported by Yaniv and Milyavsky (2007). However, the mechanism necessary for the theory of errors to be fulfilled cannot be taken for granted. Doubt concerning Delphi and the theory of errors is promoted by findings that “change toward the fed-back value also occurs when this value is false” (Woudenberg 1991: 140). Such unfavorable dynamics may occur when it is not the least knowledgeable Delphi participants that change their mind but, for instance, the least confident.

If confidence – conceived as “subjects’ beliefs about their decisional performance and their perception of the probability that their decisions are correct” (Huang et al. 2012: 440) – is well related to expertise, accuracy will improve when the least confident ones change their opinion in the direction of the estimates of the more confident (806/1501). Yet, empirical research yields mixed results. While some find that the propensity to change increases with decreasing confidence (Bolger et al. 2011; Bolger & Wright 2011) and that more confident judges tend be followed more (e.g. Sniezek & Van Swol 2001; Van Swol and Sniezek 2005), others explicitly do not find any relation between confidence and opinion change (Rowe et al. 2005; Rowe & Wright 1996). Even worse, empirical studies show that there is no significant relation between confidence and expertise/accuracy (Rowe et al. 2005; Rowe & Wright 1996, 1999). It seems confidence is more related to status (Bolger & Wright 2011) or personality traits like self-esteem.
(Rowe et al. 2005; Bolger & Wright 2011) than to expertise. That means accuracy is not necessarily increased and might even be adversely affected if the least confident participants are the ones that change their minds most. Hence, as confidence does not seem to be a proper indicator for accuracy another Delphi design recommendation is not to include any indication of confidence in Delphi feedbacks (Rowe et al. 2005; Bolger et al. 2011).

It was proposed that feedback – quantitative or argumentative – is a more reliable indicator of accuracy than statements of confidence (Bolger & Wright 2011). In general, feedback is found to be a valuable corrective feature for inaccurate first-round estimates by less knowledgeable or biased respondents (Rowe et al. 2005; Yaniv & Milyavsky 2007; Bonaccio and Dalal 2006). Yet, research on which kind of feedback provides the best indication of expertise and the most improvement of accuracy yields conflicting results. While Rowe and Wright (1996) found greater accuracy improvement for qualitative feedback (only arguments, no statistics), Rowe et al. (2005) found that qualitative feedback had no benefits – neither in terms of opinion change nor accuracy. Bolger et al. (2011) found similar results to Rowe et al. (2005). Equally contradictory, the iteration condition (just another Delphi round, neither quantitative nor qualitative feedback) once improved accuracy (Rowe & Wright 1996), and once did not (Rowe et al. 2005).

Unfortunately, Rowe and colleagues did not include a feedback that combines quantitative and qualitative features which could be recommendable as it provides the most comprehensive information. Further, they did not link their results to the distinct impact and mode of operation of cognitive biases. We aim to provide further insights on the impact of design features in the feedback-and-revision phase by applying a strong cognitive bias lens. We propose that people show substantial flexibility in selecting and processing new information (Hart et al. 2009; Jonas et al. 2006). Hence, these processes can be considerably impacted by biases (Jonas et al. 2006; Carlson & Russo, 2001; DeKay et al. 2009; Russo et al. 2008; Russo et al. 1996). Particularly, it may occur that people (with accurate first-round estimates) follow majority opinion too much (bandwagon effect), or participants (with inaccurate first-round results) change their opinion too little (belief perseverance). We will elaborate on these two biases in the next sections.

**Bandwagon Effect / Groupthink**

> **Definition**
> The bandwagon effect refers to the phenomenon that at times a person’s behavior – including decision behavior – strongly conforms to the behavior of a group or the majority of a group, merely based on the information that this thinking or behavior is pursued by the majority (Zimmermann et al. 2012). It can be seen as a pressure to adopt oneself to a standard belief or conduct (Hallowell & Gambatese 2010). In terms of Delphi processes this means that participants go along with the majority opinion instead of championing their own take (Tsikerdekis 2013). This phenomenon seems to be particularly prevalent in case of a strong initial group preference (Henningsen et al. 2006).

> **Theory / Mechanisms**
> Several theoretical arguments have been suggested in order to explain why concurrence seeking may be stronger than accuracy seeking (Bolger & Wright 2011). First, Janis’ (1973) groupthink theory is frequently cited in the context of bandwagon effects. Janis argues that people adapt to group opinion out of a moral obligation grounded in group loyalty (Janis 1973). Consequently group members do not raise controversial issues or question weak arguments
(Janis 1973), particularly in case of high group cohesiveness and homogeneity, insulation from outside experts, authoritarian leadership, and lack of methodical decision-making procedures (Janis 1973, 1982). The theory – which has been subject to harsh critique (e.g. Fuller & Aldag 1998) – is grounded on retrospective sensemaking of disastrous political decisions by face-to-face groups. It seems only limitedly applicable to Delphi panels for which the above-mentioned conditions are usually not fulfilled (e.g. Fuller & Aldag 1998).

Second, some researchers have adduced social pressure arguments. As Delphi participants remain anonymous social pressures should be less immediate and strong than in face-to-face groups (Rowe et al. 1991; Bolger & Wright 2011). Yet, they may not be eliminated entirely but still be felt and taking according effect. Due to social comparison processes and social desirability objectives people strive to perceive and present themselves in a favorable light compared to others (Whyte 1989; Myers et al. 1977). Group interaction may therefore motivate group members to follow socially desired or shared opinions (Myers et al. 1977) to minimize conflict and reach consensus. In Delphi studies this may lead to opinion change on the side of those who held minority positions in the first round (Bolger & Wright 2011), no matter how accurate this minority position was. Bolger and Wright (2011) add that those holding outlying opinions are the most likely to feel marginalized and drop out. Yet, as they challenge conventional thinking, they might be the most important participants for reaching accurate results (Bolger & Wright 2011). Particularly in our research context of future-oriented Delphi studies consensus is not an objective worth striving for (Woudenberg, 1991). Considering diverging perspectives, arguments and futures is way more valuable than unanimity (Rowe et al., 2005). Unfortunately, human psychology is first and foremost a pragmatic survival mechanism rather than a truth detection device (Friedrich 1993). Therefore, people might be more concerned with producing desirable outcomes than accurate ones (Nickerson 1998).

Third, uncertainty has been put forward as an explanation for bandwagon effects. In contexts of high uncertainty or ambiguity – like future-oriented issues – people tend to copy the approaches of others (Bolger & Wright 2011; Whyte 1989). This ‘informational social influence’ (Deutsch and Gerard 1955: 630) is particularly promising and prevalent when others are perceived to be experts (Bolger & Wright 2011). As future-oriented Delphi studies usually deal with highly uncertain or ambiguous issues and use expert panels to approach them, informational social influence may be particularly high.

In any of the three explanations, satisficing (Simon, 1979) may play a key role. People accept an easily retrievable answer that fulfills a satisfying level of plausibility and do not engage in any further search or optimizing (Myers et al. 1977). This may lead to mistaken consensus around “the first solution that greatly offends no one, even though no one may agree with that solution wholeheartedly” (Rowe et al. 1991).

> Empirics

Janis’ groupthink theory was partially supported by empirical investigations (e.g. Callaway & Esser 1984; Hodson & Sorrentino 1997). For instance, it was confirmed that group cohesiveness exerts significant influence on the quality of a group’s decision-making with highest quality being reported for groups of intermediate cohesiveness (Callaway & Esser 1984). Several researchers refined Janis’ theory (e.g. Henningsen et al. 2006; Hodson & Sorrentino 1997), e.g. regarding the impact of personality traits like group members’ uncertainty orientation (Hodson & Sorrentino 1997).
With respect to Delphi studies, Rowe et al. (2005) find that “majorities, whether accurate or otherwise, exerted a significant pull on minorities to the consensual position, even when that position was fallacious” (p. 397). This result was irrespective of the nature of the feedback provided (Rowe et al. 2005). Similar results were found by Myers et al. (1977). Bolger et al. (2011) also confirm these results as they find that panelists who provided minority opinions were more likely to change their positions and that the consequential convergence of opinion does not necessarily imply improved accuracy. They even conclude that “majority opinion is the strongest influence on panelists’ opinion change” (Bolger et al. 2011: 1671). Obviously Delphi practitioners must be aware that bandwagon effects may exert substantial influence on Delphi results. Although anonymity was found to limit bandwagon processes (Postmes & Lea 2000; Tsikerdekis 2013) convergence of opinion across Delphi iterations does not necessarily imply increased accuracy (Rowe et al. 2005). Some design features may help to mitigate the bandwagon effect.

> Cure

The major means against the bandwagon effect is the configuration of the provided feedback as it brings the information that causes participants to either change their opinion in the one or other direction, or not. Feedback must provide good cues about where the most accurate answers lie. As elaborated above, confidence or majority do not seem to be reliable indicators in this regard. Although it is an open empirical question how good people are in distinguishing good from bad advice on the basis of rationales (Bolger & Wright 2011), we agree with Bolger et al. (2011) that argumentative feedback can at least to some degree be a good cue to truth. Assuming panelists are able to distinguish good from bad advice, it might be an additional feature to let experts rate their fellow panelists qualitative inputs by quality and let the Delphi software sort arguments by this rating, i.e. the most persuasive arguments are positioned on top of the list, irrespective of whether they support or challenge a specific opinion.

Even though the studies by Rowe and Wright (1996), Rowe et al. (2005) and Bolger et al. (2011) have reported at best mixed results for the accuracy-improvement potential of qualitative feedback, we believe it to be highly valuable if properly configured. Our hope is backed by findings that “the quality of other panelists’ rationales was significantly positively correlated with the more valid tip” (Bolger et al. 2011: 1678), and that opinion change as a reaction to ‘reasons feedback’ tended to be for the better (Rowe & Wright 1996).

However, argumentative feedback should not be provided to other panelists unfiltered. Similar and duplicate entries should be eliminated in order not to disclose whether the argument is shared by many or not (Bolger et al. 2011). As panelists are not provided with any indication of majority they cannot fall victim to the bandwagon effect; they can only rethink their own estimates on the basis of other arguments’ persuasive power. Here, again, panel heterogeneity is advisable as the inclusion of heterogeneous participants with diverging perspectives challenges conventional thinking and fosters fruitful controversy (Förster & von der Gracht 2014).

Additionally, we recommend not including any statistical information in the feedback. As numerical feedback almost always provides an indication of consensus or majority it does in fact operate as a clear incentive to agree rather than to be accurate (Bolger & Wright 2011). If panelists do not see any mean, median or consensus value there is no obvious bandwagon they can jump on. Outliers – i.e. the ones most relevant to challenge conventional thinking (Bolger & Wright 2011) – are much less likely to feel marginalized, drop out or decide for mistaken opinion change because it is not readily visible they are outliers at all. This way, Delphi
participants can only evaluate the accuracy of their own initial estimates on the basis of argumentative feedback provided by their fellow panelists. Hence, the bandwagon effect is diminished.

Admittedly, this might hamper the formation of consensus. However, we do not consider consensus a proper aim for Delphi studies (Woudenberg 1991), particularly not for future-oriented Delphi studies as addressed in this paper. While many authors consider consensus something worth striving for and conceive the lack of it as undesirable (Elliott et al. 2010; von der Gracht 2008), our elaborations on the bandwagon effect elucidate the frequent tradeoff between consensus and accuracy (Janis 1973; McAvoy et al. 2013). We advocate that neither Delphi administrators nor panelists should have a too strong desire for unanimity but rather strive for stability in group opinion, e.g. two or more opinion clusters, and allow for informative dissent (Rowe et al. 2005).

Belief Perseverance

> Definition
Belief perseverance, or advice-discounting, refers to the observation that decision makers being confronted with unconfirmatory advice overweight their own judgmental performance and underweight (or discount) the available advice (Bonaccio & Dalal 2006). This overconfidence leads to final judgments being significantly closer to one's own estimate than to the advice, even if the advice is more accurate (e.g. Gardner and Berry 1995, Harvey and Fischer 1997, Yaniv and Kleinberger 2000). Similar to anchoring people systematically disregard new information to a certain degree and do not sufficiently move away from a pre-existent point of departure. In its extreme form belief-perseverance means entirely sticking with one's initial estimate while completely ignoring any kind of advice and new information (Yaniv & Milyavsky 2007).

> Theory / Mechanisms
In terms of belief adaption there are two psychological mechanisms competing with one another. On the one hand humans need some degree of stability in their beliefs in order to benefit from past experience (Drake 1983). Moreover, people value consistency and consider it an important cornerstone of rationality (Nickerson 1998). On the other hand a certain amount of flexibility and change in beliefs is necessary to benefit from new experiences and information (Drake 1983). Obviously, there is the frequent tendency to maintain a stand once taken. At times, it even seems that a person's sole objective is to defend and justify the own position (Nickerson 1998; Kauko & Palmroos 2014).

Besides arguments pointing to the primacy effect (Nickerson 1998), i.e. the overweighing of knowledge that was acquired earlier rather than later, and to anchoring-and-adjustment as described above with the judge's initial estimate serving as an anchor which is insufficiently adjusted to the advice (Harvey 2007; Block & Harper 1991; Bolger & Wright 2011; Bonaccio & Dalal 2006), there are three main lines of argumentation explaining the systematic intransigence underlying the belief-perseverance effect: selective accessibility/confirmatory search, information asymmetries, and egocentrism. We will elaborate on these explanations by referring to Delphi literature, cognitive and psychological research as well as literature on judge-advisor systems (JAS) as this research paradigm offers some similarities with Delphi procedures and particularly addresses belief-perseverance.
As in the case of anchoring and framing or desirability, belief-perseverance can be explained by selective accessibility and confirmatory search arguments. Here, the unwillingness to admit the inaccuracy of one’s own prior inputs leads people to selectively access and scrutinize information that is consistent with their initial estimates (Windschitl et al. 2013; Rabin & Schrag 1999) and discount or even ignore unconfirmatory evidence (Huang et al. 2012; Kauko & Palmroos 2014). Further, they perceive and interpret information in a manner supporting their beliefs (Nickerson 1998) and discredit sources of contradictory information (Kulik 1986).

This “hypothesis-based filtering” (Rabin & Schrag 1999: 46), where supporting information becomes appealing (e.g. Scherer et al. 2013; Krizan & Windschitl 2007) and conflicting information becomes dissonance provoking (Hart et al. 2009; Jonas et al. 2006; Kunda 1990), serves to perpetuate one’s self-conceptions (Kulik et al. 1986). With regard to Delphi that means that panelists “selectively attend to feedback that agrees with their stated position and ignore feedback that disagrees with it” (Rowe & Wright 1996: 76). Consequently, more arguments in favor of the initial position are collected, commitment to that position increases, confidence is strengthened, and opinion change becomes less likely (Rabin & Schrag 1999).

A second theoretical explanation for belief-perseverance is information asymmetry. Although the judge’s initial estimate and an advisor’s estimate may seem equally valid from an external perspective, they are not from the judge’s perspective (Yaniv & Milyavsky 2007). The judge has full access to the reasoning and evidence underlying his or her own estimate but only incomplete, if any, insight in the advisor’s rationales (Yaniv & Kleinberger 2000; Yaniv 2004; Yaniv & Milyavsky 2007; Bolger & Wright 2011). As the weight assigned to a judgment depends on the evidence that can be acquired in support of it, the differential information about the underlying justifications causes decision makers to discount the advice relative to their own opinion (Yaniv & Kleinberger 2000).

A third explanation for belief-perseverance is egocentrism (e.g. Yaniv & Kleinberger 2000; Yaniv and Milyavsky 2007). It postulates that “a judge adheres to a default belief in the inherent superiority of his or her own judgment” (Bolger et al. 2011: 1672) simply because it is his or her own (Bolger & Wright 2011). Psychological explanations contend that people do so to protect their ego (Nickerson 1998), appear consistent in social settings and maintain their self-esteem (Yaniv & Milyavsky 2007). Some argue that egocentrism is particularly likely when working with experts as they – being aware of their expert status – might be particularly unwilling to admit that their initial estimates were wildly inaccurate (Kauko & Palmroos 2014). Literature indicates that egocentrism expresses itself via egocentric trimming, i.e. the farther an opinion is away from one’s own the more it is discounted and the more is its source disparaged (Yaniv 2004; Yaniv & Milyavsky 2007). Yaniv and Milyavsky (2007: 105) state that “Trimming is indeed a good strategy that could be used beneficially to improve accuracy, as long as it is conducted objectively rather than egocentrically”.

> Empirics
Empirical evidence shows that taking advice substantially increases accuracy (Yaniv & Milyavsky 2007) and that advice is less discounted by judges that have less experience/knowledge than their advisors (e.g., Harvey & Fischer, 1997) or less than other judges (Harvey & Fischer, 1997; Yaniv & Kleinberger 2000). While this provides some support for the theory of errors, research also shows that judges discount advice to the disadvantage of accuracy gains. Although accuracy increases somewhat due to the discounted usage of advice (Yaniv & Milyavsky 2007), judges by far fail to exhaust the entire accuracy improvement potential (Yaniv & Kleinberger 2000, Yaniv 2004; Yaniv & Milyavsky 2007). Yaniv and
Kleinberger (2000) found that respondents’ and advisors’ estimates were, on average, equally accurate. Yet, just as Harvey and Fischer (1997) they find that own estimates are given a weight around 70% and advisor’s estimates are discounted to a weight of approximately 30% (Yaniv and Kleinberger 2000).

Windschitl et al. (2013) provide clear evidence for the selective accessibility and confirmatory search arguments brought forward earlier. Unlike many other studies they include an “information-buffet paradigm” (Windschitl et al. 2013: 75) in their study, i.e. after making an initial prediction, participants may choose which information out of many they use in order to verify or challenge their own prediction. They find that people significantly favor information supporting rather than contradicting their prediction (Windschitl et al. 2013). Further evidence for similar selective-exposure paradigms is reviewed by Hart et al. (2009) and Jonas et al. (2006).

Empirical research indicates that egocentrism might be more applicable as an explanation for belief-perseverance than information asymmetry as advice-discounting also occurs in novel situations where judges can be assumed not to have any evidence to support their own estimate (Cadinu and Rothbart 1996; Krueger 2003; Bolger & Wright 2011). Further, the egocentrism argument may be more compelling than anchoring-and-adjustment as advice-discounting also occurs when the advice is given before the judge even sees the decision task such that the judge cannot (mis)use his initial judgment as an anchor (Clement & Krueger 2000; Harvey & Harries 2004). Further direct support for egocentrism comes from Bonaccio and Dalal (2006) who found that “decision-makers gave greater weight to someone else’s forecasts incorrectly labeled as their own than to correctly labeled others’ forecasts” (p. 130). Bolger et al. (2011) contend that it is an open research question whether information asymmetry or egocentrism arguments are superior.

Empirical studies on belief-perseverance further show that discounting increases with a growing distance between the judge’s and the advisor’s estimate (egocentric trimming) (Yaniv 2004; Yaniv & Milyavsky, 2007) and growing distance to other advisors’ recommendations (outlying advice) (Harries et al. 2004). Some hope is given by studies reporting that better advice is less discounted than bad advice, although it is still discounted (Yaniv & Kleinberger 2000; Yaniv & Milyavsky 2007; Gardner & Berry 1995). With regard to most Delphi studies, it is particularly relevant to acknowledge that advice discounting is more prevalent in judgment than in choice tasks (Klayman et al. 1999; Soll & Klayman 2004).

> Cure
Concerning the question which design choices may help to mitigate belief-perseverance, we first have to point to the fact that Delphi’s anonymity plays a controversial role in this regard. On the one hand it offers participants the chance to change their opinion without losing face. On the other hand it undermines other participants’ (i.e. advisors’) credibility. People might be reluctant to change their own judgment in light of advice given by advisors whose identity remains undisclosed (Rains 2007; Tsikerdekis et al. 2013; Ziegler et al. 2006).

As empirical evidence showed good advice is less discounted than bad advice (Yaniv & Kleinberger 2000; Yaniv & Milyavsky 2007; Gardner & Berry 1995) belief-perseverance should be countered with high-quality argumentative feedback (Bolger & Wright 2011) in order to signal a high degree of expertise and provide a convincing challenge to others’ beliefs. However, it was found that the majority of rationales provided by Delphi panelists were of low quality (Bolger & Wright 2011; Bolger et al. 2011; Rowe et al. 2005). A large number merely reported unbacked personal views or textually repeated the quantitative estimate (‘I think that
this is more likely …’ (Rowe et al. 2005: 396)) rather than presenting compelling causal reasoning (Bolger & Wright 2011).

Hence, we recommend striving for **enhancing the quality of argumentative feedback**. A simple way to do so is giving participants one or two examples for good and poor reasons, respectively, in order to raise awareness for the significance of each type of input (Bolger & Wright 2011). Additionally, we again recommend not providing feedback in an unfiltered manner. We suggest Delphi administrators to thoroughly review panelists’ arguments and to delete uninformative inputs, i.e. low-quality inputs that do not provide causal reasoning (Bolger et al. 2011). If Delphi administrators succeed in eliciting high-quality arguments to be used as feedback, this may work against several of the above-mentioned mechanisms underlying belief-perseverance, i.e. anchoring, information asymmetry, and egocentrism.

It is argued that Delphi participants’ propensity to generate poor-quality reasoning is fostered by a lack of expertise, lack of motivation and involvement, and the nature of the task (Bolger & Wright 2011). While a proper level of expertise can be assured early on during the expert-selection procedure, motivation and involvement can be raised by offering financial rewards or social incentives like acknowledging the names of the most engaged panelists within the publication of the Delphi results (Bolger & Wright 2011). The rating and ordering of arguments by their quality (persuasiveness) that we already suggested earlier might also be effective here. The nature of the task is usually not susceptible to changes; Delphi studies addressing long-term future issues are obviously ambiguous in nature and limitedly amenable to causal reasoning.

Kauko & Palmroos (2014) suggest a **post-survey adjustment procedure** to mitigate the belief-perseverance bias mathematically. Grounded on the basic insight of belief-perseverance that “Most modifications in individual forecasts will be in the right direction but too small” (Kauko & Palmroos 2014: 315), they scale up modifications in individual forecasts ex post. Change in each answer of each respondent is multiplied by a suitable constant, in their empirical study this constant was 4.39 (Kauko & Palmroos 2014). While the identification of a proper constant remains a considerable challenge, they note that even marginal upscaling with a constant of 1.1 resulted in better forecasts than in the condition without any de-biasing (Kauko & Palmroos 2014).

If the procedure is kept secret so that respondents cannot strategically adjust their answers (Kauko & Palmroos 2014) it may be a promising solution – one being exclusively designed for Delphi studies as other techniques (like face-to-face groups) do not provide the necessary data, i.e. initial and revised estimates from each participant. However, referring back to our earlier elaborations on the bandwagon effect, we note that Delphi administrators would still have to make sure that participants adjust their initial estimates in the right direction. Otherwise the post-hoc procedure could scale up an opinion change away from the accurate value and, say, towards the faulty value provided by a mistaken majority. In order to avoid such undesirable amplification of biases, Kauko and Palmroos’ (2004) procedure would have to be combined with other design features as recommended in this paper such as high expert heterogeneity.

Researchers further suggest the use of **warning and counter-argument** to reduce belief-perseverance bias (Huang et al. 2012). As mechanisms such as confirmatory search are unconscious processes (Nickerson 1998), warning alone may already provide some cure against belief-perseverance. Block and Harper (1991) found that for subjects being warned of the potential influence of the belief-perseverance bias overconfidence was reduced (but not
eliminated). By asking people for counter-arguments Delphi administrators may go one step further. As the quality of estimates largely depends on the variety of information considered during the judgment process (Kray & Galinsky 2003), encouraging panelists to think of alternative hypotheses and counter-argument is a valid way to enhance judgmental accuracy and mitigate belief-perseverance (Nickerson 1998; Huang et al. 2012). This consider-the-opposite intervention (Windschitl et al. 2013) particularly works against confirmatory search as it provides a clear stimulus to consider information that might otherwise be neglected (Huang et al. 2012). The direct provision of counter-argument (by Delphi administrators) as suggested by Huang et al. (2012) is not recommended as it is difficult to identify those experts who are biased (since the true value is regularly unknown). One would have to supply counter-argument to all participants which would hold the risk that accurate estimates are revised as well and non-confident panelists might be more susceptible / responsive to counter-argument than confident panelists – regardless of expertise / accuracy.

Another cure against confirmatory search, and implicitly belief-perseverance, is disfluency, i.e. providing feedback in a disfluent format, as suggested by Hernandez and Preston (2013). They argue that “the effort associated with disfluency prompts a deeper, more analytical, and critical processing of the information itself” (Hernandez & Preston 2013: 178; see also Oppenheimer 2008; Alter & Oppenheimer 2008). While disfluency yields interesting effects in laboratory settings, we are reluctant to advice Delphi administrators to present feedback in a disfluent way as this might cause increased fatigue and drop-outs rather than an increase in accuracy.
**SUMMARY: DESIGN FEATURES COUNTERING COGNITIVE BIASES**

Based on the elaborations above, table 1 summarizes the major design recommendations and illustrates which design features work against which bias(es).

