Judgment and Decision Making (2023), **Vol. 18**:e25 1–31
doi:10.1017/jdm.2023.22

EMPIRICAL ARTICLE

Virtuous opinion change in structured groups

Fergus Bolger[1], Gene Rowe[2], Ian Hamlin[3], Ian Belton[3], Megan Crawford[3], Aileen Sissons[3], Courtney Taylor Browne Lūka[3], Alexandrina Vasilichi[3] and George Wright[3]

[1]Department of Management, Anglia Ruskin University, Cambridge, UK; [2]Gene Rowe Evaluations, Norwich, UK and [3]Management Science, University of Strathclyde, Glasgow, UK

Corresponding author: Fergus Bolger; Email: fbolger42@gmail.com

Received: 15 June 2023; **Accepted:** 20 June 2023

Keywords: group judgment; judge-adviser systems; Delphi; judgmental forecasting; crowdsourcing

Abstract

Although the individual has been the focus of most research into judgment and decision-making (JDM), important decisions in the real world are often made collectively rather than individually, a tendency that has increased in recent times with the opportunities for easy information exchange through the Internet. From this perspective, JDM research that factors in this social context has increased generalizability and mundane realism relative to that which ignores it. We delineate a problem-space for research within which we locate protocols that are used to study or support collective JDM, identify a common research question posed by all of these protocols—'*What are the factors leading to opinion change for the better ('virtuous opinion change') in individual JDM agents?*'—and propose a modeling approach and research paradigm using structured groups (i.e., groups with some constraints on their interaction), for answering this question. This paradigm, based on that used in studies of judge-adviser systems, avoids the need for real interacting groups and their attendant logistical problems, lack of power, and poor experimental control. We report an experiment using our paradigm on the effects of group size and opinion diversity on judgmental forecasting performance to illustrate our approach. The study found a U-shaped effect of group size on the probability of opinion change, but no effect on the amount of virtuous opinion change. Implications of our approach for development of more externally valid empirical studies and theories of JDM, and for the design of structured-group techniques to support collective JDM, are discussed.

1. Introduction

The individual has been the focus of the majority of research into judgment and decision-making (JDM); however, important decisions are often made collectively rather than individually, for example, on the basis of expert advice, in a committee or a working group of peers, or through public consultation. This tendency has increased in recent times with the opportunities for easy information exchange offered by the Internet in the form of wikis, discussion fora, social media platforms, and so forth. The desire to seek out the opinions of others reflects the social nature of humankind—and it can be argued that most decisions are made in a social context—in other words, they are influenced by, and affect, others in one way or another. We argue that JDM research that considers this social context has greater generalizability and mundane realism than research that ignores this context.

Although people consult with other people when making decisions for social reasons, such as building support for a position (Vroom & Yetton, 1973) or reinforcing group identity and motivation

(Weber & Hertel, 2007; Worchel et al., 1998), a major reason for collective JDM is to improve the quality of the decisions reached. This is based on the assumption that 'n + 1 heads are better than one'—in other words, decision quality will be enhanced relative to individuals through an increased knowledge base and different perspectives brought to bear (e.g., Davis, 1969; Kerr & Tindale, 2011; Larson, 2010; Vinokur & Burnstein, 1974). Indeed, increased accuracy of collective decision making relative to that by individuals has been demonstrated[1]. An example is the 'wisdom of crowds' (Surowiecki, 2005), where large numbers of people independently make judgments that, when aggregated, modally produce more accurate results than the estimates of the majority of the constituent crowd—this is because averaging balances out random errors in individual judgments (Galton, 1907; Herzog & Hertwig, 2009). Advantages of collective over individual deliberation beyond simple aggregation effects have also been found for groups where some measure of interaction—and thus information exchange—between group members is permitted (see, e.g., Kerr & Tindale, 2004 and Tindale & Winget, 2019, for reviews).

However, research has shown that benefits of collective JDM are not always realized, or realized as fully as they might be (Janis, 1982; Kerr et al., 1996; Simmons et al., 2011). This is largely attributable to cognitive and social psychological factors. For example, with more sources of information to consider, cognitive load may increase; thus, we may become less and less able to extract signal from noise, or simply give up trying to make reasoned decisions altogether (for reviews, see Eppler & Mengis, 2004; Hwang & Lin, 1999)—although this load could potentially be shared in groups freely interacting in real time, it is not always the case, for instance, where there is asynchronous interaction. Further, a commonly documented response by individuals to high information load is to use cognitive heuristics to reduce that load (e.g., Helgeson & Ursic, 1993; Johnson & Payne, 1985; Swain & Haka, 2000), potentially resulting in more biased judgments and decisions at an individual level, as has commonly been shown in the JDM literature. Aggregation over a set of similarly biased judgments will reinforce that bias at a group level; for example, the mean of individual judgments that are modally too high will also tend to be too high (while the apparent consensus may lead to increased confidence in the biased group judgment). Not only might cognitive load be increased by the additional opinions and reasoning to consider in group settings, but it also might be increased from the social situation itself (e.g., presentation management, evaluating others' motivations, assessing truthfulness, etc.).

Social biases can also reduce the benefits of accessing more—and more diverse—information provided by group interaction (see, e.g., Steiner, 1972 for an analysis of the effects of group processes on their productivity). Perhaps the best-known example of this is 'groupthink' (Janis, 1971; Whyte, 1952) where groups pay more attention to ensuring group cohesion than they do to accuracy, leading *inter alia* to overconfident group decisions with potentially disastrous consequences (e.g., the Bay of Pigs fiasco and the Challenger explosion, Nijstad, 2009). Further negative effects on group performance include: group members being swayed by those who are the most dogmatic (e.g., Devine et al., 2001) or dominant (Watson et al., 1998), rather than the most knowledgeable; some group members playing personal or political 'games' for their own gain rather than concentrating on the task at hand (e.g., Barge & Keyton, 1994); and individuals in groups failing to share their unique information, focusing instead on the information that is held in common (e.g., Stasser & Titus, 1985; Stasser & Vaughan, 2013). Additionally, people in groups may sometimes suffer social *inhibition* due, for example, to anxieties related to interacting with strangers (Buck et al., 1992), which reduces their motivation to perform. There can also be a diffusion of responsibility in groups, leading to reduced participation (aka 'social loafing'), with the probability of such behavior increasing with group size and degree of physical and temporal dispersion of group members (Chidambaram & Tung, 2005).

Because of the perceived benefits of collective JDM, but with the recognition of the potential negative effects just outlined, considerable efforts have been put into designing technologies that seek to

[1] We are deliberately avoiding the use of the word 'group' here as we are talking about JDM by more than one person regardless of how it is done (e.g., by interacting groups or through aggregation of individual judgments), the study of which transcends several literatures. We will discuss the distinctions between these different modes of collective JDM shortly.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

maximize the group 'process gains' while minimizing the 'process losses' (Steiner, 1972; Watson et al., 1998) including the Delphi technique and prediction markets (PMs) that we will describe shortly. Delphi and PMs can be referred to collectively as 'structured-group techniques (SGTs)' to distinguish them from unstructured, freely interacting groups where there is no attempt to control the group processes. It is these structured groups that are the primary focus of this paper.

It should now be clear that the study of collective JDM can bring practical benefits (e.g., improved decision-aiding technologies that better balance the costs and benefits of collective efforts). However, we will now argue that research in this area has theoretical benefits that cannot be attained from focusing solely on individual JDM.

Collective JDM research has largely been equated with the study of 'group decision-making'. In a recent review of this field, Tindale and Winget (2019) situate protocols[2] used in such studies along two dimensions: How much interaction or information exchange is allowed among the group members (i.e., how much individual JDM is affected by information from others); and how much influence group members have over the final decision (e.g., how directly individuals contribute to consequent policymaking).

Freely interacting groups that make consensus decisions are high on both dimensions, while opinion polls used for decision-making by others are low on both dimensions. Judge-adviser systems (JASs: see Bonaccio & Dalal, 2006 for a review), where a decision-maker consults advisers, are high in influence over decisions, but have limited information exchange, since information flows only from adviser to judge. Focus groups are high in interaction, but, like opinion polls, how the results are used is usually outside the control of group members. Meanwhile, Delphi (Dalkey & Helmer, 1963; Linstone & Turoff, 1975)—initially designed for long-term forecasting, but widely applied to collective JDM subsequently—and PMs—an application of the aforementioned wisdom of crowds, principally for short-to-medium term forecasting of economic and political events—fall somewhere toward the middle of this space since interaction between group members is restricted, while decisions are generally based on mathematical aggregation (but exact procedures, and hence position in the space, vary). These last two approaches require a little elaboration.

In the typical application of Delphi, anonymous panelists are asked to provide a judgment (or forecast, choice, solution, etc.) on the issue at hand in a first *round*. These individual judgments are then collated by a facilitator, aggregated, and fed back in a second round to the panelists of the nominal group ('nominal' because the group does not usually convene at a particular place or point in time). Panelists then consider the information (which may simply be the mean or median of the group response where quantitative values are called for, but may also include confidence and/or justifications of responses) and then provide another response. Several further iterations may take place until there is little or no change in responses. After the final round, the facilitator aggregates quantitative responses of individual panelists (or collates and/or summarizes qualitative ones), to form a single group response. Aggregated quantitative responses are usually equally weighted in Delphi, but weighting by measured expert performance is an option (see, e.g., Czaplicka-Kolarz et al., 2009; Hanea et al., 2016 for examples of performance weighting).

Interest in PMs is growing (see, e.g., Healy et al., 2010; Wolfers & Zitzewitz, 2004). PMs actually have the *most* restricted information exchange of any of the protocols discussed here: Typically, participants can only see the current market price—and the offered buying and selling prices. They then indicate their opinion of those prices by the offers they make, and the deals they strike.[3] Precisely how a PM is implemented can vary substantially, but the general principle is that each individual participant buys a prediction of an event at a price that reflects the current collective thinking of the probability of that event occurring. For example, if there are two possible outcomes 'A' and 'B', and outcome 'A' is

[2]We use the term 'protocol' to refer to a set of rules or principles for structuring the interaction between panelists—this contrasts with the specific procedures or methods that are used, which can vary between different instantiations of the protocol.

[3]PredictIt, the largest PM platform, also allows for comments, but the value of these for accurate pricing is questionable—PMs are competitive rather than cooperative, so disinformation is as likely as information.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

seen by the majority as being the more likely of the two, then outcome 'A' will command a higher price than outcome 'B'. When the event—say an election—is resolved, the price of the winning outcome will be an amount corresponding to a probability of 1 (i.e., the maximum possible price) and the losing event(s) a probability of 0. So if, for instance, you believe that there is a 70% chance of candidate A beating candidate B, you should be prepared to pay up to \$0.70 for the chance to win \$1 (i.e., the expected value of the bet is \$1 x p(.7) = 70 cents).

These different approaches to collective JDM have largely been studied without reference to each other, with different methods used, and different theories developed. These theories, and supporting evidence, are published in different journals, despite their common basic research question: *What are the factors leading to improved (or diminished) JDM performance of groups relative to individuals?* Of course, the study of the social psychological processes involved (and their interaction with properties of the method, such as type of aggregation allowed) is essential to answering this question. The answer can contribute to our understanding of the psychology of JDM as well as the construction of better technologies for collective JDM (i.e., protocols and supporting infrastructure such as IT systems, and instruments for selection and training of experts).

In this paper, we propose a research paradigm that can assist in the search for answers to the research question above across all the different approaches to group JDM identified above. This is a paradigm in Kuhn's (1962) specific sense of a theoretical or conceptual framework that guides hypothesis generation, and construction of experiments to test these hypotheses. The paradigm thus involves a general methodology within which are a variety of specific procedures that are informed by our theoretical perspective, which rests on a reframing of the general research question above from *What are the factors leading to improved (or diminished) JDM performance of groups relative to individuals?* to *What are the factors leading to opinion change for the better—or 'virtuous opinion change'—in individual agents?*

Why do we make this change? To have any improvement in JDM performance due to some intervention, an agent must experience a change of opinion from one state to a better one, however defined. This new framing allows operationalization of our basic research protocol which is that:

1. An agent expresses their initial opinion (e.g., makes a forecast or judges the value of an uncertain quantity)
2. The agent receives some information that might be pertinent to the expressed opinion (e.g., a forecast generated from a statistical system or someone else's opinion on the value of the quantity)
3. After considering this information, the agent is given the opportunity to revise their original opinion
4. Further iterations of opinion revision on receipt of new information can occur
5. The final opinion is passed on to the decision-maker (perhaps aggregated with those of other agents)

We deliberately keep our terminology general so as not to prejudice the potential application of this protocol; thus, the nature of the *intervention* is open—for example, it can be discussion by groups of various compositions, or receipt of information from the Internet or from a statistical system—so long as information is exchanged. Further, *opinions* can be judgments, beliefs, arguments, statements of fact, etc. and either quantitative or qualitative in nature[4], while *agents* must simply have expressible opinions (as defined above) that can potentially be revised on the receipt of information. Normally agents will be individual people, but they could be groups of people (e.g., think tanks or companies), or a 'chatbot', or a hybrid of people and machines.

Our research protocol has three inspirations[5]: Asch's (1951) conformity experiments, Delphi, and JAS. JAS is the simplest example of this protocol—there is just one judge who makes a judgment,

[4]However, in this paper we are primarily interested in assessing the veridicality of opinion change, so we focus on quantitative or categorical judgments that can be potentially validated.

[5]Some readers may see similarities between our work and Hammond et al.'s (1975) social judgment theory (SJT), particularly the parts of SJT relating to coordination between multiple judges. These were not explicit inspirations for our work; however, examination of similarities and differences between the approaches is mutually informative, so will be carried out in the Discussion.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

then receives some advice, usually in written form, from one (or more) adviser(s), and then is given the opportunity to 'revise and resubmit' their judgment. Delphi is very similar in design to JASs, but typically uses larger numbers of anonymous agents who are both judge *and* adviser (e.g., peers who may hold different opinions)—the focus of research has been on whether final group judgments are more accurate than individuals or standard committees. Meanwhile, social psychological research has looked at how people in interacting groups might be influenced by other members. Asch's paradigm has multiple agents making a judgment of a given stimulus in turn, where all but one of the agents are confederates of the experimenter and make an obviously incorrect judgment—of interest is to see if the participant changes his or her opinion to that of the group. These three applications of the protocol demonstrate its potential to explore the entire problem-space delineated by Tindale and Winget (2019) that we described above.

Taking our lead from 'JASs' and 'Delphi inquiry systems' (Parenté & Anderson-Parenté, 1987), for ease of reference, we wish to characterize all applications of the protocol outlined above as 'multi-agent revise and resubmit systems' (MARRSs for short). To reiterate, the basic questions being asked by researchers in the three MARRS examples above (and in PMs) are the same—What influences opinion change? And under what conditions does it change for the better (i.e., virtuous opinion change or VOC)? So, in JASs, for instance, questions include: How does perceived quality of advice affect the way judges use it? What is the optimal response in terms of VOC? In Delphi we might ask: Does feeding back confidence or reasons or both have the biggest effect on VOC? While in Asch's more freely interacting groups it was asked: How does the ratio of false to true judgments or the ambiguity of the stimulus affect VOC (opinion change *away* from the truth in this case)?

So, what are the factors potentially influencing VOC across MARRS? From reading and brain-storming we identify three basic types of factors that may influence the probability of a group member changing his or her opinion—characteristics of the individual (I), the situation (S), and the method (M)—see Figure 1.

I characteristics of the individual include: the ability to process arguments and evidence (that might lead to opinions being reinforced or changed); personality traits such as openness to alternative viewpoints; knowledge (normative versus substantive expertise) and the degree of variation in that knowledge between experts; the strength with which knowledge is held (e.g., confidence); and social factors such as social skills (e.g., persuasion) and status (e.g., giving authority to position).

S characteristics include: how much time there is available to consider evidence[6]; the complexity of responses (e.g., the number and variety of different opinions that could be expressed, either received or sent); the degree of consensus (or, alternatively, the relative balance of opposing opinions, e.g., Is there a large majority?); the existence, nature, and size of any incentives for good judgment; and task factors such as the problem difficulty[7], the type (e.g., describing the current situation or forecasting the future), and the potential validity of any answers provided by panelists (e.g., Is the answer known or knowable in principle?)

M Characteristics include: the size of the group; the social processes permitted (e.g., Is the interaction face-to-face or mediated by a facilitator?); the style of any facilitation (e.g., light touch or heavy?); the amount and type of feedback (e.g., quantitative, qualitative, or mixed?); the opportunities to revise opinion (e.g., number of Delphi rounds); and the instructions given (e.g., members may be asked to think why they might be wrong).

Many of the factors may interact with each other to influence opinion change—both within and between types. For example, time available may well interact with the number of opinions expressed (both situation factors, which we shall denote 'S'), so that the potential for new evidence to produce opinion change will only be realized if there is sufficient time for that evidence to be given full consideration.

[6]Time available could also be a deliberate feature of the method—and thus an M as well as an S factor—but this is rare in practice.

[7]It is probable that the nature of the task being performed—such as its difficulty—impacts on how panelists respond within a Delphi environment (e.g., Hackman & Morris, 1975; McGrath, 1984; Rowe & Wright, 1996; Straus, 1999).

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

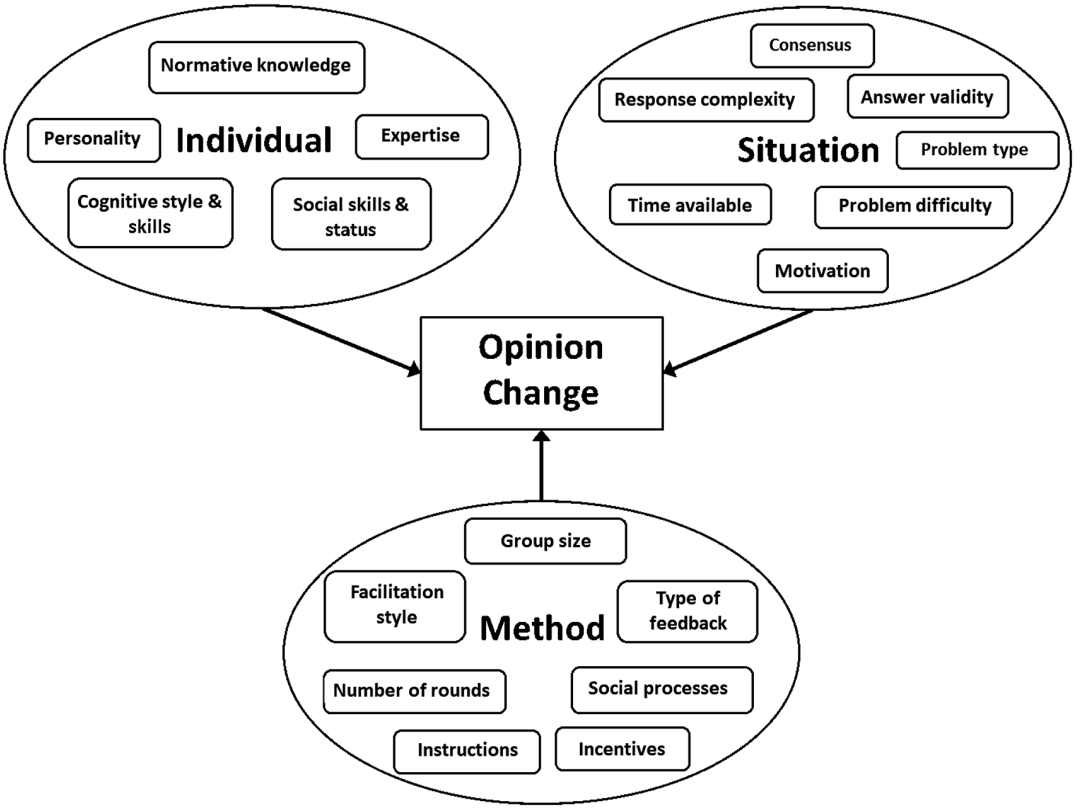

Figure 1. *Some factors potentially influencing opinion change of individuals in a MARRS.*

So, following an approach used by Bolger et al. (2011) to model opinion change in forecasters of Australian rules football (see also Bolger & Wright, 2011 for the theoretical context for this approach), if we assume that only properties of the individual, situation, and method—and/or their interactions—influence opinion change, then for each in the group, the probability of him or her changing their opinion is given by:

$$\mathrm{p\,(OPC)} = \mathrm{I} + \mathrm{S} + \mathrm{M} + \mathrm{I.S} + \mathrm{I.M} + \mathrm{S.M} + \mathrm{I.S.M} + \mathrm{error}. \quad (1)$$

This can be operationalized by a logistic regression if opinion change (OPC) is coded as a binary variable where no opinion change (between two points in time separated by a MARRS) is coded by a 0 and opinion change is coded by a 1.

But we are also interested in determining the conditions where this individual opinion change is virtuous (i.e., leads to greater accuracy)[8]. This requires a slightly different modeling approach. For binary judgments (e.g., happens or does not happen; true–false, etc.), accuracy change could be coded as 0 for no accuracy change, +1 for positive change (the judgment improves, i.e., VOC), and − 1 for negative change (the judgment gets worse). Modeling accuracy change with a (now multinomial) logistic regression equivalent to [(1)] is then possible, but its interpretation is problematic. This is because accuracy change is not just affected by the I.S.M. factors on the right-hand side, but is also largely determined by initial accuracy; if you were initially correct, then you can only get worse (if you

[8]Although operationally we define VOC as opinion change that leads to greater accuracy, we derived the term 'virtuous' from Dalkey's (1969) proto theory of opinion change in Delphi, where opinion may be 'pulled' toward the 'truth' (or toward other things).

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

change your opinion) or stay the same (if you do not)—VOC is therefore restricted to those who were initially incorrect[9]. For this reason, we propose that accuracy change is summed over several questions to create a net accuracy change (NACH) score:

$$\mathrm{NACH} = \sum_{k=1}^{N} (j_k - t_k)\,, \tag{2}$$

where j_k is an instance k of a set of judgments, t_k is the true value for the corresponding instance, and N is the total number of judgment instances in the set under consideration.

Of course, this does not in itself necessarily solve the problem of accuracy change being dependent on initial accuracy. For example, if your set of questions is extremely easy, then NACH may be minimal or non-existent. It is therefore desirable, for modeling purposes, to use as large a set of questions as is practical, and which are varied in difficulty. These problems are reduced if judgments of continuous quantities are elicited as, unless an estimate is exactly correct initially—a rare event in most contexts—there should be scope for improvement, but we still recommend aggregating over as many judgments of varied targets for each participant as possible to maximize internal and external validity of the conclusions. The second question, regarding the determinants and direction of accuracy change, can be answered with an ordinary least squares regression, thus:

$$\mathrm{NACH} = \mathrm{I} + \mathrm{S} + \mathrm{M} + \mathrm{I.S} + \mathrm{I.M} + \mathrm{S.M} + \mathrm{I.S.M} + \mathrm{error}. \tag{3}$$

Given our averred intention to study factors leading to VOC, this equation begs the question of the necessity of [EQ (1)]—why bother to model opinion change when [EQ (3)] subsumes it? The answer is that a study might uncover important information relating to influences on opinion change within a MARRS, yet fail to find any determinants of accuracy change, simply due to the difficulty of the judgment task and/or the knowledgeability of the experts used. Finding opinion change without significant accuracy improvement might suggest the technique could have value with a different task–expert mix, whereas failure to find opinion change at all implies little potential for the method.[10]

2. The simulated group response paradigm

We now propose an efficient way of studying collective JDM in MARRSs which will permit us to systematically investigate VOC in terms of the three types of factors identified in Figure 1, and their interactions. This research paradigm—called the 'simulated group response paradigm' (SGRP)—is primarily based on methods used to study JASs and, as such, facilitates the extension of the high level of experimental control manifest in JAS studies to the study of other sorts of MARRSs typically studied using traditional interacting groups.

We argue that attempts to answer questions regarding, for example, the effectiveness of Delphi have often been poorly executed (or not executed at all) because of logistic problems in recruiting interacting groups in sufficient numbers to attain reasonable power. The SGRP avoids these problems by allowing group designs to be run while retaining VOC of individual agents as the primary unit of analysis. As well as helping determine better SGTs, this can also benefit research in JASs, because including more complex designs (e.g., with multiple interacting judges) will increase the external validity of future studies in this area.

The SGRP consists of two main phases which are summarized in Figure 2. In phase 1, data are elicited (or created) for use as stimuli in phase 2—this is referred to as the 'stimulus elicitation exercise' (SEE). In phase 2, hypotheses about influences on VOC are tested in a Delphi procedure with—in

[9]However, opinion change might also be constrained by initial accuracy, for instance, if accuracy and confidence are, in fact, correlated as proposed in Dalkey's (1975) account of how opinion may change in Delphi (see also Parenté & Anderson-Parenté, 1987)

[10]Factors affecting NACH might not be the same as those affecting p(OPC). For instance, if accuracy and confidence are poorly correlated, then confidence could be an important factor in opinion change, but has little influence on NACH.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

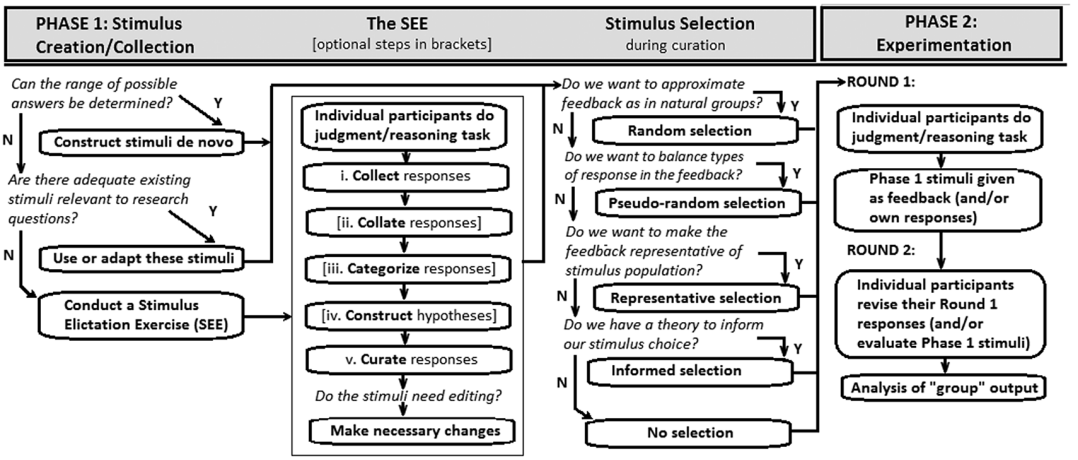

Figure 2. *The simulated group response paradigm (SGRP).*

the basic form—two rounds. In round 1 participants individually provide answers to a judgment or reasoning task. At round 2 the participants then receive a number of other answers to the same task and are invited to revise their round 1 answer. Thus, phase 2 of the SGRP is, from the point of view of a participant, identical to a standard two-round Delphi. There are, however, important differences from the perspective of the experimenter. The most significant difference is that the answers fed-back at round 2 do not derive from other group members working on the task more or less synchronously—as is usually the case—but are elicited, or created, at some earlier point in time during phase 1. A detailed exposition and discussion of the SGRP can be found in Appendix 1.

3. Experiment using the SGRP: group size and diversity

To give a specific example of how the SGRP can allow questions about effects on VOC to be answered efficiently, let us consider how we might examine the influence of group size on amount and quality of opinion change. Determining the number of participants per group is an important decision faced by any researcher wishing to use a Delphi—or indeed any group process, structured or otherwise. On the one hand, larger panels are more likely than smaller to contain the expertise required for good judgment outcomes (e.g., more accurate forecasts). On the other hand, expertise is often scarce, so recruitment of large Delphi panels comprised of individuals with appropriate knowledge may be difficult (and expensive), while large groups are harder to manage than small.

Several size ranges been suggested, for example 5–20 (Rowe & Wright, 2001), 15–60 (Hasson et al., 2000), no more than 50 (Toma & Picioreanu, 2016), or 15–30 for homogenous panels (Clayton, 1997) and 5–10 for heterogeneous panels (Delbecq et al., 1975). However, the theoretical basis for these proposals is unclear and empirical support mostly lacking. In practice, Delphi studies are run with a very wide variety of group sizes, from smaller than 10 (e.g., Stebler et al., 2015) to larger than 100 (e.g., Bisson et al., 2010), with variation driven more by the exigencies of the tasks than any principled considerations. There is probably no universal optimum group size for Delphi; rather, different combinations of tasks, available expertise, and, indeed, precise nature of the method used (e.g., type of feedback, number of iterations, etc.) will likely determine what works best in any given case.

Larger Delphi groups are likely to produce a wider range of opinions and perspectives, which may in turn generate more accurate and plausible group judgments (e.g., Hussler et al., 2011). For this to happen, we propose that VOC must occur by those with less accurate initial judgments moving toward those with more accurate initial judgments, more than vice versa, which in turn will be facilitated

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

by asking participants to provide rationales in support of their judgments (Bolger & Wright, 2011). However, where written rationales are involved, increasing the size of a Delphi group means increasing the amount of text[11] for panelists to review in each Delphi round. At some point, this may lead to information overload and consequently reduce the effectiveness of group performance.

Diversity of opinion may have a similar effect on VOC as group size—increasing VOC up to a point, but leading to overload, and thus a decrease in performance, thereafter. Further, diversity and group size are likely to interact, with a more rapid increase in information load for the group placed in the high-diversity situation than in the low. So, we hypothesize that there is a trade-off between the advantages of adding knowledge and perspectives by increasing group size and diversity of opinion, and the disadvantages of increased information load for panelists and facilitators.

To compare just two group sizes—say 5 and 9—in a traditional Delphi setup, one would need to run 60 groups per condition (for power of .8 and a moderate effect size of .35), meaning finding 840 ((60 x 5) + (60 x 9)) participants! This is probably why such a study has never been done, to our knowledge at least. In the SGRP one needs 120 participants (2 groups of 60) for the main study, plus possibly another group for the SEE—the exact number depending on the task, but is unlikely to be more than a further 60—to answer the same question, a huge improvement over the traditional method (note that increasing nominal group size does not affect the number of participants needed in the main study, but may impact on the number needed for the SEE). An added advantage of the SGRP over traditional groups is that, since participants are working individually, the experimenters do not have to coordinate panelists, which can be a substantial drain on resources.

The only two studies to date that have directly examined Delphi group size found no consistent relationship between group size and effectiveness. Boje and Murnighan (1982) compared the performance of Delphi groups of 3, 7, and 11 across two statistical and two almanac problems with answers uncertain or unknown to participants, respectively. Estimate accuracy slightly *decreased* with group size, but confidence increased, thereby inflating overconfidence, a commonly observed bias (Lichtenstein et al., 1982). However, the *group*-level analysis used here was seriously lacking in power (a maximum of seven observations per condition) and would likely have been unable to detect even large effects of group size on accuracy. Brockhoff (1975) compared groups of 5, 7, 9, and 11 panelists in a five-round Delphi. Group performance was measured using the median group error on sets of 'fact-finding' and forecasting questions. The study found no systematic effects of group size, with the ordering of groups by accuracy varying as a function of problem type and round assessed. Statistical power was even lower than in Boje and Murnighan (1982), as only one group of each size was tested. In both studies many procedural details are vague—particularly as regards feedback given to panelists—and we suspect that there was poor experimental control. The value of the findings of these two studies is consequently somewhat limited.

We conducted an experiment to investigate the effects of varying group size and levels of opinion diversity on the accuracy of short-term probabilistic forecasts.

Our specific hypotheses were as follows (which are all confirmatory):

H1: Increasing group size will lead to an increase in the probability of opinion change. It is not predicted whether this relationship will be linear or quadratic in nature.

H2: Increasing opinion diversity will lead to an increase in the probability of opinion change. It is not predicted whether this relationship will be linear or quadratic in nature.

H3: There will be an interaction between group size and opinion diversity, with the effect of group size increasing more rapidly for high-diversity groups.

H4: Increasing group size will lead to an increase in forecast accuracy up to a certain (as yet unknown) group size but a decrease in accuracy thereafter.

H5: Increasing group diversity will lead to an increase in forecast accuracy.

[11] In most Delphi applications, rationales are written, which helps to maintain anonymity.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

4. Method

4.1. Participants

A total of 282 participants were recruited online using the 'Prolific' recruitment platform. Thirty participants took part in the pre-study SEE—this provided sufficient data to prepare the stimuli for the main study—and 252 participants completed the main study.

4.2. Design

4.2.1. Pre-study (SEE)

No variables were manipulated in the pre-study. The procedure used is reported below.

4.2.2. Main study

The main study was a 5 x 2 between-subjects design. Two independent variables were manipulated in the study: Delphi group size (5 levels: 7, 10, 13, 16, and 19); and opinion diversity (2 levels: low and high, described below).

4.3. Stimuli and procedure

4.3.1. SEE

Here we collected data for the construction of simulated groups of answers and rationales that were used in the main study. This procedure was in line with the SGRP described above. Stimuli were presented online using Qualtrics survey software. Participants were given 10 short-term forecasting questions (see Appendix 2). For each question, participants were asked to answer 'yes' or 'no' and give three rationales to support their answer.

4.3.2. Main study

Participants were randomly allocated to one of 10 conditions. Stimuli were presented online using Qualtrics. The order in which participants received the 10 forecasting questions was also randomized to prevent order effects. In each condition, participants were given 10 short-term forecasting questions. The following process was repeated for each question:

1. Participants were asked to answer 'yes' or 'no', give a rationale for their answer, and provide a confidence estimate on a scale from 50 to 100 percent (an estimate of less than 50 percent would mean that the participant should have chosen the other answer).
2. Participants were then shown a set of 6, 9, 12, 15, or 18 answers taken from those collected during the pre-study, each with an accompanying rationale. In the low-diversity conditions, each 'yes' and each 'no' answer was supported by a version of the same rationale. In the high-diversity conditions, each 'yes' and each 'no' answer was supported by three different rationales (repeated multiple times in the 12, 15, and 18 answer conditions). In every case, participants were shown a majority of contrary answers and rationales (for 'no' if they answered 'yes' and vice versa), in the ratio 2:1 (4:2; 6:3; 8:4; 10:5; 12:6).
3. Next, participants were shown their original answer and rationale and given an opportunity to revise them, if they wished, after reading the other responses. Participants were also able to revise their confidence estimate at this point.
4. Lastly, participants completed an individual difference questionnaire designed to measure 'actively open-minded thinking' (AOT—Baron et al., 2015) and provided basic demographic information.

Two separate attention checks were included in the stimuli to prevent completion of the online survey by 'bots'. IP addresses of all completed responses were also checked to confirm that there were no duplicates.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

4.4. Model to be tested

In line with our approach described above, we are trying to model influences on opinion change, with the primary influences under consideration being group size (a methodological variable) and diversity (a situational variable[12]). We also considered two individual-difference variables that might affect VOC. First, we hypothesize that those indicating lower initial confidence in their forecasts should be both less expert and more likely to change opinion than those higher in confidence. Further, those more receptive to evidence should be more likely to change their opinions for the better. We therefore had two more hypotheses:

H6: Increased confidence at round one will lead to a reduction in the probability of opinion change.

H7: Participants high in AOT will be more likely to change opinion than those low in AOT.

Although these individual-difference variables may interact with each other and/or group size and/or diversity, we have no particular hypotheses about this, so, for the sake of simplicity, we will not include such interaction terms in this model. The model to be tested is then:

$$\mathrm{OPC} = \mathrm{I} + \mathrm{S} + \mathrm{M} + \mathrm{S.M} + \mathrm{error}, \tag{4}$$

where OPC is opinion change for each of the 10 forecasts made by each participant—this is a binary variable where 1 indicates a participant changed his or her forecast between round 1 and round 2, and 0 indicates no such change; I variables are AOT and round 1 confidence for each person and each forecast; S is a dummy for diversity (0 or 1 for low or high); and M is group size. Above we predicted a possible nonlinear relationship between group size and performance – this might also pertain for opinion change. Recall that the judge was always in a minority, so although the ratio of opposed to supporting opinions remained constant, the number of opinions against their position increased with group size. However, if the number of opinions to consider became too large, then their influence might have decreased. Even if this hypothesis is correct in general, as mentioned earlier, we do not know a *priori* if in this particular study it will be manifest because the inflection point may occur beyond the maximum group size tested. For this reason, we tested the shape of the relationship between group size and opinion change—a polynomial contrast revealed a reliable quadratic trend (mean difference = .034, 95% Cis [.007, .062])[13]. We therefore squared (and centered) group size before entering it into the regression.

Since the dependent variable—opinion change—is a binary variable, and the order of participants' forecasts was of no consequence since it was randomized, the appropriate analysis is a multilevel logistic regression, where question number is a level 2 identifier.

In MLM a decision needs to be made as to whether predictors are fixed or random variables. We have no particular expectation that slopes of our predictors (i.e., confidence at round 1, AOT, group size, diversity—and the interaction between group size and diversity, calculated by multiplying the two terms) will vary systematically across our heterogeneous set of ten questions (although the intercepts are likely to do so due, for example, to different levels of uncertainty across questions). For this reason, all of the predictors in our model have random intercepts and fixed slopes.

5. Results and discussion

The regression model shows that contrary to expectations (H1), probability of opinion change neither increased with group size, nor displayed an inverted U-shaped relationship. Rather there was a significant U-shaped function, with proportion of opinions changed *decreasing* from group size 7 to

[12] We distinguish methodological from situational factors in that the former are fixed by the experimenter, whereas the latter are free to vary—thus, diversity, an independent variable in our study, should be considered methodological rather than situational. However, in a natural setting diversity would not be strictly controllable by the researchers and so cannot properly be considered a methodological factor, although efforts could (perhaps should, e.g., Bolger, 2018) be made to increase diversity during the selection of experts or the procedure itself.

[13] In fact, the curve was U-shaped rather than the predicted inverted U-shape—we will return to this in the discussion of results.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

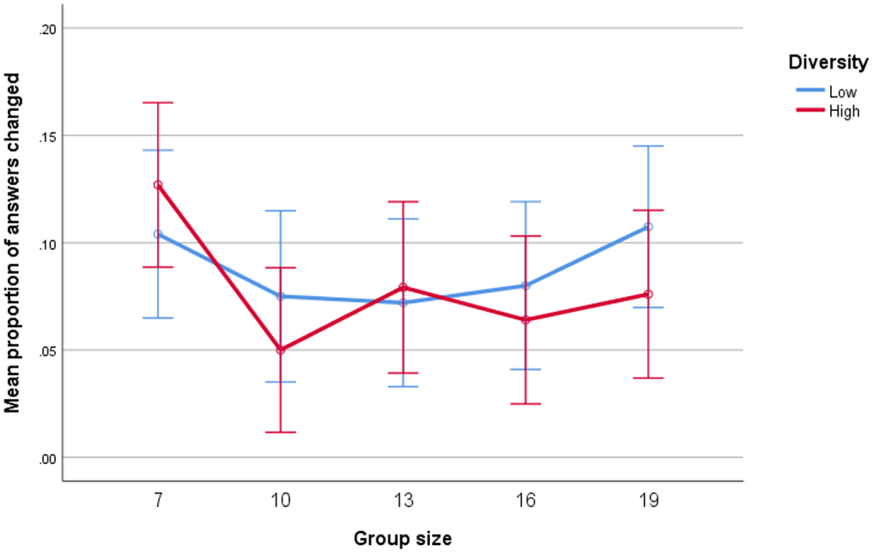

Figure 3. *Mean answer change proportion by group size and diversity. Error bars are 95% confidence intervals.*

Table 1. *Multilevel logistic regression coefficients and tests of significance.*

Model term	Coefficient	Standard error	*t*	Significance
Intercept	−3.575	0.787	−4.545	<.001***
Pre-confidence	−0.005	0.005	−1.051	.293
AOT	0.027	0.010	2.702	.007**
Diversity	−0.056	0.146	−0.383	.702
Group Size (GS)	−0.070	0.027	−2.544	.011*
GS × Diversity	0.039	0.017	2.274	.023*

*$p < .05$.
***$p < .01$.
***$p < .001$.

group size 10, then *increasing* again for groups of size 13 to 19—this curve was somewhat smoother and flatter for the low-diversity condition than that for the high-diversity condition (see Figure 3), which was manifest as a significant coefficient for the interaction between group size and diversity in the regression model, as predicted (H3) (see Table 1). Also contrary to expectations, diversity (H2) and confidence levels at round 1 (H6) did not influence opinion change. As predicted (H7), opinion change was significantly more likely for those scoring higher in AOT.

To examine the influences on VOC, we coded accuracy change as +1 if a forecast was changed from incorrect to correct, −1 if a forecast was changed from correct to incorrect, and 0 if there was no change (correctness was established by monitoring major news outlets during the forecast period). The NACH over the ten questions was calculated for each judge and was then regressed on two individual difference variables (mean confidence before receiving feedback about other 'group' member's forecasts, and AOT score), a situation variable (diversity), a method variable (group size), and the interaction of diversity and group size:

$$\mathrm{NACH} = \mathrm{I} + \mathrm{S} + \mathrm{M} + \mathrm{S.M} + \mathrm{error}. \quad (5)$$

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

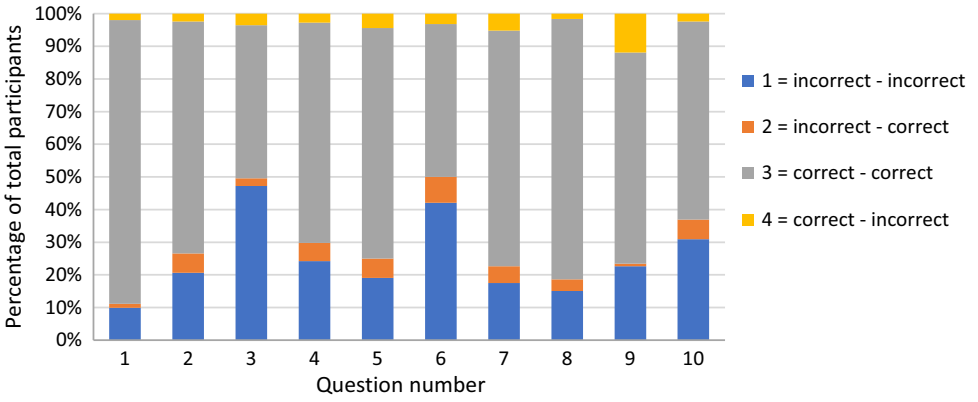

Figure 4. *Percentage of participants changing their answers between rounds, by question number.*

An MLM analysis for this model, however, produced no significant coefficients, contrary to predictions (H4 and H5).

First, we find no statistically significant effects of group size and diversity on VOC, and although the effects on opinion change are significant, they are small in practical terms—essentially people are rather unwilling to change their original answers, with only just over 1 answer in 10 being changed at most, despite being presented in every case with a majority of contrary opinions. This echoes the 'egocentric discounting' effect found in several JAS studies (see Bonaccio & Dalal, 2006 for a review), whereby judges tend to underweight advisers' opinions relative to their own (e.g., Harvey & Fischer, 1997; Yaniv & Kleinberger, 2000). DeKay (2015), p. 406) highlights 'the ubiquity of predecisional information distortion'. Here, a decision-maker's current information preferences determine his/her evaluations of the next information item. In this way, preference for a (currently) leading option is supported and that for a trailing option is degraded. Our findings are consistent with this phenomenon, since it suggests that our participants' assessments of the rationales presented to them may have been distorted to favor retaining their original judgments. Further, since the present task involved binary choices, there was no scope for participants to integrate their own and others' judgments, for example using weighting or averaging strategies that may be optimal in many cases where, as here, judges have limited ability to evaluate their expertise against their fellow group members (e.g., Bednarik & Schultze, 2015).

Features of our task could also have been limiting factors for opinion change. For example, the majority selected the correct answer to start with for eight of the ten questions (see Figure 4), while opinion was evenly split for the remaining two. In this situation, where most judges are initially right, both opinion change and thus VOC may be restricted as a result. However, this does not seem a likely explanation for the low opinion change and VOC in our case—with the exception of Q1, there were still 20%–50% incorrect initial answers (see Figure 4) so there was plenty of scope for change, yet it was not realized.

Overall, Figure 4 shows that the balance of change for the better was about the same as that for the worse, which accounts for the lack of net VOC found. For most questions VOC dominates, but opinion change for the worse substantially outweighed opinion change for the better for Q9—thus more persuasive arguments, and more opinion change, do not necessarily translate into more accurate outcomes. This may be particularly true of forecasting tasks such as the ones we used here where something may occur or not regardless of the best arguments and indicators available at the time of forecast[14]. In other tasks—such as estimating something's value on the basis of partial

[14]For this reason, in the Australian rules football forecasting study the expectations of expert sports pundits were used as the standard against which to compare our judges' forecasts rather than the true outcomes—potentially this could be done for the current geopolitical forecasting task, although it is less obvious who should be polled for the expert predictions.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

information—the amount of evidence and/or quality of arguments may be more directly related to the location of truth.

While changes in group size and diversity did have *some* effect on the amount of opinion change for this task, another feature clearly needs to be added to the method, and/or a different task used, if useful amounts of opinion change (and potentially VOC) are to be stimulated. Variables that have been either found to stimulate opinion change, or proposed to, are rationales, confidence, and majority positions. Taking these in reverse order, majority positions were controlled for in this study—there was always a 2 to 1 majority against a judge's initial position in order to encourage change—but perhaps an even larger majority might have encouraged greater change.

We proposed earlier that in order for VOC to occur, those lacking in confidence in their judgment relative to others must be more likely to change their opinion—a view that found support in the Australian rules football study described above (Bolger et al., 2011). In the current experiment, we found no evidence that those initially low in confidence in their judgment changed their opinion more; however, pre-decision confidence levels were fairly high (just over 77% on average) and we did not feed back confidence between judgment rounds, but it may be that expertise self-ratings for each judgment question or external expertise assessments are more valid indicators of true expertise than confidence. These ideas could be tested using the SGRP.

If the rationales given for forecasts had been more persuasive, more opinion change might have occurred—although in our view many of the rationales used in the study (see Appendix 3 for examples) were already quite persuasive, particularly compared to those sometimes observed in natural Delphi groups (see, e.g., Bolger et al., 2011). If persuasive rationales were also related to correct outcomes, then this would naturally translate into increased VOC. In this study we varied diversity of opinion (i.e., how many different arguments were made), rather than persuasiveness, but it would also be feasible to manipulate persuasiveness within the SGRP.

We also wish to speculate on why we found a U-shaped curve for the relationship between group size and opinion change rather than an inverted U-shape as predicted. Researchers have previously found evidence that individuals change their information search and processing strategies when faced with information overload, moving toward simpler, heuristic approaches (e.g., Cook, 1993; Helgeson & Ursic, 1993; Swain & Haka, 2000) that help to relieve the excess load. Participants in the smaller groups we observed may have used a compensatory strategy to weigh the arguments, and this produced more opinion change than in larger groups who then experienced information overload. For even larger groups, participants may have been forced to use a non-compensatory strategy—perhaps a 'majority-rules' heuristic—instead, leading opinion change to increase again. To test this possibility, it would be necessary to use some 'process-tracing' approach, such as concurrent verbal protocols or selection of arguments from an information board (Schulte-Mecklenbeck et al., 2017), which should be possible to accomplish using the SGRP.

Finally, the significant positive relationship found between participants' score on the AOT scale (Baron et al., 2015) and their propensity for opinion change further supports our conjectures regarding the role of predecisional information distortion in limiting such change, since the AOT scale includes several items testing for myside bias and predecisional distortion has been described as a measure of that bias (Baron et al., 2022).

This study moves forward the rather small literature on the effects of group size on performance—small, we argue, due to the logistic problems of constituting sufficient groups of various sizes to get sufficient statistical power using traditional methods with actual rather than nominal groups. With the SGRP we have sufficient power to have some confidence in our results; in other words, the observed effects, and lack of effects, are likely to be the case for our sample of participants, and this type of task. So, rather than looking to replicate with a larger sample, we can contemplate new experiments aimed at furthering our understanding of the factors moderating any group-size effects that may exist. In 'real-world' Delphi surveys, there tends to be substantial opinion change between rounds, which leads us to believe that it is our forecasting task, or participants, or the interaction of the two, that is

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

responsible for our major finding of rather low levels of opinion change. Thus, a starting point for future studies could therefore be analysis of the differences between our study and those 'real-world' surveys. Perhaps participants should be given cues as to the expertise levels of the other panelists (e.g., confidence in the judgment, sometimes fed back in real-world Delphis)? Perhaps non-binary forecasts might give more scope for opinion change as such change could be incremental (and most real-world Delphis use continuous or multi-point response scales)? The effects of such variables on VOC could be quickly tested on groups of varying sizes using the SGRP.

More generally, our experiment demonstrates the utility of the SGRP for investigating VOC in structured groups. While we did not find a conclusive answer to our questions about the effects of either group size or diversity, our findings do suggest that one cannot assume a MARRS such as Delphi will necessarily affect VOC as a matter of course, but likewise our results should not be interpreted as evidence that Delphi has no use. In some contexts—such as where most participants' judgments are correct from the outset, initial confidence levels are relatively high, and/or others' rationales are not sufficiently persuasive—judges' initial opinions may tend to be sticky even in the face of conflicting majority opinions. However, the effectiveness of SGTs is likely to be determined by an interaction between many human and task features; thus, a single study cannot address more than a few of these.

6. General discussion: extending JAS and beyond

The SGRP facilitates the extension of the high level of experimental control manifest in JAS studies to the study of other sorts of MARRS typically studied using traditional interacting groups. We argue that attempts to answer questions regarding, for example, the effectiveness of Delphi, have often been poorly executed (or not executed at all) because of logistic problems in recruiting interacting groups in sufficient numbers to attain reasonable power. The SGRP avoids these problems by allowing group designs to be run—with their attendant questions—while retaining VOC of individual agents as being the primary unit of analysis, thus permitting one individual to stand for one group.

The Delphi technique is very widely used in opinion survey practice, and there is a great deal of practitioner interest in applying the technique in the best possible way. For example, Belton et al.'s (2019) recent paper evaluating best practice for the use of Delphi has already been cited 202 times in the Google Scholar database (March, 2023). However, several of the paper's prescriptions for improved practice are based on advice of practitioners, rather than laboratory-based evidence, including the crucial issue of Delphi panel size, where many ranges have been suggested, as described above. Other important practice-based issues, such as the type and form of between-round feedback, whether the status and confidence of individual panelists should be revealed to other panelists, the number of Delphi rounds to be utilized, and the degree of heterogeneity to be sought in the expert panel, are also currently unsupported by laboratory-based evidence (see Belton et al., 2022 for a discussion). We believe that the SGRP will make the resolution of these practice-based issues considerably more tractable than the traditional approach using actual rather than nominal groups.

We believe that the SGRP can benefit JAS research too. It is true that there is no fundamental difference between the two approaches—the basic JAS design can be extended to answer any question that the SGRP can answer. In fact, in some cases this has been done already. For example, Hütter and Ache (2016) used a simulated advice approach to study advice taking that is very similar to what might be done in the SGRP—numerical estimates that were more or less extreme than the judge's initial estimates were generated by drawing from a pseudo-normal distribution and then fed back to participants as advice. However, an important difference between Hutter and Ache's work and how we see a typical SGRP study is that supporting rationales linked to the numerical estimates were not utilized—elicitation of qualitative responses (i.e., rationales) for feedback are a major reason for conducting an SEE in the SGRP.

