For more information on this publication, visit www.rand.org/t/TLA3082-1.

About RAND
The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. To learn more about RAND, visit www.rand.org.

Research Integrity
Our mission to help improve policy and decisionmaking through research and analysis is enabled through our core values of quality and objectivity and our unwavering commitment to the highest level of integrity and ethical behavior. To help ensure our research and analysis are rigorous, objective, and nonpartisan, we subject our research publications to a robust and exacting quality-assurance process; avoid both the appearance and reality of financial and other conflicts of interest through staff training, project screening, and a policy of mandatory disclosure; and pursue transparency in our research engagements through our commitment to the open publication of our research findings and recommendations, disclosure of the source of funding of published research, and policies to ensure intellectual independence. For more information, visit www.rand.org/about/research-integrity.

RAND’s publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif.
© 2023 RAND Corporation
RAND® is a registered trademark.

Cover: melitas/GettyImages.

Limited Print and Electronic Distribution Rights
This publication and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited; linking directly to its webpage on rand.org is encouraged. Permission is required from RAND to reproduce, or reuse in another form, any of its research products for commercial purposes. For information on reprint and reuse permissions, please visit www.rand.org/pubs/permissions.
About This Manual

The RAND Corporation developed the Delphi method in the late 1940s–early 1950s to help researchers explore the existence of consensus among experts. The method has been substantially expanded over time and is still in use today in both traditional and modified forms. Nonetheless, methodological guidance that identifies key considerations for researchers interested in using the Delphi method is thin. This document serves as a user manual to help researchers design and conduct high-quality studies using the Delphi method. The final chapter is a tool for critically assessing the quality of Delphi studies conducted by other researchers.

Center for Qualitative and Mixed Methods

RAND Global and Emerging Risks is a division at RAND that develops novel methods and delivers rigorous research on potential catastrophic risks confronting humanity. This work was undertaken by the division’s Center for Qualitative and Mixed Methods, which develops and promotes tools and methods to generate empirically based insights through exploratory data collection and analysis. For more information, visit www.rand.org/global-and-emerging-risks.html or contact the division director (contact information is provided on the webpage).

Funding

Funding for this research was provided by gifts from RAND supporters and income from operations.

Acknowledgments

We thank the Center for Qualitative and Mixed Methods and the Data Collection Design and Implementation Group at RAND for their support of this methodological work. We also thank our colleagues Aneesa Motala, Jody Larkin, and Catria Gadwah-Meaden for their help with the literature review. Finally, we are grateful to our reviewers, Lisa Rubenstein and Melissa Bradley, for their excellent comments and suggestions on the draft of this manual.
Contents

About This Manual ........................................................................................................ iii

CHAPTER 1
Introduction .................................................................................................................. 1
  Background ................................................................................................................. 1
  Purpose ....................................................................................................................... 3

CHAPTER 2
What Delphi Is ............................................................................................................. 5
  Classical Delphi .......................................................................................................... 5
  Delphi Panelists ........................................................................................................... 6
  Reasons for Using Delphi ............................................................................................. 7
  Question Items and Response Scales .......................................................................... 8
  Delphi Types ............................................................................................................... 8
  Methodological Limitations of the Delphi Method ....................................................... 12

CHAPTER 3
Design ............................................................................................................................ 13
  Justification for Using Delphi ..................................................................................... 13
  Selecting the Type of Delphi ....................................................................................... 14
  Assembling the Research Team (Roles, Responsibilities, and Expertise) .................... 14
  Specifying Panel Eligibility Criteria .......................................................................... 15
  Identifying Recruitment Strategies ............................................................................ 15
  Using Participant Retention Strategies ..................................................................... 16
  Conducting Preliminary Work ..................................................................................... 17
  Conducting a Pilot Test ............................................................................................... 18
  Deciding Which Study Features to Use ...................................................................... 18

CHAPTER 4
Data Collection .............................................................................................................. 27
  Items ........................................................................................................................... 27
  Questions ..................................................................................................................... 28
  Discussion Moderation ............................................................................................... 30
  Software Used ............................................................................................................ 31

CHAPTER 5
Data Analysis ................................................................................................................. 33
  Descriptive Statistics ................................................................................................. 33
  Stability ....................................................................................................................... 34
Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Classical Delphi Process</td>
<td>5</td>
</tr>
<tr>
<td>2.2</td>
<td>Key Characteristics of the Classical Delphi Method</td>
<td>6</td>
</tr>
<tr>
<td>2.3</td>
<td>ExpertLens as an E-Delphi</td>
<td>11</td>
</tr>
<tr>
<td>6.1</td>
<td>Forest Plot Example</td>
<td>38</td>
</tr>
<tr>
<td>7.1</td>
<td>Categories for Rating the Overall Quality of a Delphi Study</td>
<td>51</td>
</tr>
<tr>
<td>A.1</td>
<td>Recommendations for Design of Delphi Studies</td>
<td>53</td>
</tr>
<tr>
<td>A.2</td>
<td>Recommendations for Data Collection in Delphi Studies</td>
<td>54</td>
</tr>
<tr>
<td>A.3</td>
<td>Recommendations for Data Analysis in Delphi Studies</td>
<td>54</td>
</tr>
<tr>
<td>A.4</td>
<td>Recommendations for Reporting in Delphi Studies</td>
<td>54</td>
</tr>
<tr>
<td>B.1</td>
<td>DCAT Checklist</td>
<td>58</td>
</tr>
</tbody>
</table>
Introduction

As originally created in the late 1940s–early 1950s by a group of RAND Project AIR FORCE researchers, the Delphi method is an iterative, anonymous, structured, group-based elicitation technique. It was envisioned as a method for making more accurate social and technological forecasts and as a way to improve policy decisionmaking. Delphi is based on the premise that asking experts the same questions several times and sharing the other experts’ answers will help develop consensus, which can be used as a form of evidence (Dalkey, 1969). With time, the Delphi method evolved in terms of the meaning of its key characteristics, use cases, ultimate goals, and the types of individuals whose perspectives are elicited. Delphi is widely used today by a wide variety of academic disciplines, particularly the medical field, which is responsible for introducing numerous modifications of the method (Khodyakov et al., 2023).

Despite its popularity, methodological guidance for conducting Delphi studies and running Delphi panels is thin. One potential reason for the absence of such guidance is wide variation in the types of Delphi that exist nowadays. Therefore, this document serves as a user manual to help researchers understand variation in the way the Delphi method is used and to describe key Delphi design, implementation, and reporting considerations that researchers should account for to ensure high-quality Delphi studies, regardless of the type of Delphi they want to conduct. The final chapter of this document also includes a tool for critically assessing the quality of Delphi studies conducted by other researchers.

Background

The Delphi method is a structured group communication process that helps generate evidence or make decisions under conditions of uncertainty and incomplete information (Niederberger and Spranger, 2020). