Title: Improving the practical application of the Delphi method in group-based judgement: a six-step prescription for a well-founded and defensible process

Authors

Ian Belton[1*], Alice MacDonald[1], George Wright[1], Iain Hamlin[1]

Keywords

Delphi, review, methodology, judgement, expert, consensus

Abstract

This paper provides a practical, systematic approach to the design and delivery of a Delphi survey. We prescribe a sequence of six steps to do with (i) setting up the Delphi process – including selecting respondents and generating a requisite number of focal issues, (ii) software/delivery choice, (iii) developing question items and response scales, (iv) providing feedback between a requisite number of Delphi rounds, (v) preventing and dealing with panellist drop out, and (vi) analysing and presenting the Delphi yield. At each step, the Delphi administrator has a range of choice options and we provide discussion of the pros and cons of each option - in order that the overall design and delivery of a particular Delphi survey is both well-founded and defensible.

Introduction

The Delphi technique (Linstone & Turoff, 1975) has been widely cited in the literature on effective group-based judgment and decision-making, particularly in the domains of healthcare and quality-indicator development (Boulkedid, Abdoul, Loustau, Sibony, & Alberti, 2011; Hasson, Keeney, & McKenna, 2000) as providing a more useful process than traditional group meetings (Rowe & Wright, 2001). The Delphi process is designed to improve a group's access to multiple interpretations and views on a given topic of discussion, while also aiming to suppress more negative features of group discussions such as domineering individuals and opinions, which can undermine the effectiveness of these discussions (Grime & Wright, 2016; Rowe & Wright, 2001). The Delphi procedure is defined by four basic principles: anonymity, iteration, controlled feedback of responses to all group members, and statistical aggregation of individuals' responses (Rowe & Wright, 2001). In its most basic form, a Delphi procedure involves instructing a group of individuals to respond anonymously to a sequence of questions about a single quantitative estimate – such as the probability of an event occurring or the timing of when it will occur. A facilitator then combines, or aggregates, the responses into a statistical summary of the group response (e.g. median), sometimes along with the reasons given for the responses (Rowe and Wright, 2001). Individuals are then invited to submit a revised response, after considering the variety of responses received, or to resubmit their first response. This iteration and controlled feedback process continues over various 'rounds' until a consistent pattern of responses is reached, for example in the form of an obvious consensus (i.e., a general agreement within the group) or a notable dissensus.

[1] University of Strathclyde

[*] Corresponding author

The Delphi procedure has been applied repeatedly across many diverse domains – encompassing judgmental forecasting and policy-focussed decision making – and there have been reviews of the literature in various contexts, including, for example, the evaluation of healthcare quality (e.g. Boulkedid et al., 2011; Toma & Picioreanu, 2016), public policy (de Loë, Melnychuk, Murray, & Plummer, 2016) nursing (e.g. Foth et al., 2016; Keeney, Hasson & McKenna, 2001) and transport (Melander, 2018). However, there is a striking lack of consistency in what people mean by Delphi, how they choose to run Delphi studies, and how they report those studies (Humphrey-Murto & de Wit, 2019; Wright & Rowe, 2011).

Several review papers have discussed the use of certain methods or design features in the Delphi process, for example the procedure to be followed at each round of Delphi, as well as suggestions for specific features of the process, such as number of 'experts' to be used (e.g. Rowe & Wright, 2001), what constitutes an 'expert' (e.g. Devaney & Henchion, 2018; Hasson et al., 2000), how to deal with participant attrition over the multiple rounds of a Delphi application (e.g. Hsu & Sandford, 2007), what sorts of questions should be asked during the process (e.g. Toma & Picioreanu, 2016) and the type of response scale to use (e.g. McMillan, King, & Tully, 2016). However, these papers do not explicitly describe the main steps taken to complete the Delphi procedure or the most appropriate order for these steps. Two exceptions that do provide an outline of the workflow involved in a Delphi procedure are focused on concerns relevant to healthcare research (Toma and Picioreanu, 2016; Trevelyan & Robinson, 2015). In this paper, we propose a detailed and generalised template for a progressive 6-step process (see Table 1) to aid the completion of a successful Delphi procedure in any context and provide illustrative examples from the literature to support the design features and procedural specifics suggested. At many of the steps, the Delphi practitioner must make a judgment call on an aspect of their particular Delphi survey design and we provide discussion of the pros and cons of each option at each step. While we consider the order of steps given in the paper to be the most appropriate in general, we recognise that the steps do not all necessarily have to be carried out in the order described. In some cases, several of the steps can potentially be carried out in parallel (e.g. topic generation, software testing and expert selection).

The Six Steps of Delphi

Table 1: 6-Step Delphi Application

The Six Steps of a successful Delphi application

Step 1: Setting up a Delphi process

- Determine the overall goals of the exercise
- Choice of experts
    - Use heterogeneous experts
    - At a minimum, use 5-20 experts
    - Consider using an expert nomination process
- Initial Considerations
    - Geographical dispersion of experts
    - Relative expense and time-demand of Delphi versus alternatives
    - Severity of disagreements amongst experts
- The survey

- Generate issues for consideration in the Delphi survey by open-ended questioning
- Write an introduction and a closure
- Restrict the survey to what can be answered in 30 minutes
- Estimate the Delphi time-line
- Pilot the survey

Step 2: Developing question items and response scales

- Decide on the number of issues to explore
- Creating questions
  - Formulate clear, concise questions, and group by issue explored
  - Start with simple questions
  - Are the capabilities of the panellists matched to the questions posed?
- Formulate clear response formats
  - When taking measurements, choose between categorical, ordinal or interval scales
  - If using ordinal, Likert-type scaling, decide on even or odd number of response categories
  - Define the end points of the response category

Step 3: Software delivery choice

- Paper and pencil?
- E-mail?
- Web-based?
- Tailor-made or off-the-shelf?
- Conventional Delphi rounds or real-time?

Step 4: Providing feedback to panellists

- Provide median responses
- Provide either range or inter-quartile range of responses for each question
- Elicit and utilise respondents' rationales for their numerical responses
- Remove indicators of the prevalence of majority opinions
- Develop and apply a criterion of consensus
- Continue polling until responses show stability, generally 3 rounds are enough
- Be alert to continuing dissensus in panellists' opinions

Step 5: Preventing and dealing with panellist drop-out

- Note that self-rated experts tend not to drop out

- Use social rewards
- Consider using financial rewards
- Use personal communications with panellists
- Note that the greater the number of rounds, the greater the degree of dropout

Step 6: Analysing and presenting the Delphi data

- Make use of descriptive statistics to describe the data
- Note that small sample sizes and non-random sampling limit statistical analyses
- Use non-parametric statistical analyses
- Make use of graphical representations of data
- Integrate Delphi results with knowledge of the broader picture provided by other, perhaps more quantitative, research

Step 1: Setting up a Delphi process

Determining the overall goals

First and foremost, when initially designing a Delphi process, you must consider what issues or topics you wish to include as the focal point of discussion in the Delphi process – what is it that you want to gain expert opinion on? This choice has an impact on the selection of expert panellists and the design of the survey targeting this chosen audience. This initial evaluation has been termed the 'Exploration' stage (Linstone & Turoff, 2002). Often, the starting point is a review of the relevant literature to confirm the issues that should be addressed by the Delphi process (e.g. Czaplicka-Kolarz et al., 2009; Li, Chen, & Kou, 2017; Mokkink et al., 2010; Novakowski & Wellar, 2008). After identifying the initial issues to be discussed or debated, it is standard practice on a Delphi procedure to 'pilot' these issues in order to gain an insight into how these issues are viewed by the potential panellists (e.g. Day & Bobeva, 2005; Hasson et al., 2000). Piloting also allows participants to gather their thoughts on the topic and bring anything else forward that they may wish to discuss as part of the Delphi process (Toma & Picioreanu, 2016; Hasson et al., 2000). This initial procedure is often conducted by distributing open-ended questionnaires to the selected experts or panellists (e.g. Day & Bobeva, 2005; Keeney et al., 2006; Hasson et al., 2000), or holding face-to-face workshops with a sub-set of panellists (e.g. Frewer et al., 2011; Landeta, Barrutia, & Lertxundi, 2011). This is a beneficial step in the Delphi process as it allows for the gathering of a substantial range of viewpoints on the topics that are to be evaluated. Open-ended questions can also bring to light important aspects of the topic that that have been missed by the research team (Toma & Picioreanu, 2016). Where it is particularly important to focus on specific issues and/or avoid opinion bias, a more structured questionnaire-based approach can be used, but the open-ended approach is more typical for this stage of the process (Toma & Picioreanu, 2016).

In larger-scale Delphi projects, identifying the topics to include in the survey can often be a complex multi-method, multi-stage process that combines, for example, bibliometric analysis, workshops, and scenario exercises, as well as extensive testing of the Delphi items. It is beyond the scope of this paper to discuss those processes in detail but illustrative examples include Cuhls, Blind and Grupp (2002),

Gheorghiu et al., 2017; Li, Chen and Kou (2017) and NISTEP (2009). Some studies report and name the exploratory step as Round 1 of a formal Delphi process (e.g. Geist, 2010; Hung et al., 2008; van der Steen et al., 2013) but this can create confusion when reporting and discussing the Delphi survey results – for example, a "three-round Delphi" would then contain two iterations of the survey question-answering.

Once the findings from the exploration phase have been obtained using one or more of the methods summarised above, the responses received can then be collated into a list of issues/items/questions to be included in the main Delphi survey. A good Delphi survey should include a clear, concise introduction and ending and we suggest that the content should be restricted to what can reasonably be answered in 30 minutes. We discuss how to structure the questions in Step 3, below. We recommend piloting the main Delphi survey before sending it out to the panel (c.f., Gordon, 2003; Mitchell, 1991) by sending an initially designed survey to a small sub-set of the larger panel.

Selecting your experts

The appropriate selection of panellists is a key stage in the design of any Delphi (Hsu & Sandford, 2007). As a Delphi study involves gathering expert opinions on a particular topic, the recruitment of panellists is usually informed by the level of their domain knowledge in the relevant topic area (Hsu & Sandford, 2007; Hasson et al., 2000), hence the term 'expert' panellists (McKenna, 1994). There is no "magic formula" for expert selection (Keeney, Hasson & McKenna, 2006, p.209) and panels tend to be purposive or convenience samples rather than representative, random samples from particular populations (de Loë et al., 2016; Devaney & Henchion, 2018). In general, researchers should use their common sense to identify sensible selection criteria for respondent selection that are likely to satisfy the study's intended target audience. Typical criteria used in Delphi research include having a specified number of relevant academic publications (e.g. Hallowell & Gambatese, 2010; Mokkink et al., 2010), professional experience/activity in the field of interest (e.g. Cuhls et al., 2002; Morrison & Barratt, 2010; Toma & Picioreanu, 2016) and/or membership of relevant organisations/institutions (e.g. Czaplicka-Kolarz, Stańczyk, & Kapusta, 2009; Melnyk, Lummus, Vokurka, Burns, & Sandor, 2009; NISTEP, 2009). Devaney and Henchion (2018) used a conceptual continuum of 'closeness' to the topic of interest (the Irish bioeconomy) to identify not only 'traditional' experts such as academics but also stakeholders with 'subjective closeness' to the issues such as producers and representative bodies. A final consideration when recruiting experts is their likely level of engagement with the topic. Experts' degree of interest in answering the questions in the Delphi survey predicts both initial response rate and subsequent drop-out rate (Hasson et al., 2000).

We recommend that if the potential pool of Delphi respondents is large, for example the heads of University business schools, then the professional organizations to which individuals belong should be asked to nominate Delphi participants (for example, the UK Association of Business Schools or the European Foundation for Management Development). Nominated representatives are less likely to drop-out of the multiple round Delphi process because of the official nomination, especially if the name of the participant is included in the final write-up of the Delphi results report (see Step 5, below)

In terms of the number of expert panellists that are required, many ranges have been suggested: 5-20 (Rowe & Wright, 2001), 15-60 (Hasson et al., 2000), no more than 50 (Toma & Picioreanu, 2016; Witkin & Altschuld, 1995), or 15-30 for homogenous Delphi panels (Clayton, 1997) and 5-10 for heterogeneous panels (Delbecq, Van de Ven, & Gustafson, 1975). Empirical research suggests that the lower end of these ranges may be adequate (Boje & Murnighan, 1982; Belton et al., 2019; Brockhoff, 1975). In practice, the number of expert panel members recruited for Delphi studies varies widely and

is essentially informed by the specific aims and intricacies of the study such as the topic of concern and the area and level and range of expertise to be solicited (Clayton, 1997). In a recent review of 63 Delphi studies, de Loë et al. (2016) reported a huge range of panel sizes, from fewer than 10 up to more than 1000. In the field of technology foresight, national-level surveys can involve many thousands of expert panellists, for example in Japan (2900 participants – NISTEP, 2009), The Republic of Korea (5450 participants – Choi & Choi, 2017) and China (nearly 3000 – Li et al., 2017).

Other Delphi design issues also impact on the group size that is likely to be appropriate. For example, if it is proposed to feed back panellists' rationales (or a selection thereof) after each round as well as statistical summaries, this may be a limiting factor since the volume of material for participants to review after each round can quickly become unmanageable. In an experimental Delphi study involving yes/no geopolitical forecasting questions, Belton et al. (2019) found some (limited) evidence for a reduction in performance of individuals who had to review 18 single-sentence rationales as well as their associated forecasts, compared to those reviewing only 6, 9, 12 or 15 rationales. Automated methods for filtering or organising rationales such as Dynamic Argumentative Delphi (DAD – Gheorghiu et al., 2017 – see Step 3, later, for more detail) can help address this issue. Drop-out/attrition rates will be linked to panellist task demands, as demonstrated by Franklin and Hart (2007) - we will explore this issue in detail under Step 5, later.

It is generally desirable to use a heterogeneous panel of experts rather than a homogenous panel. One of the 'bestselling' characteristics of the Delphi method is that because it involves a group of panellists interacting and making decisions about particular topics or issues, the potential for bias inherent in a single opinion becomes less of a threat (Rowe & Wright, 2001, 2011). In more heterogeneous samples, participants will have more varied opinions and experience of the topic or issues in question, and are therefore more likely to represent the variety of perspectives that exist on a particular topic and achieve more accurate and plausible judgements (Bolger & Wright, 2011; Rowe & Wright, 2001; Spickermann, Zimmermann, & von der Gracht, 2014). Further, empirical research suggests that opinion diversity can improve the accuracy of the Delphi yield (Belton et al., 2019; Hussler, Muller, & Rondé, 2011). Therefore, we would strongly recommend the use of heterogeneous groups during a Delphi. Diversity can be encouraged by selecting experts who differ on a set of relevant criteria such as sector (academic, industry or government), field of expertise and/or demographics (e.g. Cuhls et al., 2002; Gheorghiu et al., 2017). But, be aware that the use of heterogeneous panel membership (for example, care-home managers and, also, care home residents) is, perhaps, likely to result in continuing dissensus over Delphi rounds (for example, views on the quality of care-home care). We return to this topic in Step 6, below. If the Delphi panel is viewed as too homogeneous then it is possible to induce heterogeneity of viewpoints in the initial stages of Delphi using adversarial techniques such as role-playing, devil's advocacy or dialectical enquiry (Bolger & Wright, 2011).

Practicalities of running a Delphi

It is important to consider the practicalities of a Delphi before embarking on the process. It is often the case that the expert panellists are situated in different locations, and so being able to meet physically to discuss the topic of concern can be difficult. One great benefit of a Delphi process is that it can be delivered mainly online or involve over-the-internet communication channels, and so participants in different locations can easily take part and communicate with each other (Boulkedid et al., 2011). A Delphi can take a substantial amount of time to run (e.g. weeks to months, depending on the scale of the survey and the degree of automation in its administration – McMillan et al., 2016; NISTEP, 2009). In some cases, the number of iterations or rounds to achieve opinion stability (i.e., consistency of responses between successive rounds may be greater than anticipated (Dajani, Sincoff, & Talley, 1979),

therefore requiring an unanticipated continuing time commitment from participants, which can increase drop-out rates (Hsu & Sandford, 2007). Additionally, using participants over a number of rounds will increase the expense of the study relative to 'one-off' group decision-making techniques, such as Nominal Group Technique (NGT). However, because a Delphi questionnaire can be administered over the internet, costs such as participant travel expenses will be zero, therefore making Delphi a more financially viable method to employ in comparison to an NGT method for example (Hsu & Sandford, 2007; McMillan et al., 2016). We discuss the administering of Delphi surveys further under Step 2.

Another issue to bear in mind is that although the objective of Delphi is to achieve stability of responses from a group of experts in a particular field to reach a more accurate judgement, this stability of responses may not be in the form of a group consensus; it may alternatively be in the form of a continuous dissensus or disagreement (de Loë et al., 2016; Rowe & Wright, 2001). Panellists may provide varying but, at the same time, stable responses or opinions to each other during Delphi iterations, therefore never resulting in a convergence of opinion (Rowe & Wright, 2001). Alternatively, stability itself may not be achievable (Landeta & Barrutia, 2011). The risk of asking participants to provide opposing responses on a topic over and over again should be carefully assessed before commencing a Delphi study, as this could undermine the process by causing drop-out or attrition amongst continually disagreeing respondents. However, the Delphi group leader, or 'facilitator', has autonomy to choose not to hold another round if disagreements still remain after a few iterations (Rowe & Wright, 2001). Note also that attrition rates may influence the types and consistency of responses obtained and agreement produced over subsequent Delphi rounds regardless (Rowe & Wright, 2001; Toma and Picioreanu, 2016) – for example, if those panellists with extreme viewpoints decide, after round 1 feedback, to drop out of the Delphi process. We discuss how to deal with attrition rates in more detail in Step 5 below.

A research team proposing to run a Delphi exercise needs to be well-organised and efficient and careful project planning is required to ensure the smooth running of the process (Hasson et al., 2000). An important element of planning for Delphi is determining a realistic time-line for the process. The time-line should allow for recruiting experts, developing and testing the survey, and distributing, collecting and analysing multiple Delphi rounds. Experts must be given a reasonable period to respond to each round and the collation of responses between rounds can also be time-consuming, depending on the format chosen for responses and level of feedback provided after each round. Keeney et al. (2006) note that each round can take up to eight weeks to complete in practice; Gordon (2003) suggests 3 weeks between rounds while Eggers and Jones (1998) propose 4 weeks.

Additionally, researchers must ensure that all data collected as part of a Delphi exercise is handled in accordance with relevant data protection legislation. In the EU, the General Data Protection Regulation (GDPR) imposes various restrictions regarding where participants' data can be stored and how it must be managed.

In summary, our recommendations for Step 1 are:

- Confirm that a Delphi process is definitely the best option given the topic of interest, your research goals, your financial and logistical resources, and the time available
- Consider using an expert nomination system to identify an appropriate group of experts for the panel
- Use at least 5 experts; between 5 and 20 may be sufficient but using more is acceptable if it is practicable to do so

- Select a heterogeneous group of experts or, if this is not possible, consider inducing heterogeneous opinions using adversarial techniques
- Generate issues for consideration in the Delphi survey using open-ended questioning
- Use basic language when constructing items in Delphi surveys to help reduce any language barriers or lack of understanding
- Include a clear introduction and ending and restrict the questions to an amount that can reasonably be answered in 30 minutes
- Pilot the survey
- Make sure you have a clear plan for the whole Delphi exercise, including a time-line; the process may take longer than you anticipate

Step 2: Developing question items and response scales

The language used to formulate the questions to be asked in a Delphi study, and the response options available to panellists, are features which must be carefully deliberated, with close reference and consideration to the overall aims and hypotheses of the study.