The effect of qualitative reasons in support of quantitative advice have rarely been studied in JASs despite Von Swol & Sneizek's (2005) finding that such reasons had a major positive impact on advice acceptance (surprisingly, written reasons had a greater impact than those delivered face-to-

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

face). Further, Jungermann (1999) points out that real consultants will usually explain and justify their recommendations with arguments—this practice in its various forms has been described by decision researchers (e.g., Janis & Mann, 1977; Lipshitz, 1993; Montgomery, 1989; Simonson, 1989). Where more elaborated advice has been given in JASs it has focused on the persuasiveness of the message channel—verbal versus written—rather than characteristics of the message itself, for example, the quality of the arguments. Thus, one way in which SGRP provides an explicit advance on JASs is through the SEE where qualitative responses can be elicited for feedback in the main study (and categorized or manipulated to test specific hypotheses), so the procedure for producing simulated 'advice' is integral to the SGRP, but not JASs.

Hütter and Ache (2016) claim to be the first to study sampling of advice in JASs. For sampling to be investigated, there need to be several sources of advice to sample. In JASs, studies do not regularly consider multiple advice, but in a Delphi this is normal. Of course, JASs can easily be extended to the case of multiple advice, but the SGRP affords it by offering explicit procedures for generating that multiple advice—in other words, our paradigm normalizes more complex, and we argue, realistic, experimental designs than JASs. So, sampling strategies could easily be studied within the SGRP by, for instance, using an information board (e.g., Mouselab: Payne et al., 1993) to record the sequence in which judges search for information and/or what sources they choose to consult in what order (perhaps with costs to searching imposed). Another obvious manipulation would be to compare sequential information presentation—as can be the case in 'real time Delphi' (Gordon & Pease, 2006), where judges can post their judgments and see those judgments already posted, whenever they like—with parallel presentation, as in a traditional Delphi where panelists can only post and review judgments once per round. The sequence in which information is received is not just an issue in evaluating different forms of Delphi, but also in JASs, for example in the work of DeKay (2015) discussed above, on how decisions can affect the evaluation of subsequent information.

Our work has several similarities to social judgment theory (SJT: Hammond et al., 1975). Both offer frameworks for investigating influences on human judgment, and both offer methods to assist this study. SJT is based on Brunswik's lens model (Brunswik, 1952) of the relationship between judges and their environments, where it is proposed that judges learn to use a set of cues (e.g., symptoms) to make judgments of a criterion (e.g., the existence or extent of a disease). In SJT methods are provided for discovering judges' cue-criterion 'judgment policies', whereas in the SGRP we provide methods for finding the influences on VOC—in both approaches judgment strategies are described as linear models. Determining the relative influences of environmental and personal features on judgment accuracy is also crucial to both approaches. Finally, both approaches allow for the study of collective judgment—in SJT's case, in the form of learning policies from others ('interpersonal learning') and resolving differences in policies so as to reach a consensus ('interpersonal conflict') (see Dhami & Mumpower, 2018; Dhami & Olsson, 2008 for a review). Although neither learning nor consensus reaching have been discussed in the current paper, we believe that SGRP holds considerable promise for research in both areas. Regarding interpersonal learning, we wish to note that the degree to which behavioral change observed in groups is accompanied by belief (cf. policy) change has been a central question ever since the Asch experiments. In the case of consensus, although we have focused on improving accuracy of group judgments in this paper, consensus is often a desirable outcome, and may be the best outcome where there is no 'right' answer.

There are some differences between the SGRP and SJT, though. The most important difference is that SJT assumes judgment tasks where there is at least one identifiable cue for predicting a criterion, and the predictive validity of that cue can be determined empirically. We argue that this assumption does not hold for many of the tasks where structured groups of experts are used—what precisely constitutes a cue may be unclear, while the means for determining cue validity may be similarly obscure or unavailable. For example, consider the forecasting question we used in our study 'Putin will send troops into territory belonging to another country' (see Appendix 2). This was answered with reference to, *inter alia*, specifics of past Russian actions and Putin's intentions and character, which although

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

could be cast as cues, might better be treated as arguments, which are more or less persuasive, rather than more or less valid. While what constitutes a cue is a debate for another forum, what is clear is that even if we do take such things to be cues, their rare, or sometimes unique, nature renders learning of their validity by repeated experience of cue-criterion correspondences difficult if not impossible. In practice, we usually call in panels of experts to make judgments exactly in those instances where there is little data for statistical approaches to be used, and it is on these types of problem for which we see the greatest benefit for SGRP research.

In conclusion, two further thoughts. First, we believe that research into opinion formation and change in *online* groups is particularly timely due to the huge increase in group-working over the Internet resulting from the recent pandemic, which is likely to remain more common than beforehand. We submit that research using VOC modeling and the SGRP can help investigate how such online information exchanges and opinion makers can lead to improved decision-making in organizations and wider society (e.g., for promoting desired environmental and health behaviors, or increasing public participation in government) rather than perpetuating—or even amplifying—bias, prejudice, and misinformation. Second, an exciting possibility is the prospect of researchers throughout the world using SEEs for a range of problems (judgmental, decision-making, and problem solving) and then making their databases accessible more widely. This approach could significantly shorten the research process (as researchers would not all have to conduct their own SEEs), aiding consistency, and potentially leading to a quantum leap in the number of rigorously conducted, well-analyzed studies. Results from such research could then be used inform the conduct of real-world structured-group instantiations seeking specific answers to strategy, policy, and other issues, thereby increasing the likelihood of producing high-quality answers to important questions.

Data availability statement. The dataset for this article can be found at https://osf.io/3wsfj/.

Author contribution. The work reported here is part of a large project with many personnel. All authors listed here are part of a team who contributed to the larger project in some way or another, for example, developing software and training, constructing experimental materials, running participants, and analyzing data. It was agreed at the outset that all team members should be included on all outputs regardless of their specific contribution. The first two authors were principally responsible for the writing of this paper, the remaining authors have assisted with data analysis and/or commented on drafts, and are listed in alphabetical order.

Funding statement. This research is based on work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), under Contract [2017-16122000003]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Competing interest. The authors declare none.

References

Asch, S. E. (1951). Effects of group pressure on the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership and men* (pp. 177–190). Pittsburgh, PA: Carnegie Press.

Barge, J. K., & Keyton, J. (1994). Contextualizing power and social influence in groups. In L. R. Frey (Ed.), *Group communication in context* (pp. 85–106). Hillsdale, NJ: Lawrence Erlbaum.

Baron, J., Isler, O., & Yilmaz, O. (2022, February 4). Actively open-minded thinking and the political effects of its absence. https://doi.org/10.31234/osf.io/g5jhp.

Baron, J., Scott, S., Fincher, K., & Emlen Metz, S. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, *4*(3), 265–284.

Bednarik, P., & Schultze, T. (2015). The effectiveness of imperfect weighting in advice taking. *Judgment and Decision Making*, *10*(3), 265–276.

Belton, I., Cuhls, K. & Wright, G. (2022). A critical evaluation of 42, large-scale, science and technology foresight Delphi surveys. *Futures & Foresight Science*, *4*, e2118. https://doi.org/10.1002/ffo2.118

Belton, I., MacDonald, A., Wright, G., & Hamlin, I. (2019). Improving the practical application of the Delphi method in group-based judgment: A six-step prescription for a well-founded and defensible process. *Technological Forecasting and Social Change.*, *147*, 72–82.

Bisson, J. I., Tavakoly, B., Witteveen, A. B., Ajdukovic, D., Jehel, L., Johansen, V. J., Nordanger, D., Garcia, F. O., Punamaki, R.-L., Schnyder, U., Sezgin, A. U., Wittmann, L., & Olff, M. (2010). TENTS guidelines: Development of post-disaster psychosocial care guidelines through a Delphi process. *British Journal of Psychiatry*, *196*, 69–74.

Boje, D. M., & Murnighan, J. K. (1982). Group confidence pressures in iterative decisions. *Management Science*, *28*(10), 1187–1196.

Bolger, F. (2018). The selection of experts for (probabilistic) expert knowledge elicitation. In L. Dias, A. Morton, & J. Quigley (Eds.), *Elicitation. International Series in Operations Research & Management Science* (Vol. 261, pp. 393–443). Cham, Switzerland: Springer.

Bolger, F., Stranieri, A., Wright, G., & Yearwood, J. (2011). Does the Delphi process lead to increased accuracy in group-based judgmental forecasts or does it simply induce consensus amongst judgmental forecasters? *Technological Forecasting and Social Change*, *78*, 1671–1680.

Bolger, F., & Wright, G. (2011). Improving the Delphi process: Lessons from social psychological research. *Technological Forecasting and Social Change*, *78*, 1500–1513.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*, 27–151.

Brockhoff, K. (1975). The performance of forecasting groups in computer dialogue and face-to-face discussion. In H. Linstone & M. Turoff (Eds.), *The Delphi method: technique and applications*. London: Addison-Wesley.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago, IL: University of Chicago Press.

Buck, R., Losow, J. I., Murphy, M. M., & Costanzo, P. (1992). Social facilitation and inhibition of emotional expression and communication. *Journal of Personality and Social Psychology*, *63*(6), 962.

Chidambaram, L., & Tung, L. L. (2005). Is out of sight, out of mind? An empirical study of social loafing in technology-supported groups. *Information Systems Research*, *16*(2), 149–168.

Clayton, M. J. (1997). Delphi: a technique to harness expert opinion for critical decision-making tasks in education. *Educational Psychology*, *17*(4), 373–386.

Cook, G. J. (1993). An empirical investigation of information search strategies with implications for decision support system design. *Decision Sciences*, *24*, 683–699.

Czaplicka-Kolarz, K., Stanczyk, K., & Kapusta, K. (2009). Technology foresight for a vision of energy sector development in Poland till 2030. Delphi survey as an element of technology foresighting. *Technological Foresight & Social Change*, *76*(3), 327–338.

Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi Method to the use of experts. *Management Science*, *9*(3), 458–467.

Dalkey, N. C. (1969). The Delphi Method: An experimental study of group opinion. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_memoranda/RM5888.html (accessed 15 November 2022).

Dalkey, N. C. (1975). Toward a theory of group estimation. In H. A. Linstone & M. Turoff (Eds.), *The Delphi method: techniques and applications* (pp. 236–261). Reading, MA: Addison-Wesley.

Davis, J. H. (1969). *Group performance*. Reading, MA: Addison-Wesley.

DeKay, M. L. (2015). Predecisional information distortion and the self-fulfilling prophecy of early preferences in choice. *Current Directions in Psychological Science*, *24*(5), 405–411.

Delbecq, A. L. & Van de Ven, A. H. (1971). A group process model for problem identification and program planning. *Journal of Applied Behavioral Science*, *7*, 466–491.

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). *Group techniques for program planning*. Glenview, IL: Scott, Foresman, and Co.

Devine, D. J., Clayton, L. D., Dunford, B. B., Seying, R., & Pryce, J. (2001). Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, Public Policy, and Law*, *7*(3), 622.

Dhami, M. K., & Mumpower, J. (2018). Kenneth R. Hammond's contributions to the study of human judgment and decision making. *Judgment and Decision Making*, *13*, 1–22.

Dhami, M. K., & Olsson, H. (2008). Evolution of the interpersonal conflict paradigm. *Judgment and Decision Making*, *3*, 547–569.

Economist (2016). Who said Brexit was a surprise? *The Economist*, 24 June. ISSN 0013-0613.

Edwards, W., Lindman, H., & Phillips, L. D. (1965). Emerging technologies for making decisions. In F. Barron, et al. (Eds.), *New Directions in Psychology II* (pp. 261–325). New York, Holt: Rinehart & Winston.

Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from organizational science, accounting, marketing, MIS, and related disciplines. *Information Society*, *20*(5), 325–344.

Galton, F. (1907). Vox populi. *Nature*, *75*(7), 450–451.

Gelman, A. & Rothschild, D. (2016). Something's odd about the political betting markets. *Slate*, 12 July. ISSN 1091-2339.

Gordon, T. J., & Pease, A. (2006). RT Delphi: An efficient, "round-less", almost real time Delphi method. *Technological Forecasting and Social Change*, *73*(4), 321–333.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

Hackman, J., & Morris, C. (1975). Group tasks, group interaction processes, and group performance effectiveness: A review and proposed integration. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. *8*, pp. 45–99). New York: Academic Press.

Hammond, K. R., Stewart, T. R., Brehmer, B., & Steinmann, D. O. (1975). Social judgment theory. In M. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes* (pp. 271–312). San Diego, CA: Academic Press.

Hanea, A., McBride, M., Burgman, M., Wintle, B., Fidler, F., Flander, L., Twardy, C. R., Mascaro, S., & Manning, B. (2016). InvestigateDiscussEstimateAggregate for structured expert judgement. *International Journal of Forecasting*, *33*(1), 267–269.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, *70*, 117–133.

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing*, *32*(4), 1008–1015.

Healy, P. J., Linardi, S., Lowery, J. R., & Ledyard, J. O. (2010). Prediction markets: Alternative mechanisms for complex environments with few traders. *Management Science*, *56*(11), 1977–1996.

Helgeson, J. G., & Ursic, M. L. (1993). Information load, cost/benefit assessment and decision strategy variability. *Journal of the Academy of Marketing Science*, *21*(1), 13–20.

Herzog, S., & Hertwig, R. (2009). The wisdom of many in one mind improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231–238.

Hussler, C., Muller, P., & Rondé, P. (2011). Is diversity in Delphi panelist groups useful? Evidence from a French forecasting exercise on the future of nuclear energy. *Technological Forecasting and Social Change*, *78*(9), 1642–1653.

Hütter, M., & Ache, F. (2016). Seeking advice: A sampling approach to advice taking. *Judgment and Decision Making*, *11*(4), 401.

Hwang, M. I., & Lin, J. W. (1999). Information dimension, information overload and decision quality. *Journal of Information Science*, *25*(3), 213–218.

Janis, I. L. (1971). Groupthink. *Psychology Today*, *5*(6), 84–90.

Janis, I. L. & Mann, L. (1977). Decision making: A psychological analysis of conflict, choice, and commitment. New York: Free Press.

Janis, I. L. (1982). Groupthink: Psychological studies of policy decisions and fiascos. Boston: Houghton Mifflin.

Johnson, E. J., & Payne, J. W. (1985). Effort and accuracy in choice. *Management Science.*, *31*(4), 395–414.

Jungermann, H. (1999). Advice giving and taking. In *Proceedings of the 32nd annual Hawaii international conference on systems sciences* (pp. 1–11). IEEE.

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgment: comparing individuals and groups. *Psychological Review*, *103*(4), 687–719.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*, 623–656.

Kerr, N. L., & Tindale, R. S. (2011). Group-based forecasting: A social psychological analysis. *International Journal of Forecasting*, *27*, 14–40.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: Chicago University Press.

Larson, J. R. (2010). In search of synergy in small group performance. Hove, East Sussex: Psychology Press.

Levingston, I. (2016). Why political polls and betting odds disagree with each other so much, *CNBC*, 28 July.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge: Cambridge University Press.

Linstone, H. A., & Turoff, M. (1975). *The Delphi method: Techniques and applications*. London: Addison-Wesley.

Lipshitz, R. (1993). Decision making as argument-driven action. In G. A. Klein, J.. Orasanu, R. Calderwood & C. E. Zsambok (eds.), *Decision making in action: Models and methods* (pp. 172–181). Norwood, NJ: Ablex.

McGrath, J. E. (1984). *Groups: Interaction and performance*. Englewood Cliffs, NJ: Prentice Hall.

Montgomery, H. (1989). From cognition to action: The search for dominance in decision making. In H. Montgomery & O. Svenson (eds.), *Process and structure in human decision making* (pp. 23–49). New York: Wiley.

Moscovici, S. (1985). Social influence and conformity. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (3rd ed., pp. 347–412). New York, NY: Random House.

Nijstad, B. A. (2009). *Group performance*. New York, NY: Psychology Press.

Nyberg, E. P., Nicholson, A. E., Korb, K. B., Wybrow, M., Zukerman, I., Mascaro, S., Thakur, S., Oshni Alvandi, A., Riley, J., Pearson, R., Morris, S., Herrmann, M., Azad, A., Bolger, F., Hahn, U., & Lagnado, D. (2021). BARD: A structured technique for group elicitation of Bayesian Networks to support analytic reasoning. *Risk Analysis 42*, 1155–1178. https://doi.org/10.1111/risa.13759

Parenté, F. J., & Anderson-Parenté, J. K. (1987). Delphi inquiry systems. In G. Wright & P. Ayton (Eds.), *Judgmental forecasting* (pp. 129–156). Chichester: Wiley.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). Appendix: The Mouselab system. In *The adaptive decision maker* (pp. 264–278). Cambridge, UK: Cambridge University Press.

Rothstein, H., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, England: Wiley.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

Rowe, G., & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting*, *12*, 73–89.
Rowe, G. & Wright, G. (2001). Expert opinions in forecasting: The role of the Delphi technique. In J. S. Armstrong (ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 125–144). Boston, MA: Kluwer Academic.
Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., & Jinks, C. (2018). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality and Quantity*, *52*(4), 1893–1907.
Schulte-Mecklenbeck, M., Johnson, J. G., Böckenholt, U., Goldstein, D. G., Russo, J. E., Sullivan, N. J., & Willemsen, M. C. (2017). Process-tracing methods in decision making: On growing up in the 70s. *Current Directions in Psychological Science*, *26*(5), 442–450.
Simmons, J. P., Nelson, L. D., Galak, J., & Frederick, S. (2011). Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research*, *38*(1), 1–15.
Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research*, *16*, 158–174.
Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, *48*(6), 1467–1478.
Stasser, G., & Vaughan, S. I. (2013). Models of participation during face-to-face unstructured discussion. In E. H. Witte & J. H. Davis (Eds.), *Understanding group behavior: consensual action by small groups* (Vol. 1, pp. 165–192). New York: Psychology Press.
Stebler, N., Schuepbach-Regula, G., Braam, P., & Falzon, L. C. (2015). Use of a modified Delphi panel to identify and weight criteria for prioritization of zoonotic diseases in Switzerland. *Preventative Veterinary Medicine*, *121*, 165–169.
Steiner, I. D. (1972). *Group process and productivity*. New York: Academic Press.
Straus, S. G. (1999). Testing a typology of tasks: An empirical validation of McGrath's (1984) group task circumplex. *Small Group Research*, *30*(2), 166–187.
Surowiecki, J. (2005). *The wisdom of crowds*. New York: Anchor Books.
Swaab, R. I., Phillips, K. W., & Schaerer, M. (2016). Secret conversation opportunities facilitate minority influence in virtual groups: The influence on majority power, information processing, and decision quality. *Organizational Behavior and Human Decision Processes*, *133*, 17–32.
Swain, M. R., & Haka, S. F. (2000). Effects of information load on capital budgeting decisions. *Behavioral Research in Accounting*, *12*, 171–198.
Tindale, R. S., & Winget, J. R. (2019). Group decision-making. In *Oxford Research Encyclopedia of Psychology*. Retrieved from https://oxfordre.com/psychology/display/10.1093/acrefore/9780190236557.001.0001/acrefore-9780190236557-e-262;jsessionid=7A66CF810B8F11B07FA9FA03C62243A8 (accessed 23 October 2022).
Toma, C., & Picioreanu, I. (2016). The Delphi technique: Methodological considerations and the need for reporting guidelines in medical journals. *International Journal of Public Health Research*, *4*(6), 47–59.
Van Swol, L. M. & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British Journal of Social Psychology*, *44*(3), 443–461.
Vinokur, A., & Burnstein, E. (1974). The effects of partially shared persuasive arguments on group induced shifts: A group problem solving approach. *Journal of Personality and Social Psychology*, *29*, 305–315.
Vroom, V. H., & Yetton, P. (1973). *Leadership and decision-making*. Pittsburgh, PA: University of Pittsburgh Press.
Watson, W. E., Johnson, L., Kumar, K., & Critelli, J. (1998). Process gain and process loss: Comparing interpersonal processes and performance of culturally diverse and non-diverse teams across time. *International Journal of Intercultural Relations*, *22*(4), 409–430.
Weber, B., & Hertel, G. (2007). Motivation gains of inferior group members: A meta-analytical review. *Journal of Personality and Social Psychology*, *93*(6), 973–993.
Whyte Jr. , W. H. (March 1952). Groupthink. *Fortune*. 114–117.
Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, *18*, 107–126.
Worchel, S., Rothgerber, H., Day, E. A., Hart, D., & Butemeyer, J. (1998). Social identity and individual productivity within groups. *British Journal of Social Psychology*, *37*, 389–413.
Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*, 260–281.

Appendix 1: The SGRP in detail

Phase 1: Stimulus creation/collection

There are three basic options at phase 1: create stimulus items to be used in phase 2 *de novo*; adapt some existing stimuli to suit your purpose; or elicit responses that can be used in phase 2 during a pilot study that we refer to as a 'stimulus elicitation exercise' or SEE. In some circumstances it might be clear

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

what possible responses there could (sensibly) be from participants in a Delphi. For example, perhaps the task of interest is judging the value of an uncertain quantity (but with a known or soon-to-be-known value) that can only fall within a certain range, and your research questions concern the potential biasing effects of different distributions of answers around the true value (e.g., generally too high or too low). Similarly, qualitative responses might realistically fall only in certain categories, or types, which could be organized in different ways to, for example, investigate systematically the relative contributions of the draw of credibility or of a majority. A further possibility could be to generate responses according to some theoretical model of experts, tasks, or some combination of the two. For instance, potential distributions of judgments of uncertain quantities might be generated on the basis of assumption that the experts are Bayesian, or deviate from Bayes in plausible ways, for example, 'conservative' updaters (e.g., Edwards et al., 1965) for tasks that vary in terms of the quality of the feedback available. In other circumstances there may be an existing set of stimuli from published research that can be used to test a new research question (or perhaps attempt full or partial replication of an existing finding). Another possibility is to use responses already collected in a previous SEE—by yourself or other researchers—indeed, we encourage researchers in this area who use the SGRP to store any stimuli they elicit in a publicly accessible database that can be reused by interested parties in future studies. If stimuli cannot credibly be created *de novo* and cannot be drawn from previous research, then it is necessary to conduct an SEE.

The SEE

This begins with what is essentially a traditional first-round Delphi process. In this, participants are instructed (as per a typical Delphi) to answer one or more questions of interest with an expectation that their answers will be used as feedback to a nominal group: It is important that participants have the expectation that their answers will be used in this way so as to motivate them to produce responses that have value to others, and thus keep the process as similar to a Delphi as possible. Meanwhile, participants might range from the totally naïve to those with considerable expertise on the topic of concern, depending on the research questions and the practicalities of recruitment. Under normal circumstances, however, we advise that the SEE participants be drawn from the same population as those who will be used in phase 2 (again to maintain compatibility with the standard Delphi process[15]).

In this initial step—'**Collect** responses' in Figure 2—information collected in addition to answers could include rationales for judgements (pro arguments), rationales for alternatives (asking participants, e.g., 'Why might your answer be wrong?'), confidence ratings, self-ratings of participants' own expertise on the topic, and so on. Responses to one or more personality questionnaires might also be collected in order to establish participants' characteristics (see Figure 1)[16]. Importantly, the SEE is an opportunity to collect as much potentially relevant information as possible, not only on factors that are of immediate interest to the researcher, but also on factors that may be of interest in the future.

In our view, this first step is then best regarded as a broad data acquisition stage and accordingly we propose an optional second step—'**Collate** responses' in Figure 2—which involves *creating* a *database* of SEE responses for each question posed. In a third potential step—'**Categorize** responses' in Figure 2—you can apply some *categorization or manipulation* to the data collected. For example, one analysis might simply involve conducting a word count of participants' rationales—something as simple as *argument length* might prove an influencing factor on opinion change over rounds (the main criterion measure). However, more sophisticated analyses could also be done, such as rating arguments for their *quality* or *complexity* or *persuasiveness*. The collated material in the SEE database is a resource that can be used in a series of studies, and can be re-analyzed and rated according to new hypothetically

[15]It is possible to imagine exceptions to this rule, though. For instance, you might wish to run an experiment with a setup more akin to JAS than Delphi, in which case you could use experts in the SEE and naïve participants in phase 2. See the Q&A section in the Discussion for commentary on wider use of the SGRP.

[16]For example, you might want to feed back responses from particular kinds of participants at phase 2.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

interesting factors that emerge later. Indeed, there is even the possibility that large databases could be made open source, for use by the entire research community.

It should by now be apparent that the SGRP can be used in at least two different ways: In a *deductive* mode the experimenter will start with one or more hypotheses and develop stimuli to test these, perhaps necessitating an SEE if appropriate stimuli cannot be found elsewhere or constructed *de novo*; alternatively, in an *inductive* mode, data can be collected prior to the formation of hypotheses. In the latter case, after categorization there would be a fourth step—'**Construct** hypotheses' in Figure 2—where research questions can be identified and specific hypotheses formulated (of course, in a deductive mode you would skip this step since you would already have constructed your hypotheses at the outset)[17].

In a fifth SEE step—Figure 2, '**Curate** responses'—the appropriate stimulus material for the desired Delphi experiment (i.e., quantitative and/or qualitative responses to the task) is selected, and possibly edited, for use as feedback to real participants in the second Delphi round of phase 2. For quantitative forecasting, this editing typically involves computing some statistics, such as the mean and standard deviation of the forecasts, which are then fed back alongside, or instead of, the original responses. Where qualitative responses are given, including rationales, the facilitator might, for instance, edit the language to assist comprehension or maintain anonymity of the panelist, or combine several similar arguments into one so as to reduce majority/minority effects.

Stimulus selection

In the SGRP, the experimenter will usually need to select $n - 1$ stimuli (i.e., the group size chosen for the study less the target participant) from the larger set (produced by the SEE or otherwise generated) for presentation to the participants at phase 2, round 2. There are various ways to do this selection, the choice of which should be determined by the purpose of the study being conducted.

The first approach is **Random Selection** [18]. If a proposed study is largely exploratory, with no *a priori* predictions or insights, then the simplest approach is to randomly select the requisite data from the SEE database. For example, if there are 20 entries in a database and it is intended to study nominal Delphi groups of size six, then data (solutions, rationales) from a quarter of the sampled participants (five) would be obtained from the database using a random selection process, and the process repeated *for each nominal group created* (i.e., the set of responses fed back to a single participant at phase 2). Random selection is the selection approach that is most akin to an ordinary Delphi panel and is therefore well suited for running SGRP control groups against which performance of participants in experimental conditions—with manipulated (i.e., non-random) responses—can be compared.