<table>
<thead>
<tr>
<th>Panel composition</th>
<th>Framing &amp; anchoring</th>
<th>Desirability bias</th>
<th>Bandwagon effect</th>
<th>Belief perseverance</th>
</tr>
</thead>
<tbody>
<tr>
<td>high heterogeneity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>inclusion of mavericks</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avoid pyramid search</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>role-playing</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Warning, eliciting counter-argument, ensuring proper understanding of Delphi</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Participants’ traits</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓ / X</td>
</tr>
<tr>
<td>high expertise</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>high cognitive abilities</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high involvement</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-hoc procedure</td>
<td>Ø</td>
<td>✓</td>
<td>Ø</td>
<td>✓</td>
</tr>
<tr>
<td>Unrelated order</td>
<td>✓</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
</tr>
<tr>
<td>Feedback</td>
<td>Ø</td>
<td>Ø</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>argumentative only</td>
<td>Ø</td>
<td>Ø</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no double entries</td>
<td>Ø</td>
<td>Ø</td>
<td></td>
<td></td>
</tr>
<tr>
<td>no non-causal entries</td>
<td>Ø</td>
<td>Ø</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

✓ Positive effect, Ø No effect, X Detrimental effect

**Table 1:** Main design features and their effects on cognitive biases.

The most important design recommendation derived from our cognitive biases perspective on Delphi processes is the composition of a panel that is very heterogeneous, includes mavericks...
and does not apply pyramid search. If heterogeneity cannot be sufficiently ensured by selecting participants, it is recommended to assign roles to participants in order to create artificial heterogeneity (Yaniv 2011; Green and Armstrong 2011). Although panel heterogeneity has been suggested before (Ecken et al. 2011; Mannix & Neale, 2005; Sommers, 2006) our analysis shows that it is valuable in mitigating all of the major biases that are at play in Delphi studies. As heterogeneity provides a wide range of frames, anchors, and desirability perspectives, it enhances initial estimates that “bracket” the real value (Förster & von der Gracht 2014). Some participants’ estimates will be positioned above, some below the accurate value. Such heterogeneity-induced bracketing mitigates biases that would take effect in more homogeneous panels of like-minded participants (Förster & von der Gracht 2014). Ecken et al. (2011) provide an example how heterogeneity leads to bracketing and, hence, to the elimination of the desirability bias. Bracketing equally applies to framing (Yaniv 2011) and anchoring. However, it is especially the different perspectives of such heterogeneous groups that is valuable when identifying potential future outcomes.

Of course, Delphi administrators should avoid providing any frame, anchor, or desirability impulse via the Delphi itself because such an input would be received by all panelists irrespective of their diversity. That means heterogeneity predominantly works against pre-existing frames, anchors, and desirability perspectives but not against Delphi-induced ones. In order to avoid Delphi-induced frames or anchors, we recommend presenting Delphi questions or projections in an unrelated order, i.e. each question or projection should be as contextually detached from its preceding one as possible.

In stage 4 of the standard Delphi process (feedback and revised estimates) the divergent perspectives provided by a heterogeneous panel create fruitful disagreement, evoke controversial thinking and promote judgmental accuracy (Yaniv 2011).

Warning participants of the existence and effects of biases and asking them for counter-arguments also has the potential to reduce the impact of each bias. However, it needs to be noted that, in order to take proper effect, each bias would require a different bias-specific warning, respectively (Furnham & Boo 2011). Yet, confronting participants with several individual warnings is supposed to be rather annoying for them. Delphi administrators are recommended to ensure proper understanding of Delphi procedures and purposes at the side of the participants. Furthermore, they could focus on one or two biases that are particularly likely to occur for the specific judgmental task of their Delphi, and limit warning to these biases. Contrastingly, asking participants to think about and consider counter-arguments can be done in a general manner that works for all biases alike.