First, we would recommend that you are clear in the issues or sub-topics of the overarching topic or question you wish to explore within your Delphi group; the choice of issues will naturally be influenced by the results of the exploration phase – discussed in Step 1 – as well as the researcher resources available and the researcher expertise in the topic area considered (Hasson et al., 2000). The type of questionnaire and the number of issues that researchers wish to explore in a Delphi study will also depend on how extensively the topic or issue has been considered in previous literature or how much it has already been the focus of public debate and opinion. Sometimes, the area of interest will be relatively focused and so a limited number of items will be sufficient (e.g. 10-15 – Agell et al., 2015; Hallowell & Gambatese, 2010). In other cases, particularly where there has been relatively little research in the field or where the survey needs to be exhaustive in scope, many more questions may be needed (e.g. 50-85 topics, 7 questions on each topic – NISTEP, 2009). The only real limit is on the time and effort that panellists can reasonably be expected to spend completing the survey. If including questions on many topics, break them into sub-topics and present them in a clear and logical sequence, with questions thought to be easiest to understand and answer placed at the start of the sequence.

With regard to formulating the specific questions, poorly worded questions can induce bias of various kinds (Hallowell & Gambatese, 2010; Loveridge, 2002; Winkler & Moser, 2016). For example, consider the positive framing of the following question: "By how much should personal income tax be increased to provide community support for the aged to continue living in their own homes?" Note that the option of decreasing income tax is not given as explicitly available. Furthermore, employing surveys that are long and intricate can also affect response rates (e.g. Frewer et al., 2011) and may even cause false dissensus (Salancik, Wenger, & Helfer, 1971; Scheibe, Skutsch, & Schofer, 1975). Obtaining qualitative responses to an initial set of open-ended questions can help researchers to subsequently develop suitably structured questions that will elicit well-focussed quantitative and qualitative responses from participants in later Delphi rounds (Frewer et al., 2011). Moreover, using basic phrasing when constructing questions for Delphi surveys will help to overcome any language barriers when working with panellists who have been internationally sourced (Frewer et al., 2011; Rowe & Wright, 2001, 2011). Alternatively, a multi-language approach can be attempted but this creates other difficulties (Frewer et al., 2011).

Typically, a Delphi survey will be based around a set of questions that require numerical responses. The most commonly used type of response scale in Delphi study surveys is a rank-ordered or Likert-type scale (e.g. Aengenheyster et al., 2017; Elwyn et al., 2006; Makkonen, Hujala, & Uusivuori, 2016; Ogden, Culp, Villamaria, & Ball, 2016). Depending on the type of questions to be asked and the depth, or grain, of responses required, we would recommend employing either an ordinal response scale (i.e., a rank-ordered response scale, e.g. not important to most important), a categorical response scale (i.e., utilising different categories, e.g., "yes", "no", or "don't know"), or an interval response scale (i.e., a numerical scale, e.g., a confidence estimate – 0%, 5-10%, 10-20%, …).

Once you have selected the most appropriate response scale type, the next step is to decide on the number of response options (Toma & Picioreanu, 2016). More response options promote more fine-grained measurement; too many options risk increasing random error owing to the potential failure of respondents to interpret the subtle differences in meaning between options. Most Delphi studies use scales with either 5, 7 or 9 response options. Generally, previous research suggests that 7-point scales are the most reliable (Cicchetti, Showalter, & Tyrer, 1985; Preston & Colman, 2000; Weng, 2004). Many researchers include an even number of response options on agreement scales, so as to provide participants with a mid-point reflecting a neutral level of agreement. Although mid-points are sometimes the favoured option among those who are not motivated to think carefully about which option to choose, they are necessary to accommodate those respondents who have well-considered, genuinely neutral opinions. Depriving respondents of an option that accurately indexes their true opinion will ultimately reduce the measurement accuracy of the item (Furr, 2011).

When constructing scales, researchers should also consider whether to provide verbal labels for all response options or only for those at the poles. We recommend providing verbal labels for all options, because doing so produces scales of better psychometric quality (Krosnick, Judd, & Wittenbrink, 2005). Labelling all response options has been found to reduce mis-responding (Weijters, Cabooter, & Schillewaert, 2010), presumably because of the resultant reduction in cognitive load and increased clarity regarding the meaning of the response categories (Krosnick, 1991; Swain, Weathers, & Niedrich, 2008).

As well as collecting numerical responses, researchers may also choose to allow panellists to provide written rationales in support of their responses (discussed further in Step 4) and/or include open response fields for additional comments or questions (e.g. APEC Center for Technology Foresight, 2010; Cuhls et al., 2002; Gheorghiu et al., 2017). If providing rationales, participants must be warned not to disclose information that would compromise their anonymity (such as, "As a Professor of Epidemiology at the University of London, my view is that…").

Overall, this step in the Delphi process is particularly important. The structure and format of questionnaires directly influence the nature of the responses obtained from panellists and so should be carefully considered when designing a Delphi process and its materials. Survey designers are advised to recognise the importance and complexity of designing ad hoc scales, and seek further information from specialist resources, such as Furr (2011) and Krosnick and Presser (2010). Our recommendations for Step 2 are:

- Carefully formulate the Delphi questions to be of interest to the focal respondents, using neutral language and expression
- Where possible, and depending on the level of responses required, use Likert-type scales
- 7-point Likert-type scales allow fine-grained measurement without overwhelming respondents with too many response options

- Consider including a neutral or "I don't know" option
- Consider providing verbal labels for all response items

Step 3: Software delivery choice

Once the Delphi survey has been designed, it is important to consider *how* to administer the survey (e.g. pencil-paper task, email distribution, or via an online tool) and the type of Delphi procedure to use (e.g. conventional, modified or real-time – Aengenheyster et al., 2017).

A conventional Delphi procedure has traditionally been administered in 'pen-paper' style (Hasson et al., 2000; Toma & Picioreanu, 2016). Until the early 1990's, Delphi questionnaires were invariably delivered by post to panellists' addresses, which subsequently induced very low response rates and high attrition (Mitchell, 1991). The traditional approach was also slow, required participants to have an acceptable writing style, and necessitated a substantial quantity of sorting and filing of paperwork for the research team (Hasson et al., 2000), especially when collecting large numbers of surveys. Most Delphi practitioners now employ the use of electronic methods (e.g. online tools and/or offline documents delivered by email) to administer the surveys during a Delphi procedure (de Loë et al., 2016; Foth et al., 2016). Online and email-based methods are fast and efficient and can enhance participation rates and reduce the workload for facilitators/researchers considerably (Boulkedid et al., 2011; Gnatzy, Warth, von der Gracht, & Darkow, 2011; Toma & Picioreanu, 2016). Electronic options include email distribution of off-line (e.g. Word) surveys, manually facilitated online surveys using survey software such as Qualtrics (www.qualtrics.com) or SurveyMonkey (www.surveymonkey.com), or online tools that manage the whole Delphi process. Online tools can be custom-built for a specific project (e.g. Gheorghiu et al., 2017). Alternatively, several proprietary tools for online Delphi currently exist, including the Delphi Decision Aid (http://armstrong.wharton.upenn.edu/); Surveylet (www.calibrum.com); eDelphi (www.edelphi.org/); Mesydel (www.mesydel.com); and the Global Futures Intelligence System (www.millennium-project.org/projects/global-futures-intelligence-system/). (For a recent comparative review of several tools, see Aengenheyster et al., 2017). In general, off-the-shelf propriety tools are less customisable than tailor-made, individually developed software, although the latter are likely to be more costly and time-consuming. Also, if the focal Delphi survey is to be a one-off event, gaining sufficient command of a less-than-intuitive proprietary tool that has a relatively small number of users needs to be considered carefully.

The use of online survey administration also has its challenges, such as needing to have constant access to the internet and problems with saving survey progress (Toma & Picioreanu, 2016). In their paper regarding methodological considerations of a Delphi and suggestions for an improved approach to the procedure overall, Toma and Picioreanu (2016) suggested using an electronic questionnaire, sent by email for example, that does not require an internet connection to complete. In addition, online approaches may not always be more time- or cost-efficient. Geist (2010) compared 'pen-paper' and online approaches and found that while the cost of running the online survey was lower, the up-front expense of developing the online survey tool made this option more expensive overall (although the tool could be re-used in subsequent surveys with little additional expense). When choosing between electronic approaches, researchers should take account of the PC/IT skills that may be required for both participation in and management of online survey methods (Hasson et al., 2000). A simple online form may be more user-friendly – and therefore effective – in some contexts than a more complex, real-time tool, for example. In addition, sending emails to participants containing the survey is a much more personal and direct approach than instructing them to access a webpage (Boulkedid et al., 2011) and so could reduce drop-out rates for some panellists.

An important factor which affects how a Delphi questionnaire is administered is the type of Delphi procedure used. Traditionally, the Delphi method involves participants responding to a questionnaire or survey and subsequently receiving feedback over a series of distinct rounds which takes place over a number of weeks (Delbecq et al., 1975). It has been specifically recommended to give participants two to eight weeks to respond at each round (Delbecq et al., 1975; Keeney et al., 2006). A 'real-time' Delphi process (e.g. Gary & von der Gracht, 2015; Gnatzy, Warth, von der Gracht, & Darkow, 2011; Gordon & Pease, 2006) is a variation of the technique that involves the whole process taking place within a single "round-less" time window, rather than a series of time windows with "sync points" at the end of each round. Participants in a real-time Delphi receive feedback of those aggregated results that are contemporaneously available, immediately after they have made their own Delphi response (Aengenheyster et al., 2017) and can proceed through all Delphi rounds in one session if they wish. The benefit of this process is that it allows for more flexible participation, whereby participants do not all have to be available to respond within a set time period for each round and they can edit their responses numerous times before a given deadline (Aengenheyster et al., 2017).