Often, we will want to exert more control over the stimuli that are fed back than with random selection. For instance, we may not have a very large SEE database relative to the degree of potential variation in stimuli. In this case, random selection could lead to some kinds of responses being represented infrequently, or not at all. Here, we could balance the types of responses by randomly selecting from within types ('stratified' sampling). For example, we may wish to balance right and wrong responses, long and short, or quantitative and qualitative. These and other methods for constructing stimulus sets which, while not strictly random, attempt to maintain some degree of randomness we refer to as **Pseudo-Random Selection** [19].

[17]If you have hypotheses at the outset, then they may well shape the SEE, for example, if you wanted to manipulate quality of arguments, then you might ask SEE participants to each produce several arguments with the hope that those produced later will be less convincing than the first ones produced. However, it is not essential to construct hypotheses before the SEE, for example, phase 1 rationales for forecasts could be categorized in some way after the fact—perhaps as different types of causal explanation—then one could formulate a phase 2 experiment to test hypotheses about the relative influence of these different types of explanation on VOC.

[18]We use the term 'selection' to refer to determining responses to be used, to distinguish from 'sampling', which we use to refer to choice of participants

[19]A tutorial on sampling methods is beyond the scope of this paper: We simply wish to give a sense of the main options available to researchers wishing to select stimuli in the SGRP. Please note that we are using 'pseudo-random' in a looser sense

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

A third approach is to make a **Representative Selection**, in which the participants' responses are selected from the SEE database based on the degree to which database participants, or their responses, are representative of their respective populations in the real world, on certain criteria deemed of relevance to the researcher. For example, consider a question in which participants were asked in an SEE to choose between two options (e.g., 'Which is further north, Rome or New York?'): if two thirds of the population choose New York and one third Rome, then a representative selection approach would lead to using responses from the database that matched that ratio. Representative selection could be based on all sorts of factors, from demographic or personality characteristics of database participants to numerical or qualitative aspects of elicited responses. Of course, there are difficulties with implementing representative selection, not least in determining what patterns are 'representative of the population in the real world' as this may depend on there being a large body of relevant previous research, while justifying relevant criteria for judging representativeness can likewise be problematic.

A fourth approach to selecting feedback from the database is more apt when you have a specific research question in mind—for instance, you want to study the effects of particular types of feedback to answer 'what if?' questions—in other words, using the SGRP in the deductive mode. We refer to this as **Informed Selection**: Data are chosen from the database informed by the research questions. As a case in point, one issue that has been of concern in social psychology more generally (as well as in Delphi research) is the extent to which people are affected by whether they are in the majority or minority with regard to their initial opinion (e.g., Asch, 1951; Moscovici, 1985; Swaab et al., 2016). If you wished to test the factors that might be likely to make a minority more influential than is usual, then clearly you would need to select from the SEE database an imbalance of responses, so that most argued for one option, judgment, or problem solution, rather than another/others. The option(s) most likely to represent the minority position among the experimental sample could be determined from the SEE data.

As already mentioned, stimuli collected in SEE might be edited before being used as feedback in the experimental Delphi. When using informed selection, this editing might go beyond simply correcting spelling or grammatical errors, or amending language to appear more professional or easier to read, to actually constructing new feedback. This might be done by significantly amending responses or by actively writing novel feedback, perhaps to fill conceptual holes in your dataset, or indeed, introducing somewhat outrageous or outlying arguments if there was a wish to study, for example, the effect on possible VOC of sensational or extreme opinions (e.g., 'fake news'). Importantly, even if you are extensively modifying the stimuli, the SEE still fills an important role by revealing the views of a typical sample and the language and terminology used—and hence acts as a benchmark against which the novelty of devised material could be contrasted.

Finally, if we have not performed an SEE, and simply have a set or sets of responses that we have constructed for a particular purpose, then no selection will be required. Alternatively, it is possible that a SEE database might throw up essentially very few different responses (or different *types* of response) for some tasks and thus every single response (or exemplar of every type of response) might be used as feedback in Phase 2 (i.e., little or no stimulus selection).

Phase 2: Experimentation

Main study

In this phase, the main experimental[20] Delphi study is conducted. We suggest using two rounds, because most response change occurs after the first round and because this is also experimentally easier (although other designs are possible—see the Discussion section below). Usually, a new set

than the more formal one referring to number generation by a precise mathematical procedure that yet still satisfies at least one statistical test for randomness.

[20]We use the terms 'experimentation' and 'experimental' here deliberately to emphasize that although phase 1 might be either deductive or inductive, we see phase 2 as being primarily a deductive one (i.e., we would expect that researchers normally enter phase 2 with a hypothesis or hypotheses to test). This is in accordance with our research agenda as set out in the Introduction.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

of participants are recruited for the main study (i.e., those who have not previously taken part in the SEE[21]): As already mentioned, in order to reproduce a typical Delphi panel of peers, these participants should be sampled from the same population as those from whom responses were collected in the SEE. At round 1, each participant individually responds to the target problem(s). Note that these first-round responses could be collected and further added to the SEE database, thus enriching it: Although the feedback for the second round would have already been identified, the responses could potentially provide stimuli for future experiments. Next, all of the participants in round 2 receive feedback, specific to their experimental condition, from a set of anonymous, nominal others (ostensibly those in 'their group'), to consider and use as a basis to refine their own initial judgements (or not).

Main study analysis

Finally, the responses from the phase 2 experimental study are analyzed to answer the research questions. Importantly, as we have already discussed, the level of analysis is usually at the individual level, in which the effects of specific, controlled feedback on each of the *n* participants in each condition can be established: In other words, the amount of opinion change is recorded for every participant and assessed as to whether it is 'virtuous' or not. Since we are focusing on individual outputs rather than on group, and thus have independent observations, we only need *n* participants rather than approximately *n* times the group size for inferential statistics to be performed with sufficient power to enable the research questions to be answered with an acceptable degree of confidence.

Aside from this greater efficiency, we posit that the main advantage of the SGRP over 'real' Delphi groups is in *experimental control*. What the approach enables us to do is essentially answer 'what if' questions: *what if* participants had certain qualities, or the feedback was of a certain type, or the process was structured in a certain way, or the questions to be answered were of a certain type? Answering these questions not only allows academic learning about what factors in nominal groups are effective in stimulating virtuous opinion change (or any change), as well as testing theoretically interesting hypotheses as to *why* these may prevail, but also has practical implications. If we have a 'real-world' Delphi problem to solve, with real experts, the appropriate SGRP studies should be able to inform exercise sponsors about the type of experts to recruit, in what numbers, and using which processes, in order to enhance the *likelihood* of getting improved judgments. Compared to this, traditional Delphi empirical studies with 'real' groups essentially rely upon serendipitous recruitments and interactions to yield convincing and publishable results, and absence of control means that all most studies can say is that 'on this particular occasion, this particular assemblage of participants happened to be aided by this particular variation of the Delphi process'[22]. As to the question of whether a similar result would be obtained under any minor variation of the factors, such traditional studies would likely be mute.

Discussion: Some answers to questions that might be asked of the SGRP

What is the relationship between the SGRP and traditional structured-group research?

We anticipate that many readers will be interested to know the extent to which the SGRP can substitute for traditional structured-group research with 'real' groups. Predictably, perhaps, the answer is 'it depends. . .'. While it is possible to imagine variants of SGRP that are to all intent and purposes identical to traditional Delphi—and we will touch upon such designs briefly below in the answers to a couple of the other questions raised—such an approach loses much of the point of the SGRP. This point being—as discussed above in relation to the paradigmatic nature of SGRP—*to assist the design of tightly controlled experiments for answering questions about influences on opinion change* and, as such, the

[21] It is possible, although logistically problematic, to reuse SEE participants in the main study—see the Q&A in the Discussion.

[22] Journals have a preference for research that reveals findings in which we can have some confidence, usually indicated by demonstrating the statistical significance of claimed differences between conditions, with a bias against studies that yield null results (e.g., Rothstein et al., 2005). We can only speculate on the number of unpublished empirical studies of Delphi languishing on researchers' hard drives.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

principal raison d'etre of SGRP is *not* to simulate real groups but rather to remove aspects of real groups that prevent the efficient investigation of some important research questions: Using SGRP to simulate Delphi with real groups inevitably means reintroducing that inefficiency and poor control. We made a similar point earlier with regard to the consequences of allowing direct interaction between participants in a structured group; thus, this is clearly an even greater issue for those techniques which do—the nominal group technique (NGT, e.g., Delbecq & Van de Ven, 1971) and Investigate Discuss Estimate Aggregate (IDEA, e.g., Hanea et al., 2016)—but less of one for those methods where interaction is more constrained than Delphi: PMs —a form of crowdsourcing (e.g., Surowiecki, 2005)—and JASs.

All this notwithstanding, there may still be circumstances when you would wish to use more complex or 'realistic' SGRP designs, as we will illustrate shortly. Meanwhile, considering a basic SGRP with just two rounds in phase 2, as presented in Figure 2, although from the viewpoint of a participant the experience is the same, there is an important difference between the *output* of an SGRP 'group' and that of a traditional structured group. In basic SGRP the output is an individual's opinion (potentially) revised after seeing several other opinions, but for traditional structured groups the output is the aggregate of the (potentially) revised opinions of *all* the group members[23]. So, the SGRP arrangement is more akin to JASs than Delphi: A participant is offered some advice and determines whether or not to revise his or her original view. Since our basic motivation for developing the SGRP is to assist research into factors affecting opinion change, this 'limitation' does not trouble us. Further, as we have discussed above, JASs are themselves important judgment-aiding technologies and the subject of a significant research literature (see Bonaccio & Dalal, 2006, for a review).

Can SEE participants also take part in the main study?

Participants in the SEE can be used in a number of ways. First, they could continue to be used in a traditional study, for example, receiving feedback from their 'real' nominal group(s) in a second round (and so on). This might be a sensible ploy if real experts have been recruited, and the results would still be of use in establishing their particular views on the central topic of the survey. Second, they could, after the first round, be informed that there is in fact no second round (but see the point below about deception), with their involvement ending at that point. Third, they could continue to be used in the study proper, but rather than receiving feedback from some *pre-determined* nominal group (a subset of the recruited panelists), they would then receive feedback from their own SEE database (or indeed, data from previous entries in this database) according to the experimental intent and design of the study. The latter approach seems the most efficient use of recruited panelists, although there might be logistical difficulties in rapidly entering their preliminary data into the SEE database and identifying apt feedback without the length of inter-round gap becoming excessive. This approach might therefore be best for studies involving Random Selection, and least good for those involving Informed Selection (the latter needing the greatest amount of data treatment).

Can there be more than two rounds in the SGRP study?

The SGRP as described here relates to two-round protocols only, which are prevalent in practice. However, it is perfectly possible that the approach could be extended for use in studies with three or more rounds. There are a number of ways in which this could be accomplished, including: use of confederates; constructing plausible revised answers; selecting appropriate revised answers from previous studies; and having the same participants respond over the multiple SGRP rounds while maintaining the integrity of the virtual groups they see at each round (cf., reusing SEE participants as discussed in the answer to the previous question): The precise method used would depend on

[23] In which case you have just one data point per group. An alternative for interacting groups would be to take each individual group member's revised opinion as the unit of analysis, but unlike the SGRP output these observations are *not* independent of each other, so doing this does not help much to increase analytic power.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

the research questions being posed and the characteristics of the experts and judgment task. Such studies might reveal important additional influences on VOC, such as how people with particular personal characteristics respond to certain types of feedback; however, adding further rounds increases complexity and potentially reduces some of the benefits of SGRP relative to traditional structured-group methods with real groups, as we argued in the answer to the first question above. We therefore submit that extending the SGRP to additional rounds is superfluous at this time, when there is still so much to learn about changes by individuals to initial feedback from nominal others.

How many participants do you need in the SEE, and of what sort?

If external validity is a concern, then it is important to ensure that the dataset used in the main experiment truly includes the full and relevant range of responses—which might not be the case if the SEE were conducted with an especially small sample, or with a sample that was in some way unusual or biased (e.g., experts from a single country or industry where the issue is of relevance to multiple countries or industries). As an example, consider making a forecast of a geopolitical event. You might conduct the SEE with, say, 30 participants, asking for rationales regarding why that event might occur and why it might not occur. If a small range of rationales were uncovered that were frequently mentioned, then you might feel confident that you had acquired most of the issues via the SEE. However, if there were a large range of rationales produced, few of which were repeated, then this might cause some concern that the elicitation exercise had not fully captured the real-life richness of factors on which people would likely judge that particular problem. In such a case, we would recommend further sampling of subjects in a rolling SEE until 'saturation' is reached—a concept used in qualitative social science research where data are gathered (e.g., in interviews or focus groups) until no further significant themes are produced. For an extensive discussion of the saturation concept and methods, see Saunders et al. (2018). Unfortunately, since each judgment or reasoning task is different, devising a universal rule for calculating SEE sample size seems to us infeasible. The critical point about sample size for the SEE is how many different valid responses (or response types) there are in the population. If only one response, or type of response, is possible (or few, as in the case of n-alternative forced choice tasks), then you may not need to do an SEE at all (you may need a SEE to elicit rationales for these responses, though). However, in most cases there will be a finite number of possible valid responses, but that number will not be known a *priori*—we contend that saturation is the only way to find out[24]. An additional point here in favour of the SGRP in comparison with traditional methods is that in the SGRP the researcher is explicitly considering the characteristics of the stimuli, and attempting to make them representative of the population of possible responses, whereas in the traditional approach the researchers are hostages to fortune in this respect and, for example, in relatively small samples of homogeneous judges it is likely that a restricted range of responses such as rationales will be tendered.