Participants’ personal traits, i.e. high expertise, high cognitive abilities, and high involvement, take effect with regard to the mitigation of framing and anchoring, the desirability bias, and the bandwagon effect but may have controversial effects on belief perseverance. While these traits foster the provision of high-quality arguments that serve against belief perseverance of other participants, highly experienced panelists – being aware of their strong knowledge base – might be particularly reluctant to change their mind and preserve their beliefs instead. Post-hoc procedures only take effect for the specific bias they were designed for. However, they should not negatively impact the other biases as they are applied after the completion of the Delphi.

Finally, feedback can naturally not take effect with regard to framing and anchoring or the desirability bias as those operate prior to the provision of feedback. However, feedback is of
enormous relevance to the mitigation of the bandwagon effect and belief-perseverance (Wright & Rowe, 2011). It confronts participants with “a new anchor, a different frame, or a piece of disconfirming information, [and] can trigger beneficial thought processes” (Yaniv & Milyavsky 2007: 119). As argued above, we recommend providing only argumentative feedback, i.e. no statistics that could pull panelists to the majority. Further, double entries of arguments should be deleted, also to avoid bandwagon-specific majority effects, and non-causal arguments should not be provided in order to enhance feedback quality. This way, participants can focus on the content of high-quality arguments and thoroughly consider a proper opinion change instead of following the crowd or just sticking to their initial estimate.

We might also note that the addressed cognitive biases do not take effect in isolation but may occur in parallel as well as in sequence and impinge on each other. There are several kinds of potential co-effects. First, participants’ initial estimates (Delphi stage 3) may be simultaneously influenced by anchoring and framing as well as desirability bias. Second, it may of course be that initial estimates are biased by framing and anchoring or desirability and participants subsequently also fall victim to either the bandwagon effect or belief-perseverance. Third, biases prevalent in the initial estimates (Delphi stage 3) may even be polarized in subsequent feedback and revision rounds (Delphi stage 4) by multiple biases interfering with one another. For instance, a Delphi panel comprising a majority of like-minded experts sharing a certain desirability perspective will probably provide initial estimates being desirability-biased. The consequently biased feedback could make desirability contagious (Ecken et al. 2011) as other experts converge towards this biased feedback value (Kerr & Tindale 2011; Yaniv 2011), e.g. because the bandwagon effect makes them move away from their un-biased minority opinion (while the belief-perseverance bias keeps the majority participants at their erroneous position).

CONCLUSION

Research provides fairly diverging evaluations of the accuracy of Delphi studies in general and has identified contradictory results concerning the effects of certain design features (Rowe & Wright 1996; Rowe et al. 2005). These inconclusive results supposedly mainly stem from the discretion in design choices researchers have when planning and conducting Delphi studies. While being difficult when striving for a general accuracy assessment of Delphi studies, its flexibility is also one of Delphi’s major advantages. We tried to take advantage of design flexibility of Delphi studies by proposing several features whose advantages and drawbacks in countering cognitive biases we discussed in this study. We structured our analysis along the typical process steps of a Delphi study, and indicated which biases may occur at which stage of the process. We addressed the two main cognitive biases impacting initial Delphi estimates (stage 3); i.e. framing and anchoring as well as the desirability bias, and the two main biases taking effect during feedback and revision activities (stage 4); i.e. the bandwagon effect and belief-perseverance. We explained the mode of operation of each effect as well as underlying mechanisms leading to the respective psychological and cognitive phenomena. For each bias we also discussed several design features that may serve as remedies against unfavorable effects. We finally recommended a set of design features that partially mitigate the effects of several biases in parallel. As there is not much literature focusing on the enhancement of the accuracy of future-oriented Delphi studies through the avoidance of cognitive biases, we tried to contribute to this research stream by applying a strong cognitive bias perspective on Delphi processes and design features. This perspective leaves still much room for further research such as empirical investigations of the new design features recommended in this study,
analyses of the interdependencies between different cognitive biases being at play in parallel, or considerations of other distinct biases that might be relevant in specific research contexts. …
References

Bolger, Fergus / Stranieri, Andrew / Wright, George / Yearwood, John (2011): Does the Delphi process lead to increased accuracy in group-based judgmental forecasts or does it simply induce consensus amongst judgmental forecasters?, in: Technological Forecasting & Social Change, Vol. 78, No. 9, pp. 1671-1680.