Real-time participation has the potential to improve response rates, since participants can take part when it suits them during a single time-frame (Gary & von der Gracht, 2015; Gnatzy et al., 2011), rather than responding within a series of time windows, and do not have to wait for feedback until all panellists have completed a given round. This feature may be useful for people who have conflicting priorities such as work or family commitments; however, this procedure can only be conducted via the internet (Aengenheyster et al., 2017). In addition, a 'round-less' real-time approach means that the whole Delphi process can plausibly be completed much more quickly than using a round-based design. However, the iterative, round-by-round structure is a key characteristic of Delphi, which was purposely designed to minimise group bias (Rowe & Wright, 2001). Another issue for real-time Delphi is that if qualitative rationales are provided as well as numerical responses, the qualitative data can quickly become unwieldy and cannot easily be tidied by a facilitator without losing the real-time aspect. One proposed solution is DAD (Gheorghiu et al., 2017), in which supporting arguments (sampled prior to the main survey via a pilot survey) are, in the main survey, dynamically ranked based on the number of votes they receive from panellists who have participated up to that point. For a similar aggregation and collation approach to organising and summarising arguments within a real-time Delphi, see Airaksinen, Halinen, & Linturi, 2016. Early research suggests that using a real-time tool may not reduce the efficacy of the Delphi process (Gnatzy et al., 2011) but further empirical work is needed to confirm the convergent validity between real-time and traditional round-based Delphi methods.

Another type of Delphi that can be used is a modified Delphi procedure, which involves a physical meeting of panellists between rounds (Boulkedid et al., 2011). Boulkedid et al. (2011) reviewed a large number of Delphi studies in order to make concrete recommendations about how to apply this technique when selecting healthcare quality indicators. They found that over half of the studies they reviewed used a modified Delphi technique in that participants met at least once during the study procedure. They argued that although this process negates one of the crucial characteristics of Delphi – namely avoiding the presence of dominant voices when making decisions (Rowe & Wright, 2001) – the meeting between panel members could have a positive outcome, such as the more personable sharing of information and opinions. Boulkedid et al. (2011) therefore suggested that panel members could meet after the last Delphi round, if required, to clarify any misunderstandings of option or to explore the different options when a consensus has not been achieved, for example.

Overall, our recommendations for Step 3 are:

- Use either email and/or an appropriate online tool to administer Delphi surveys for ease of access for participants and for ease of editing and collating responses electronically
- Take care when determining the specifics of the approach to use, for example whether to circulate offline documents via email, or use one or other type of online survey tool

- Base your choice of procedure (e.g. traditional, modified or real-time) on relevant factors including the resources and capabilities of the researchers and the demographics and resources (e.g., quality and degree of internet access) of the proposed panellists
- If time is of the essence, a real-time approach might be appropriate or if the responses provided by a traditional Delphi process are insufficiently clear, a modified process involving face-to-face contact could prove useful

Step 4: Providing feedback to panellists

Providing feedback

The type of feedback given to panellists after the first round is a key design issue (Rowe & Wright, 2001; Boulkedid et al., 2011), since it has implications for how panellists respond to subsequent rounds and for the overall effectiveness of a Delphi in gaining a group consensus response or stability of responses. Since Delphi studies, for the sake of brevity and efficiency, often result in a substantial amount of response data being omitted from the data summary, feedback must be carefully considered and controlled (Hasson et al., 2000). The Delphi administrator must decide the most appropriate statistics to summarise and report responses. For numerical measures such as Likert-type scales, the reporting of central tendencies at the feedback stages of a Delphi study is beneficial as it allows panellists to see how their response compares to the group response as a whole (Hasson et al., 2000). Measures of central tendency can efficiently and effectively depict an aggregated response from a number of panellists (Hasson et al., 2000; Hsu & Sandford, 2007). The mean and standard deviation of panellists' responses to a particular Likert-style question item are sometimes used for feedback to panellists in Delphi applications but this is generally not considered an appropriate way of summarising ordinal data and should be avoided (Jamieson, 2004; Liddell & Kruschke, 2018; Norman, 2010). We recommend providing the median and inter-quartile range for responses made to individual ordinally-measured question items - this is common practice in Delphi studies and is well-supported in the academic Delphi literature (e.g. Hsu & Sandford, 2007; Rowe & Wright, 2001).

Allowing participants to provide explanations or rationales as to why they provided a particular response in a Delphi round – in other words obtaining additional qualitative data – can add to the quality and relevance of feedback (Bolger, Stranieri, Wright, & Yearwood, 2011; Meijering & Tobi, 2016; Rowe & Wright, 2001). Of course, adding qualitative feedback also creates challenges around how to aggregate such responses: facilitator aggregation may introduce researcher bias (Franklin & Hart, 2007), while attempting to maintain the integrity of individual responses can make later Delphi rounds unwieldy (Briedenhann & Butts, 2006; Green, Jones, Hughes, & Williams, 1999; see also the DAD method (Gheorghiu et al., 2017), discussed earlier in Step 3, for a proposed solution to this issue).

Although a desirable outcome of Delphi is typically for a group of panellists to reach consensus, the presence of a majority opinion or response *during* a Delphi process can have a less desirable effect on subsequent responses since the prevalence of majority opinion can, by itself, cause opinion change (Bolger et al., 2011; Rowe & Wright 2011). In their study exploring the effects of majority opinion and high levels of confidence on opinion change and response accuracy, Bolger et al. (2011) observed that panellists who were less confident in their response during a Delph process, including those who provided a minority response, were more prone to altering their responses than panellists who were more confident, and who provided a majority response. Therefore, making salient a majority opinion mid-Delphi-process can be problematic and should be avoided (Bolger et al., 2011). It is important to

remove any indicators of the prevalence of majority opinions and levels of confidence, in order to promote opinion change in response to compelling rationales rather than opinion change towards either the majority and/or towards highly confident views (Rowe & Wright, 2011).

The ideal number of rounds to hold during a Delphi procedure is also a topic of debate in the literature (Rowe & Wright, 2001). The main function of a Delphi study is to investigate whether panellists can reach a level of consensus regarding a particular topic or issue, and the Delphi literature suggests that panellists generally move towards a level of agreement, showing less diversity of opinion, over two or three rounds of Delphi (Erffmeyer, Erffmeyer, & Lane, 1986; Rowe & Wright, 2001). For example, in their systematic review of Delphi studies used to help identify healthcare quality indicators, Boulkedid et al. (2011) observed that consensus was typically achieved after two to three iterations. We recommend, however, that the appropriate number of rounds should ideally be determined by looking for a pattern of stability, i.e. once the panellists show a level of stability in their individual responses the Delphi procedure can cease (Rowe & Wright, 2001; von der Gracht, 2012). We deal below with the specifics of how to determine when a stable consensus has been reached. In cases where practical constraints (e.g. time, cost, expert availability) require a set number of Delphi rounds to be specified up front, we propose that three rounds will generally be enough to identify a level of consensus, or level of continuing dissensus (Rowe & Wright, 2001; Boulkedid et al., 2011).

Reaching consensus (where possible)

Given the importance for most Delphi studies of reaching a group consensus, it is essential for researchers to decide how they wish to define and operationalise that consensus. Ironically, although the issue has been much debated (e.g. Boulkedid et al., 2011; Hsu & Sandford, 2007; von der Gracht, 2012), there is no agreed position amongst Delphi researchers (Diamond et al., 2014; Mitchell, 1991). It is generally recognised that a panel's opinions must first be relatively stable before consensus can be meaningfully assessed (Dajani et al., 1979; Scheibe et al., 1975; von der Gracht, 2012). In line with von der Gracht (2012), we recommend measuring both stability and consensus on a round-by-round basis and continuing until acceptable levels of both are achieved (or a stable dissensus).

Frequently used consensus criteria include the same or similar opinion being reported by a pre-determined percentage of panellists, e.g. 75% (Keeney, 2000; Hasson, 2000) or 80% (Green et al., 1999; Toma & Picioreanu, 2016), particular levels of statistical dispersion as measured by inter-quartile ranges (Landeta, 2006; von der Gracht & Darkow, 2010), various inferential statistics (Brender, Ammenenwerth, Nykanen, & Talmon, 2006; Schmidt, 1997), or more complex mathematical approaches based on fuzzy theory (e.g. Agell et al., 2015; Duru, Bulut, & Yoshida, 2012). For reviews of the options available, see von der Gracht (2012) and Kalaian & Kasim (2012). The criteria for establishing consensus should ideally be specified a priori although this does not always happen: Diamond et al.'s (2014) review of 100 Delphi studies found that only 43 specified a threshold for consensus in advance. As regards measuring stability, options include a Chi square test for independence of responses (Dajani et al., 1979) or testing for changes in the coefficient of variation across rounds (von der Gracht, 2012).

Delphi researchers should give careful consideration to their choice of consensus criteria: stricter criteria will give the results greater validity but will make the measurement of consensus harder, if not impossible, to achieve – for example, if 100% agreement between panellists is made a requirement. Above all, researchers should ensure that the chosen approach provides a level of confidence in the outcome that is suited to the needs of the research topic (Keeney et al., 2006). In some cases, such as

where a critical decision must be made, very high levels of consensus may be required, whereas in other cases general opinion trends in one direction or another may be more appropriate (Keeney et al., 2006).

If there is continuing disagreement between panellists after a number of rounds, the group's facilitator should be able to make a judgement call as to whether they feel the panellists have been given a fair opportunity to explain the reasons underpinning their continuing dissensus (Rowe & Wright, 2001), whilst remaining sensitive to increasing the time commitment of participants and increasing potential drop-out rates due to the unresolved conflict of opinions (Rowe & Wright, 2001; e.g. Toma & Picioreanu, 2016). In some contexts, consensus may not be the only (or even the main) goal of a Delphi survey: it may be more important to understand whether consensus on a given topic is possible or appropriate and to stimulate debate. This is particularly the case in so-called "policy Delphis" that aim to identify views on policy alternatives (e.g., Cuhls, 2015; de Loë et al., 2016; Franklin & Hart, 2007) – for example, panellist viewpoints on alternative airport siting options.

In conclusion, when giving feedback to panellists in a Delphi procedure between rounds:

- We recommend providing a statistical summary of the responses received, the median values and inter-quartile ranges if possible (depending on the data collected and the measures used), whilst also allowing participants to provide qualitative explanations for their quantitative responses
- Choose a consensus criterion that is not too onerous, say 75%, and also test for stability
- Regarding the optimal number of iterations, we recommend that three rounds should generally be sufficient to allow a pattern of stability to emerge from panellists
- We would advise caution as regards explicitly revealing the prevalence of majority opinions during the Delphi process and emphasise the importance of being sensitive to consistent dissensus amongst participants, rather than consensus

Step 5: Preventing and dealing with panellist drop-out

A concern with any empirical study involving direct input from participants is the attrition rate and possible lack of participant engagement, particularly in studies requiring numerous stages of participant input over a long period of time. This is especially the case in a Delphi process, where participants have to provide input over a number of rounds or iterations (Boulkedid et al., 2011; Rowe & Wright, 2011). The Delphi administrator must be aware of the issues surrounding attrition and make it their goal to prevent participants withdrawing from the study at too early a stage, for example after receiving median responses and rationales from round 1.

The success of a Delphi procedure depends on the participants who agree to take part being able to commit enough of their time to the entire procedure (Hasson et al., 2000; Keeney et al., 2001) and to respond actively at each round (Goluchowicz & Blind, 2011). A large attrition rate can directly influence the level of consensus subsequently obtained within the group and therefore the legitimacy of the results, especially as those with dissenting views may be more likely to drop out (Humphrey-Murto & de Wit, 2019). We suggest using incentives (e.g. financial, social recognition) to help motivate participants to take part initially, and to help secure their involvement throughout the Delphi process. Toma and Picioreanu (2016) suggested using a choice of financial rewards (e.g. money/vouchers given at either the beginning or the end of the study procedure, participation in a prize lottery, or a donation to an organisation/charity of their choice) which participants could select as compensation for their participation and continued input. However, there is an ongoing debate regarding the value of financial

incentives for research participation (Göritz, 2010). For example, there is evidence that financial incentives can increase response rates and reduce drop-out rates (Frick, Bächtiger, & Reips, 2001; Göritz, 2008; Pforr et al., 2015) but they may also damage the quality of the responses obtained, for example by making participants less conscientious (Barge & Gehlbach, 2012; Heerwegh, 2006) or by creating a perception, at the Delphi write-up stage, that panellists' opinions have been 'bought' by the survey organisers.

Another way of potentially reducing the drop-out rate during a Delphi procedure is for the facilitators to employ personal communication methods to reach out to participants, maintaining a constant connection, and making participants feel involved in the process. As part of their paper addressing methodological issues that can occur while designing a Delphi template for studies involving primary care workers, Toma and Picioreanu (2016) proposed that Delphi practitioners take a more personal approach, contacting all panellists through motivating phone calls and circulating a study newsletter every two weeks throughout the study to maintain a high response rate and completion rate. This suggestion is supported by McKenna (1994) who discovered that conducting face-to-face interviews at round 1 increased the chances of participants returning postal questionnaires in round 2; panel members appeared to appreciate the personal touch. Turnbull et al. (2018) found that giving panellists repeated reminders by email and text message helped achieve a high response rate (>90%) throughout the Delphi process and the majority (96%) of participants did not find the reminders annoying. This direct personal communication approach with participants is likely to be more labour-intensive and will take longer than simply sending automated surveys to personally unknown volunteers. Overall, the type of communication techniques employed can have a profound impact on the results obtained (Hasson et al., 2000). In addition, anonymity – allowing a participant to change their opinion without loss of face – is an important design principle of Delphi and personal contact with participants can undermine anonymity to some degree (depending on how it is done).

A further issue that can decrease participant input is a lack of understanding of the task in hand or of the topic – see Step 3, earlier. In their study examining the degree of consensus about statements relating to general interdisciplinary practice using a Delphi procedure, Holey, Feeley, Dixon and Whittaker (2007) found that there was a decrease in the number of participants' comments as each round passed, which was attributed partly to the participants not understanding the statements they were given as feedback between Delphi rounds. With this in mind it is crucial that a facilitator clarifies with participants that they understand the material/topic they are discussing during a Delph procedure and that any uncertainties are resolved. Level of participant understanding can potentially be assessed prior to a planned Delphi survey, for example using a focussed questionnaire (Sinha et al., 2011).

Experts are also more likely to participate and continuously respond if they see the purpose and relevance of the Delphi exercise, or if the decision to be made as a result of the Delphi has an impact on them personally (Hasson et al., 2000 – see also Step 1, earlier). As part the analysis of a series of exploratory case studies using Delphi, Goluchowicz and Blind (2011) observed that there were more self-rated experts in the topic of interest participating and providing responses in the second rounds of the Delphi procedures in comparison to self-rated non-experts. This suggests that participants who conceptualise themselves as knowledgeable in a given topic are more likely to respond in subsequent Delphi rounds, perhaps because they are more likely to feel the Delphi exercise is relevant to them.

Overall, based on our experience and the evidence presented in the literature, we would recommend that researchers:

- Consider using of a financial or social reward (e.g. give social recognition of participation by naming the respondent as a representative of a particular community of respondents) to reduce possible attrition rates during a Delphi procedure but be aware of the possible downsides of using participant payment
- Adopt a more personal approach when communicating with participants, both during rounds and between rounds to maximise continued willingness to participate
- Be aware of how engaged participants are with the topic of interest, and the extent to which they see it (and the ultimate outcome of the Delphi process) as relevant to them, as this will likely have an impact on the responses given

Step 6: Analysing and presenting the Delphi data

The single most important outcome of any study, regardless of the methodology used, is the data produced. The literature on Delphi studies demonstrates that there are many ways in which Delphi study data can be analysed and reported (Hasson et al., 2000) and that the data presented can be open to alternative interpretations (Franklin & Hart, 2007). Data outputs from a Delphi study can be analysed in a variety of ways and can incorporate both qualitative and quantitative forms of data analysis (Hsu & Sandford, 2007; Hasson et al., 2000). The most commonly used form of descriptive statistical analysis in Delphi studies is the measurement of central tendency (i.e., mean, median and mode) and level of dispersion (e.g., standard deviation and inter-quartile range) (Hasson et al., 2000). Since Delphi studies very often use Likert-type response scales, which generate ordinal data, non-parametric summary measures of central tendency (median and inter-quartile range) are appropriate (Jamieson, 2004; Liddell & Kruschke, 2018; Norman, 2010). Inferential non-parametric statistical analysis may also be useful in some situations (e.g. Linebach, Tesch, & Kovacsiss, 2013), where Delphi participants have been randomly sampled. We discuss this important issue in detail, later.

Qualitative data produced by a Delphi, for example arguments for (and against) the consensus opinions reached, can be analysed and reported in various ways (de Loë et al., 2016). Qualitative analytical methods such as content analysis (e.g. Setty, Padmanabhan, & Natarajan, 1987; von der Gracht & Darkow, 2010) and thematic analysis (e.g. Collins et al., 2010; Toma & Picioreanu, 2016) may be appropriate here. Content analysis lends itself to graphical presentation of the data – for example, bar charts of the percentage of arguments containing particular themes or content – while thematic analysis permits richer, more detailed exploration of the issues arising from the data. Software such as NVivo can assist in the management, analysis and presentation of qualitative data (see www.qsrinternational.com/nvivo/home).

The results of a Delphi study can also be reported in many different formats (de Loë et al., 2016; Hasson et al., 2000). Of the 63 Delphi studies reviewed by de Loë et al. (2016), 12 presented a written narrative only, 43 included a table of descriptive statistics such as central tendency values for each item, and 29 used graphical forms of one kind or another with or without statistics (bar graphs, plot graphs, boxplots, dendrograms etc.). Other graphical display options include scatterplots (Gary & von der Gracht, 2015) and radar charts (Czaplicka-Kolarz et al., 2009). It can also be helpful to display the sequence and type of responses across Delphi rounds, whether in tabular fashion (e.g. Holey et al., 2007; Kim et al., 2014), graphically (e.g. Bisson et al., 2010; Ward, Stebbings, Sherman, Cherkin, & Baxter, 2014) or both (e.g. Becker & Roberts, 2009). We recommend making use of a graphical data-summary format. It is crucial to remember who the audience for the Delphi reporting is and whether they will need to be given explicit guidance on how to translate and process the data summary presented to them (Hasson et al., 2000),

and how the findings link to the purpose and methodology of the study. For example, the intended reader of the report may not understand numerical summary measures such as the inter-quartile range, etc.

In practice, Delphi surveys often form part of a larger process, as can be seen in the national-level foresight exercises used for Future-oriented Technology Analysis (e.g. Choi & Choi, 2015; Cuhls et al., 2002; Georghiou, 1996; Gheorghiu et al., 2017; Li, Chen, & Kou, 2017; NISTEP, 2009). In such cases, the results of the Delphi survey must be analysed and presented in a format appropriate for the subsequent steps in that process, which may include workshops (Georghiou, 1996), a scenarios exercise (Choi & Choi, 2015), roadmapping (Cuhls, de Vries, Li, & Li, 2015) and/or other methods.

An important issue to bear in mind when analysing and summarising data from a Delphi study is that of researcher bias, particularly when there is a large quantity of data produced (Hasson et al., 2000). Due to the large amount of processing and interpretation of experts' responses by researchers, there is a risk of the researchers having too much free reign over the response data outcomes generated and reported (Franklin & Hart, 2007; Green et al., 1999). Franklin and Hart (2007) used a 'member check' procedure to avoid researcher bias, whereby the expert panellists read the Delphi-collated individual responses and moderated the researchers' report writing.

Throughout the design, facilitation and reporting of a Delphi study, it is crucial to consider the reliability and validity of the study, in order to maximise research rigour and credibility (Rowe & Wright, 2011). Reliability is the extent to which a study procedure delivers similar outcomes (e.g., answers) each time it is repeated (e.g., asked) (Hasson et al., 2000). Validity can be defined as the degree to which a test measures what it purports to measure (Coulacoglou & Saklofske, 2018). Various sub-categories of reliability and validity have been identified in the psychometric literature (e.g. Cronbach & Meehl, 1955; Coulacoglou & Saklofske, 2018). Lack of space prevents a detailed discussion here but specific sub-categories will be defined where discussed in the following paragraphs.