Are we not deceiving the participants in both phases of the SGRP?

Ideally, panelists recruited for an SEE should be informed that they are part of a nominal group in the same way they would be in a 'real' structured-group technique (SGT) exercise to ensure a similar level of motivation and expectation that their responses will be read and used by others (as they will be). If SEE participants are not to receive feedback and have subsequent responses collected, then it might well suffice to inform them that they are simply providing information for others rather than being in a nominal group *per se*: Whether this reduces the quality of their responses is an empirical question, though it is difficult to see why this should have a devastating effect on what they produce. Even if

[24]Incidentally, it is worth noting that a small number of different responses can be combined into a large number of unique groups (i.e., where each different response is included just once per group). For example, although 10 different responses only give 210 unique groups of 4 experts, and 45 groups of 8 experts, if you increase the number of responses to 30 then the number of unique groups is 27,405 and 5,852,925 for 4 and 8 member groups, respectively. See https://stattrek.com/probability/combinations-permutations.aspx.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

the experimenter wished to take no chances and initially suggest that SEE participants *would* receive feedback in a Delphi group, and then subsequently tell them they would not, full debriefing at the end of the exercise would likely be a sufficient ethical response, with mild deception not generally considered problematic in psychological studies. This also goes for participants in the main SGRP study (phase 2), although it is likely that their instructions concerning receiving information from nominal others can be framed in an accurate manner in all but the most contrived scenarios (i.e., where feedback is wholly or largely created *de novo* rather than being based on the actual responses of SEE participants).

How relevant is the SGRP for understanding opinion change in groups beyond Delphi and JAS?

The mapping of the second phase of the SGRP onto a prototypic 2-round Delphi or JAS protocol should be clear, and we have indicated how additional rounds of a Delphi could potentially be accommodated; however, the relevance to understanding opinion change in other SGTs —or groups more generally— may not be so apparent.

The main way in which Delphi and JASs deviate from the other SGTs and natural groups is that they more strictly control the nature of the interaction between group members. For example, the so-called NGT is like Delphi in that it requires participants to initially state their opinions individually, but then these opinions are fed back and discussed in a facilitated face-to-face session, before a final aggregation of opinion with equal weighting. Investigate Discuss Estimate Aggregate (IDEA) is very similar to the NGT except that the discussion is usually performed online and may not be completely free[25] (i.e., in IDEA's original formulation its main purpose is simply to ensure that all participants have a similar understanding of the problem)—the final aggregation is weighted by performance of each participating expert on a pre-test. Some forms of Delphi, such as real-time Delphi (RTD: Gordon & Pease, 2006) also have less strictly controlled interaction between participants than classic Delphi.

Allowing freer interaction raises the possibility of biases returning to the process. For example, some people might dominate the discussion if facilitation is poor in the NGT or IDEA, or because of early and/or repeated entry in RTD. On the other hand, some complex problems, or where there are a few highly expert panelists whose opinion is divided between several options, might be difficult to resolve without free information exchange, so the possible costs in terms of bias risk may be compensated by benefits in terms of efficiency and positive 'user experiences' (see Nyberg et al., 2021 for an example). For simpler problems, or with many less-expert experts, limiting the interaction may be the most efficient and satisfactory approach. What is 'best' will likely depend on an interaction between characteristics of the experts and the task, as well as the goals of the exercise—application of the modeling approach presented here in conjunction with the SGRP should help match elicitation protocols to experts and tasks.

Two ways in which 'direct' interactions—either in a 'loose' SGT or a traditional group—might differ from the controlled feedback of Delphi include: the formation of coalitions between group members, and the creation of different roles and/or hierarchies (e.g., leaders and followers). As we discussed in Introduction, these are both potential sources of bias and thus the reason for being excluded from Delphi. However, we do not know how biasing these things actually are in practice, nor if their effect is always negative. For example, perhaps in some circumstances emergence of a strong leader can help break a deadlock and facilitate better outcomes? Or a minority voice might become stronger on finding an ally and thus overturn a mistaken majority? We believe that the SGRP could help answer these questions by enabling the effects of specific features of interactions between group members to be investigated. For instance, coalitions of similar opinions can be explicitly created and compared against no explicit coalitions, or no possible coalitions (i.e., a set of disparate viewpoints). Similarly, roles can be allocated to both real and simulated participants and compared with conditions with different role allocations or no role allocations. Although such manipulations may also be possible with real interacting groups by, for instance, using confederates, or preselecting group members on the basis

[25] IDEA is a relatively new technique, but has already been implemented somewhat differently in its few extant applications.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

of particular traits or opinions, it is likely to be much easier and more efficient (e.g., require fewer participants) with the SGRP.

Slowing down or reducing the amount or quality of information exchange as a result of restricting the interaction in structured groups, might also have negative effects on performance—these can also be investigated in the SGRP. For instance, written presentation of reasons could be compared to spoken, graphical, or various combinations thereof, while quantity and quality can also readily be manipulated, as we have already discussed above. Naturalistic exchanges such as conversations between experts would be harder to simulate, but not impossible. For instance, interchanges between pairs of experts could be captured in the SEE and then presented during the main study, or participants could be allowed to select from a menu of question topics with pre-scripted answers also taken from an SEE. Performance could then be compared to that of a few real groups, thereby permitting the sources of any detriment in the simulated versions to be readily identified by virtue of the small remaining differences between the simulated and naturalistic conditions. In this way the SGRP can also shed light on processes of opinion change in directly interacting groups.

Finally, we wish to briefly discuss how the SGRP might be used to investigate prediction markets (PMs) - as we previously noted, interest in PMs as an SGT is increasing in recent years, Since PMs typically have very restricted interaction between participants they are quite well suited to investigation by the SGRP. Further, PMs usually have many contributors—in fact they need a large number to be effective—and their identities and individual opinions will be unknown, so there is anonymity. However, unlike Delphi there is no requirement for participants to consider and then express their opinions privately before looking at the market prices, which are an aggregate of all opinions expressed so far, so, in contrast to a classic Delphi, when participants join, they already can see what all previous entrants have said (in aggregate form). There is also no explicit mechanism for revision although there is nothing to stop a participant entering repeatedly by making multiple trades in the market on the basis of new information (or changes in prices). PMs are therefore potentially subject to a source of bias deliberately excluded from Delphi, namely the possibility of being led by the opinions of early movers in an information cascade, and this—combined with the very limited information exchange possible (the prices) and their competitive rather than cooperative nature[26]—could be behind some notable failures of PMs (e.g., Brexit and the 2016 US presidential election) (e.g., see Gelman & Rothschild, 2016; Levingston, 2016; Economist, 2016).

PMs are therefore like a Delphi where aggregate information is fed back (e.g., mean and range of prices) and nothing else. Since PMs usually have hundreds or thousands of participants, it would be difficult for participants to process information from individual traders (even if there was an incentive for useful information to be tendered), but it is possible to imagine a PM where enriched feedback is given, for example, the variation in prices, or a summary of rationales for particular 'unusual' prices. The SGRP could be used to investigate whether such innovations improve the markets. While PMs are fairly easy to recruit to in sufficient numbers to operate reliably—being 'fun' and mostly without the requirement for participants to have particular expertise—they lack the control needed to get to the nitty gritty of what works best and why. The SGRP permits the manipulation of important aspects of traders (e.g., expertise, confidence, competitiveness), markets (e.g., number of options, predictability, consensus), and platforms (e.g., amount, type, and rapidity of feedback) in experimental designs that have sufficient power to answer research questions with hundreds rather than thousands of participants in the main study. But do we not need to run thousands in the SEE in order to get realistic data for feedback in the main study? Not necessarily. While one option would certainly be to use data from real PMs—either previously run for analogous markets or one specifically run for the SEE—an alternative could be to build models of markets of various kinds and then use these models to generate simulated

[26]These properties are in contrast to, for example, traditional Delphi, where, if a panelist believes his or her judgment to be correct he or she tries to persuade the other panelists of this with good arguments. In PMs there is neither incentive nor opportunity to persuade others that they have the price wrong—in fact you sometimes stand to gain from *not* stating your true price, for example, pushing the price up then short selling (easier to do than for a traditional market since you usually know when a PM will be resolved).

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

trades, etc. The models could be validated using historic data and then used to produce plausible data for feedback in SGRP main studies.

Appendix 2: Stimulus materials for group size and diversity study

Nineteen questions were pre-tested for spread of answers and quality of rationales given. The following 10 questions were chosen for use in the study.

In the next month:

1. Kim Jong Un will publicly announce that North Korea will give up its nuclear weapons.
2. The UK Prime Minister Theresa May will face a vote of 'no confidence' in the House of Commons.
3. The personal details of over 1,000 customers will be stolen from a UK bank or financial-service provider.
4. The UK will expel one or more Russian diplomats.
5. A category 5 hurricane (most dangerous) will make landfall on the US mainland.
6. China and the US will begin trade talks designed to de-escalate the trade war.
7. An African country will announce that it is prepared to set up a processing centre for migrants to the EU.
8. Russian President Vladimir Putin will send troops into territory belonging to another country (other than Syria).
9. The Israeli army will kill more than 10 Palestinians.
10. Twenty or more people will die from ebola in Africa.

Appendix 3: Example rationales given as feedback to participants

Q8. Russian President Vladimir Putin will send troops into territory belonging to another country *(other than Syria)*.

[Rationales are grouped into 4 different types A–D for each forecast, 'Yes' or 'No'—only rationales for 'Yes' responses are shown here].

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

"Yes"	A

1. The invasion of Ukraine has happened with no severe repercussions for Russia. Putin believes that the EU and UN are impotent and unlikely to take any real action if he occupies new territory around the fringes of Russia.
2. Past and ongoing approval by Western countries of use of troops in Ukraine mean that Putin is likely to think he can get away with it again.
3. History of troops being sent into Georgia and Ukraine without causing any real problems for Putin.
4. Russia sees the West as weak and so I think Putin believes that he can repeat past behaviours such as the movement of troops into Ukraine.
5. Russia has done it before in Georgia, Ukraine and the Crimea so it could easily do it again. I think an attack of this kind is more likely than not.
6. It is not unlikely, given the pattern of Putin's past transgressions such a move would be logical.
7. Putin has shown himself willing to take such action in the past and the EU has done very little to suggest that it would intervene in any major way.
8. The West has allowed Putin to interfere in Ukraine and Crimea in recent years and so it is likely that he will try something similar in the near future.
9. Russia has managed to invade other countries' territory (e.g. Ukraine) in the past without any difficulty and so such a move would make sense.
10. History suggests Putin will continue to expand Russian borders as long as he keeps getting away with it.
11. Ukraine, Crimea show a pattern of behaviour which I believe will continue, given that the World Cup is now over.
12. Putin has done this consistently in the past and I think this autumn may well see another act of aggression from Russia.

https://doi.org/10.1017/jdm.2023.22 Published online by Cambridge University Press

B

1. I think it is highly possible that Russia will invade the Baltic States within the next month. Russia already has many thousands of troops stationed in Belarus, and the outlook is not good.
2. The Russian military has been steadily building up "formidable conventional forces" along the Russian border with NATO countries in recent months.
3. There are reports in the press that Putin has recently sent more troops to Russia's borders with the Baltic States.
4. It's reported that Russia has been building up its military base in Kaliningrad, which borders the NATO states Lithuania and Poland.

C

1. Putin has repeatedly stated that he wants to reunite the old Soviet bloc. I believe he will continue to pursue this goal by invading other former Soviet countries.
2. It has been predicted that Putin will continue his vast game of chess with NATO with his next move being to invade the Baltic nations of Estonia, Latvia, and Lithuania, which were previously part of the USSR.
3. Putin's ultimate goal is to reclaim as much of the original Soviet bloc from NATO as he can without provoking a major conflict.
4. Putin wants to re-instate the might of Soviet Russia and his plan is to take back territory from countries that were formerly under Russian control, e.g., the Baltic nations.

D

1. Putin is a soldier before a leader, I think he is aggressive at heart and so this is more likely than not.
2. He is a big player and likes to be seen to "throw his weight about" in an aggressive way.
3. I believe Putin is an aggressive leader who uses actions of this kind to maintain his domestic power and prestige.

Cite this article: Bolger, F., Rowe, G., Hamlin, I., Belton, I., Crawford, M., Sissons, A., Taylor Browne Lūka, C., Vasilichi, A., and Wright, G. (2023). Virtuous opinion change in structured groups. *Judgment and Decision Making*, e25. https://doi.org/10.1017/jdm.2023.22