Some researchers have claimed that Delphi enhances reliability compared to other methods (e.g. Clayton, 1997; Gordon, 1992). Others have argued that there is no evidence for the reliability of the Delphi method, since there is no way of confirming whether two or more panels who went through the same Delphi process would produce the same results (inter-observer reliability – Hasson et al., 2000; Keeney et al., 2001; Williams & Webb, 1978). However, Duffield (1993) and Kastein et al. (1993) provide some empirical support for inter-observer reliability: the studies each compared two Delphi panels with similar characteristics who started with the same information and found that they agreed on 93% and 88% of items, respectively (see also Quintana et al., 2000). Studies have attempted to assess the stability of Delphi results over time (test-retest reliability – e.g. Helmer, 1968; Uhl, 1975) but this is arguably inappropriate for Delphi, since participants are supposed to change their opinion during the process (Okoli & Pawlowski, 2004). Intra-respondent test-retest reliability can be measured at approximately the same point in time if the same Delphi question is asked twice but in different formats – perhaps by reversing the wording of a particular question posed to respondents and presenting the repeated question at a later point within a long Delphi questionnaire. However, to our knowledge, such sophisticated application of psychometric protocols has not been utilised to evaluate reliability within Delphi questionnaires. The psychometric literature is concerned with the development of questionnaires that will be used in selection decisions, for example intelligence tests, aptitude tests and personality tests. Such human resource decision making requires questionnaires that have high reliability and validity. Well-conducted Delphi applications should in our view, strive for the same standards. For this reason, we recommend – if time permits, see Step 1, above – that important Delphi questions are posed

in two forms (perhaps using reversed wording) in each Delphi round and the reliability of intra-respondent responses are verified.

The content validity of a Delphi study (the extent to which the study addresses all and only the issues relevant to the topic of interest – see Step 1, above) can be enhanced by ensuring that the panel is a representative sample of relevant experts who are motivated to participate (Goodman, 1987; Hasson et al., 2000) and giving panellists adequate opportunity to comment on the items that are being considered for inclusion in the survey (Okoli & Pawlowski, 2004). Conversely, a study will likely have lower content validity if response rates over Delphi rounds are low since this could be an indication that the questions posed generated a lack interest in the focal respondents (Rowe, Wright & Bolger, 1991).

Turning to Delphi's predictive validity (i.e. accuracy), we note that Delphi applications often involve the generation of long-term forecasts (e.g., Rowe and Wright, 2011) with, say, 25-year time horizons – which can only be evaluated for validity by comparing the forecasts with subsequent events (e.g. Dalkey, 1969; Ono & Wedermeyer, 1994; Parente & Anderson-Parente, 2011). In addition, many Delphi applications do not result in forecasts (since they are policy-focussed; Rowe & Wright, 2011) and the Delphi process can produce self-fulfilling prophecies, in that policy changes are subsequently made in line with the Delphi yield (De Meyrick, 2002). Fortunately, the face validity (or "social validity" – Landeta, 2006) of a Delphi study should not be discounted, since it is important that the stakeholders in the process are satisfied with the outcome of the study and are prepared to apply its findings. One of Delphi's strengths is that it not only helps experts reach consensus but can also help them understand how and why they reached that point.

The term validity is sometimes used in the Delphi literature in a sense that does not fit neatly into the psychometric sub-categories but is best captured by the term "external validity", used in psychological research to mean the extent to which the conclusions of a study can be generalised to the wider population (e.g. Davis, 2005). This is a somewhat similar concept to inter-observer reliability. Several aspects of Delphi design can improve external validity, such as using as large, representative and heterogeneous a sample of experts as possible (Bolger & Wright, 2011; Goodman, 1987) and ensuring panellists provide reasoned arguments in support of their judgments - such that a consensus that develops over Delphi rounds utilises all the information held by the individual members of the Delphi group.

It is important to note that the analysis of data produced from a Delphi study is influenced substantially by the make-up of the samples used in these studies. Delphi respondent samples are often comprised of groups of experts who have been chosen on a criterion sampling basis (Hasson et al., 2000), or have been chosen because for their expertise knowledge in a particular area or topic (Hsu & Sandford, 2007). Leaving this crucial issue of non-random sampling aside, the size and heterogeneity of the group used in a particular Delphi study can be used as a proxy to support the generalisability of its results (Hasson et al., 2000). In fact, studies have used a range of sizes in attempts to overcome challenges to the reliability and rigour of subsequent inferences (Hasson et al., 2000). However, it should be recognised that while utilising a larger Delphi group size will, to some extent, defend the later use of a Delphi yield for the generalisation of results, the sampling method – when not obtained by random selection – remains vulnerable to critiques of generalisability of the Delphi findings. Developing this point, Lincoln and Guba (1985) argue that the aim of qualitative research is not generalisability, it is transferability (Lincoln & Guba, 1985; see also e.g., Anney, 2014; Kuper, Reeves, & Levinson, 2008). Transferability involves asking whether the research describes the phenomenon in sufficient detail to allow readers to evaluate the extent to which the conclusions drawn are transferrable to other times, contexts, and people. The process can be viewed as a collaborative exercise between researcher and reader (Polit & Beck,

2010) and for that reason is sometimes referred to as "reader generalizability" (e.g., Misco, 2007). For some Delphi studies, thinking in terms of transferability may offer a more helpful way of evaluating the study's ultimate value.

Overall, it is important to remember that there is no single, standard form of the Delphi technique (Woudenberg, 1991) and so there is little value in debating whether Delphi is a reliable or valid technique in general terms. The correct question is whether any given Delphi design can be said to be valid and or reliable, and to what extent (Rowe & Wright, 2001). While the language of reliability and validity often applied to Delphi has its origins in psychometrics, Delphi is not a psychometric instrument but is a practical, practitioner-derived method for gauging group-based, subjective judgment. What is needed is much more empirical research to explore the conditions under which a particular Delphi process is likely to be more or less effective (Bolger & Wright, 2011; Bolger et al., 2011; Rowe & Wright, 1999).

Our recommendations for analysing and reporting Delphi study data are as follows:

- Use appropriate descriptive statistics to summarise the central tendency and spread of Delphi yield
- Complement numerical summaries with visual summaries such as bar charts and boxplots
- Be aware of potential challenges to your study's reliability and validity - if you intend to generalise the results of your Delphi survey

Conclusion

In this paper we have outlined what we believe are the six crucial procedural steps of conducting a successful Delphi study. We have proposed a template of the progressive six-step process and have provided evidence from the literature to support our evaluation of the design features and procedural specifics we think are imperative to Delphi. Based on this evidence we have given specific recommendations about specific feature that should be considered when carrying out a Delphi study – such as the choice of expert panellists, the type of response scales that should be used, the questions that should be asked as part of the Delph survey, how feedback should be provided and how the response data from the experts should be analysed and disseminated. We realise that the procedure of a Delphi study can be heavily influenced by the topic or area of concern (e.g. healthcare, education, foresight, logistic problem-solving), but we recommend that practitioners apply these six steps when conducting a Delphi in order to maximise the rigour of the process and the effectiveness and accuracy of the judgement data produced.

Acknowledgments

This paper was funded in part by the European Union Agency for Fundamental Rights.

References

Aengenheyster, S., Cuhls, K., Gerhold, L., Heiskanen-Schüttler, M., Huck, J., & Muszynska, M. (2017). Real-Time Delphi in practice – A comparative analysis of existing software-based tools. *Technological Forecasting and Social Change*, *118*, 15-27.

Agell, N., Ganzewinkel, C. J., Sánchez, M., Roselló, Prats, F., & Andriessen, P. (2015). A consensus model for Delphi processes with linguistic terms and its application to chronic pain in neonates definition. *Applied Soft Computing, 35,* 942-948.

Airaksinen, T., Halinen, I., & Linturi, H. (2017). Futuribles of learning 2030: Delphi supports the reform of the core curricula in Finland. *European Journal of Futures Research, 5*(2).

Anney, V. N. (2014). Ensuring the quality of the findings of qualitative research: Looking at trustworthiness criteria. *Journal of Emerging Trends in Educational Research and Policy Studies*, 5(2), 272-281.

APEC Center for Technology Foresight (2009). *Report of Delphi analysis for the low carbon beyond 2050 project.* Retrieved from the APEC CTF website: http://www.apecctf.org/index.php/publications.html

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education, 53*(2), 182–200.

Becker, G. E., & Roberts, T. (2009). Do we agree? Using a Delphi technique to develop consensus on skills of hand expression. *Journal of Human Lactation, 25*(2), 220-225.

Belton, I.K., Sissons, A., Rowe, G., Bolger, F., Crawford, M., Hamlin, I., … Wright, G. (2019). *The Delphi technique: Does (group) size matter?* Manuscript in preparation.

Bisson, J. I., Tavakoly, B., Witteveen, A. B., Ajdukovic, D., Jehel, L., Johansen, V. J., … Olff, M. (2010). TENTS guidelines: Development of post-disaster psychosocial care guidelines through a Delphi process. *The 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., 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*(9), 1671-1680.

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

Boulkedid, R., Abdoul, H., Loustau, M., Sibony, O., & Alberti, C. (2011). Using and reporting the Delphi method for selecting healthcare quality indicators: A systematic review. *PLoS ONE*, *6*(6), 1-9. doi:10.1371/journal.pone.0020476

Brender, J., Ammenwerth, E., Nykänen, P., & Talmon, J. (2006). Factors influencing success and failure of health informatics systems: A pilot Delphi study. *Methods of Information in Medicine, 45,* 125–136.

Briedenhann, J., & Butts, S. (2006). Application of the Delphi technique to rural tourism project evaluation. *Current Issues in Tourism, 9*(2), 171-190.

Brockhoff, K. (1975). The performance of forecasting groups in computer dialog and face-to-face discussion. In H. A. Linstone & M. Turoff (Eds.), *The Delphi method: Techniques and applications* (pp. 291-321). Reading, MA: Addison-Wesley.

Choi, M., & Choi, H. (2015). Foresight for science and technology setting in Korea. *Foresight and STI Governance, 9*(3), 54-65.

Cicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement*, *9*(1), 31-36.

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

Collins, J. A., Hanlon, A., More, S. J., Wall, P. G., Kennedy, J., & Duggan, V. (2010). Evaluation of current equine welfare issues in Ireland: causes, desirability, feasibility and means of raising standards. *Equine Veterinary Journal, 42*, 105–113.

Coulacoglou, C., & Saklofske, D. H. (2018). *Psychometrics and Psychological Assessment: Principles and applications.* London: Elsevier.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Cuhls, K. (2015). *Lessons for policy-making from foresight in non-European countries. Policy paper by the Research, Innovation and Science Policy Experts (RISE).* Retrieved from the RISE website: http://ec.europa.eu/research/openvision/pdf/rise/cuhls-lessons_policy_making.pdf

Cuhls, K., Blind, K., & Grupp, H. (2002). *Innovations for our future. Delphi '98: New foresight on science and technology.* Heidelberg; Physica Verlag/Springer.

Cuhls, K., de Vries, M., & Li, H., & Li, L. Roadmapping: Comparing cases in China and Germany. *Technological Forecasting and Social Change, 101*, 238-250.

Czaplicka-Kolarz, K., Stańczyk, 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 Forecasting & Social Change, 76,* 327-338.

Dajani, J. S., Sincoff, M. Z., & Talley, W. K. (1979). Stability and agreement criteria for the termination of Delphi studies. *Technological Forecasting & Social Change, 13*, 83–90.

Dalkey, N. C. (1969). *The Delphi method: An experimental study of group opinion. Document Number RM-5888-PR.* Santa Monica, CA: Rand Corporation.

Davis, S. F. (Ed.) (2005). *Handbook of research methods in experimental psychology.* Oxford: Blackwell Publishing Ltd.

Day, J., & Bobeva, M. (2005). A generic toolkit for the successful management of Delphi studies. *The Electronic Journal of Business Research Methodology*, *3*(2), 103-116.

De Loë, R.C., Melnychuk, N., Murray, D., & Plummer, R. (2016). Advancing the state of policy Delphi practice: A systematic review evaluating methodological evolution, innovation, and opportunities. *Technological Forecasting and Social Change, 104,* 78-88.

De Meyrick, J. (2002). The Delphi method and health research. *Health Education, 103*(1), 7-16.

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

Devaney, L., & Henchion, M. (2018). Who is a Delphi 'expert'? Reflections on a bioeconomy expert selection procedure in Ireland. *Futures, 99,* 45-55.

Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology, 67*, 401-409.

Duffield, C. (1993). The Delphi technique: a comparison of results obtained using two expert panels. *International Journal of Nursing Studies, 30*(3), 227-237.

Duru, O., Bulut, E., & Yoshida, S. (2012). A fuzzy extended Delphi method for adjustment of statistical time series prediction: An empirical study on dry bulk freight market case. *Expert Systems with Applications, 39*(1), 840-848.

Eggers, R. M., & Jones, C. M. (1998). Practical considerations for conducting Delphi studies: The oracle enters a new age. *Educational Research Quarterly, 21*(3), 53-67.

Elwyn, G., O'connor, A., Stacey, D., Volk, R., Edwards, A., Coulter, A., ... & Butow, P. (2006). Developing a quality criteria framework for patient decision aids: Online international Delphi consensus process. *BMJ*, *333*(7565), 417.

Erffmeyer, R. C., Erffmeyer, E. S., & Lane, I. M. (1986). The Delphi technique: An empirical evaluation of the optimal number of rounds. *Group & organization studies*, *11*(1-2), 120-128.

Foth, T., Efstathiou, N., Vanderspank-Wright, B., Ufholz, L, Dutthorn, N., Zimansky, M., & Humphrey-Murto, S. (2016). The use of Delphi and Nominal Group Technique in nursing education: A review. *International Journal of Nursing Studies, 60,* 112-120.

Franklin, K. K., & Hart, J. K. (2007). Idea generation and exploration: Benefits and limitations of the policy Delphi research method. *Innovative Higher Education*, *31*(4), 237-246.

Frewer, L. J., Fischer, A. R. H., Wentholt, M. T. A., Marvin, H. J. P., Ooms, B. W., Coles, D., & Rowe, G. (2011). The use of Delphi methodology in agrifood policy development: some lessons learned. *Technological Forecasting and Social Change*, *78*(9), 1514-1525.

Frick, A., Bächtiger, M.-T., & Reips, U.-D. (2001). Financial incentives, personal information, and drop out in online studies. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 209–219). Lengerich: Pabst Science.

Furr, R. M. (2011). *Scale construction and psychometrics for social and personality psychology*. London: SAGE Publications.

Gary, J. E., & von der Gracht, H. A. (2015). The future of foresight professionals: Results from a global Delphi study. *Futures, 71,* 132-145.

Geist, M. R. (2010). Using the Delphi method to engage stakeholders: A comparison of two studies. *Evaluation and Program Planning, 33,* 147-154.

Georghiou, L. (1996). The UK technology foresight programme. *Futures, 28*(4), 359-377.

Gheorghiu, R., Dragomir, B., Andreescu, L., Cuhls, K., Rosa, A. Curaj, A., & Weber, M. (2017). *New horizons: Data from a Delphi survey in support of European Union futures policies in research and innovation.* Retrieved from https://ec.europa.eu/jrc/sites/jrcsh/files/fta2018-paper-c3-cuhls.pdf

Gnatzy, T., Warth, J., von der Gracht, H., & Darkow, I. (2011). Validating an innovative real-time Delphi approach – A methodological comparison between real-time and conventional Delphi studies. *Technological Forecasting & Social Change, 78*, 1681-1694.

Goluchowicz, K., & Blind, K. (2011). Identification of future fields of standardisation: An explorative application of the Delphi methodology. *Technological Forecasting and Social Change*, *78*(9), 1526-1541.

Goodman, C.M. (1987). The Delphi technique: a critique. *Journal of Advanced Nursing,* 12, 729-734.

Gordon, T. J. (1992). The methods of futures research. *Annals of the American Academy of Political & Social Science, 522,* 25-35.

Gordon, T.J. (2003). The Delphi method. In J.C. Glenn, T.J. Gordon (Eds.), *Futures Research Methodology Version 2.0*. Washington: American Council for the United Nations University.

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

Göritz, A. S. (2008). The long-term effect of material incentives on participation in online panels. *Field Methods, 20*, 211–225.

Göritz, A. S. (2010). Using lotteries, loyalty points, and other incentives to increase participant response and completion. In S.D. Gosling, & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 219-233). Washington, DC, US: American Psychological Association.

Green, B., Jones, M., Hughes, D., & Williams, A. (1999). Applying the Delphi technique in a study of GP's information requirements. *Health and Social Care in the Community, 7*(3), 198-205.

Grime, M. M. & Wright, G. (2016). Delphi. In Brandimarte, P., Everitt, B., Molenberghs, G., Piegorsch, W. & Ruggeri, F. (Eds.), *2016 Wiley StatsRef: Statistics Reference Online.* New York, N.Y.

Hallowell, M. R. & Gambatese, J. A. (2010). Qualitative research: Application of the Delphi method to CEM research. *Journal of Construction Engineering and Management, 136*(1), 99-107.

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

Heerwegh, D. (2006). An investigation of the effect of lotteries on web survey response rates. *Field Methods, 18* (2) (2006), pp. 205-220

Helmer, O. (1968). The Delphi method: An illustration. In J. R. Bright (Ed.). *Technological forecasting for industry and government* (pp. 117-122). Englewood Cliffs, NJ: Prentice-Hall.

Holey, E. A., Feeley, J. L., Dixon, J., & Whittaker, V. J. (2007). An exploration of the use of simple statistics to measure consensus and stability in Delphi studies. *BMC medical research methodology*, *7*(1), 52.

Hsu, C.-C., & Sandford, B. A. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research & Evaluation*, *12*(10), 1-8. Retrieved from http://pareonline.net/getvn.asp?v=12&n=10

Humphrey-Murto, S., & de Wit, M. (2019). The Delphi method – more research please. *Journal of Clinical Epidemiology, 106,* 136-139.

Hung, H., Altschuld, J. W., & Lee, Y. (2008). Methodological and conceptual issues confronting a cross-country Delphi study of educational program evaluation. *Evaluation and Program Planning,* 191-198.

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

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*, 1217–1218.

Kalaian, S. A., & Kasim, R. M. Terminating Sequential Delphi Survey Data Collection. *Practical Assessment Research and Evaluation*, *17*, 1-10.

Kastein, M. R., Jacobs, M., Van Der Hell, R. H., Luttik, K., & Touw-Otten, F. W. M. M. (1993). Delphi, the issue of reliability: A qualitative Delphi study in primary health care in the Netherlands. *Technological Forecasting & Social Change, 44,* 315-323.

Keeney, S., Hasson, F., & McKenna, H. P. (2001). A critical review of the Delphi technique as a research methodology for nursing. *International Journal of Nursing Studies*, 38, 195-200.

Keeney, S., Hasson, F., & McKenna, H. P. (2006). Consulting the oracle: Ten lessons from using the Delphi technique in nursing research. *Journal of Advanced Nursing, 53*(2), 205-212.

Kim, C. H., Park, J. O., Park, C. B., Kim, S. C., Kim, S. J., & Hong, K. J. (2014). Scientific framework for research on disaster and mass casualty incident in Korea: Building consensus using Delphi method. *Journal of Korean Medical Science, 29,* 122-128.

Knapp T.R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research, 39,* 121–123.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213−236.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Eds.). *Handbook of Survey Research* (pp. 263-313). West Yorkshire, England: Emerald Group.

Krosnick, J.A., Judd, C.M., & Wittenbrink, B. (2005). Attitude measurement. In D. Albarracin, B.T. Johnson, & M.P. Zanna (Eds.), *Handbook of Attitudes and Attitude Change* (pp. 21-76). Mahwah, NJ: Erlbaum.

Kuper, A., Reeves, S., & Levinson, W. (2008). An introduction to reading and appraising qualitative research. *BMJ: British Medical Journal, 337(*7666), 404-407.

Landeta, J. (2006). Current validity of the Delphi method in social science. *Technological Forecasting & Social Change, 73,* 467-482.

Landeta, J., & Barrutia, J. (2011). People consultation to construct the future: A Delphi application. *International Journal of Forecasting, 27*(1), 134-151.

Landeta, J., Barrutia, J., & Lertxundi, A. (2011). Hybrid Delphi: A methodology to facilitate contribution from experts in professional contexts. *Technological Forecasting & Social Change, 78,* 1629-1641.

Li, N., Chen, K., & Kou, M. (2017). Technology foresight in China: Academic studies, governmental practices and policy applications. *Technological Forecasting and Social Change, 119*, 246-255.

Liddell, T. M., & Kruschke, J. K. (2018). Analysing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328-348.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry* (Vol. 75). California, USA: Sage Publications.

Linebach, J. A., Tesch, B. P., & Kovacsiss, L. M. (2013). *Nonparametric statistics for applied research.* New York, NY: Springer.

Linstone, H.A., Turoff, M. (1975). *The Delphi method: Techniques and applications.* Reading, MA: Addison-Wesley Publishing Co.

Loveridge, D. (2002). *On Delphi questions: Ideas in Progress*. Manchester: The University of Manchester.

Makkonen, M., Hujala, T., & Uusivuori, J. (2016). Policy experts' propensity to change their opinion along Delphi rounds. *Technological Forecasting and Social Change*, *109*, 61-68.

McKenna, H. P. (1994). The Delphi technique: a worthwhile research approach for nursing. *Journal of advanced nursing*, *19*(6), 1221-1225.

McMillan, S. S., King, M., & Tully, M. P. (2016). How to use the nominal group and Delphi techniques. *International Journal of Clinical Pharmacy*, *38*, 655-662.

Meijering, J. V., & Tobi, H. (2016). The effect of controlled opinion feedback on Delphi features: Mixed messages from a real-world Delphi experiment. *Technological Forecasting and Social Change, 103*, 166-173.

Melander, L. (2018). Scenario development in transport studies: Methodological considerations and reflections on Delphi studies. *Futures, 96,* 68-78.

Melnyk, S. A., Lummus, R. R., Vokurka, R. J., Burns, L.J., & Sandor, J. (2009). Mapping the future of supply chain management: A Delphi study. *International Journal of Production Research, 47*(16), 4629-4653.

Mitchell, V. W. (1991). The Delphi technique: An exposition and application. *Technology Analysis & Strategic Management*, *3*(4), 333-358.

Misco, T. (2007). The frustrations of reader generalizability and grounded theory: Alternative considerations for transferability. *Journal of Research Practice, 3*(1), 1-11.

Mitchell, V. W. (1991). The Delphi technique: An exposition and application. *Technology Analysis and Strategic Management, 3*(4), 333-358.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … de Vet, H. C. W. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19,* 539-549.

Morrison, A. P., & Barratt, S. (2010). What are the components of CBT for psychosis? A Delphi study. *Schizophrenia Bulletin, 36*(1), 136-142.

NISTEP (2009). *The 9th science and technology foresight – contribution of science and technology to future society – the 9th Delphi survey. NISTEP report no. 140*. Retrieved from http://www.nistep.go.jp/en/?page_id=3800

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Science Education, 15*, 625-632.

Novakowski, N., & Wellar B. (2008). Using the Delphi technique in normative planning research: Methodological design considerations. *Environment and Planning A, 40,* 1485-1500.

Ogden, S. R., Culp, W. C., Villamaria, F. J., & Ball, T. R. (2016). Developing a checklist: consensus via a modified Delphi technique. *Journal of Cardiothoracic and Vascular Anesthesia*, *30*(4), 855-858.

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & management*, *42*(1), 15-29.

Ono, R., & Wedemeyer, D. J. (1994). Assessing the validity of the Delphi technique. *Futures, 26*(3), 289-304.

Parente, R., & Anderson-Parente, J. (2011). A case study of long-term Delphi accuracy. *Technological Forecasting & Social Change, 78,* 1705-1711.

Polit, D.F., & Beck, C.T. (2010). Generalization in quantitative and qualitative research: Myths and strategies. *International Journal of Nursing Studies, 47,* 1451-1458.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, *104*(1), 1-15.

Quintana, J.M., Arostegui, I., Azkarate, J., Goenaga, J.I., Elexpa, X., Letona, J., & Arcelay, A. (2000). Evaluation of explicit criteria for total hip joint replacement. *Journal of Clinical Epidemiology, 53*(12), 1200-1208.

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting, 15,* 353-375.

Rowe, G., & Wright, G. (2001). Expert opinions in forecasting: the role of the Delphi technique. In J.S. Armstrong (Ed.), *Principles of Forecasting* (pp. 125-144). US: Springer.

Rowe, G., & Wright, G. (2011). The Delphi technique: Past, present, and future prospects – Introduction to the special issue. *Technological Forecasting and Social Change*, *78*, 1487-1490.

Rowe, G., Wright, G., & Bolger, F. (1991). Delphi: A reevaluation of research and theory. *Technological Forecasting and Social Change, 39,* 235-251.

Salancik, J. R., Wenger, W., & Helfer, E. (1971). The construction of Delphi event statements. *Technological Forecasting & Social Change, 3,* 65–73.

Scheibe, M., Skutsch, M., Schofer, J. (1975). Experiments in Delphi methodology. In H.A. Linstone, M. Turoff (Eds.), *The Delphi method: Techniques and applications* (pp. 262–287). Reading, MA: Addison-Wesley Publishing Co.

Schmidt, R. C. (1997). Managing Delphi surveys using nonparametric statistical techniques. *Decision Sciences, 28*(3), 763-773.

Setty, K. P. S., Padmanabhan, S., & Natarajan, R. (1987). A national energy-conservation policy Delphi: report of the findings. *Technological Forecasting & Social Change, 31,* 257–267.

Sinha, I. P., Smyth, R. L., & Williamson, P. R. (2011). Using the Delphi technique to determine which outcomes to measure in clinical trials: Recommendations for the future based on a systematic review of existing studies. *PLOS Medicine, 8*(1), e1000393.

Sokolov, A. (2008). *Science and technology foresight in Russia: Results of a national Delphi.* Paper presented at the third international Seville seminar on future-oriented technology analysis: Impacts and implications for policy and decision-making, Seville, 16-17 October 2008.

Spickermann, A., Zimmermann, M., & von der Gracht, H. A. (2014). Surface- and deep-level diversity in panel selection – Exploring diversity effects on response behaviour in foresight. *Technological Forecasting & Social Change, 84,* 105-120.

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008, Feb). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45*, 116−131.

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.

Trevelyan, E.G., & Robinson, N. (2015). Delphi methodology in health research: How to do it? *European Journal of Integrative Medicine, 7*, 423-428.

Turnbull, A. E., Dinglas, V. D., Friedman, L. A., Chessare, C. M., Sepúlveda, Bingham, C.O., III, & Needham, D. M. (2018). A survey of Delphi panellists after core outcome set development revealed positive feedback and methods to facilitate panel member participation. *Journal of Clinical Epidemiology, 102,* 99-106.

Uhl, N. P. (1975). *Consensus and the Delphi process.* Paper presented at the annual meeting of the American Educational Research Association (Washington, D.C., April 1975). ERIC document ED 104201.

Van der Steen, J. T., Radbruch, L., Hertogh, C. M. P. M., de Boer, M. E., Hughes, J. C., Larkin, P.,… Volicer, L. (2014). White paper defining optimal palliative care in older people with dementia: A Delphi study and recommendations from the European Association for Palliative Care. *Palliative Medicine, 28*(3), 197-209.

Von der Gracht, H. A. (2012). Consensus measurement in Delphi studies: Review and implications for future quality assurance. *Technological Forecasting & Social Change, 79,* 1525-1536.

Von der Gracht, H. A., & Darkow, I. L. (2010). Scenarios for the logistics services industry: A Delphi-based analysis for 2025. *International Journal of Production Economics, 127*, 46–59.

Ward, L., Stebbings, S., Sherman, K. J., Cherkin, D., & Baxter, G. D. (2014). Establishing key components of yoga interventions for musculoskeletal conditions: A Delphi survey. *BMC Complementary and Alternative Medicine, 14,* 196-208.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27(*3), 236–247.

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*(6), 956-972.

Williams, P.L., & Webb, C. (1978). The Delphi technique: a methodological discussion. *Journal of Advanced Nursing, 19*, 180-186.

Winkler, J., & Moser, R. (2016). Biases in future-oriented Delphi studies: A cognitive perspective. *Technological Forecasting & Social Change, 105,* 63-76.

Witkin, B. R., & Altschuld, J.W. (1995). *Planning and conducting needs assessments: A practical guide.* Thousand Oaks, CA: Sage.

Woudenberg, F. (1991). An evaluation of Delphi. *Technological forecasting and social change*, *40*(2), 131-150.

Wright, G., & Rowe, G. (2011). Group-based judgmental forecasting: An integration of extant knowledge and the development of priorities for a new research agenda. *International Journal of Forecasting, 27*(1), 1-13.

Biographical endnotes

Ian Belton is a qualified lawyer who recently received a PhD in psychology from Middlesex University. Ian's research focuses on the psychology of human judgment and decision-making. In particular, he examines decisions made in organisational contexts such as the legal, defence and security sectors. He also has an interest in methods for structuring group judgment and the elicitation of expert judgment.

Alice MacDonald has an MA in Psychology from the University of Aberdeen and also an MSc in Research Methods in Psychology from the University of Strathclyde. She has an interest in Education, Educational Psychology and learning development.

George Wright is a psychologist with an interest in how judgments and decisions are made in the face of uncertainty about the future. Are these judgments and decisions sometimes flawed? If so, can behavioural and management science techniques improve their quality? George is editor of both the Journal of Behavioral Decision Making and of the new start journal, Futures & Foresight Science.

Iain Hamlin received a PhD in Psychology from Lancaster University and an MA in Psychology from the University of Edinburgh. His research interests focus on various issues in the domain of social perception and social interaction.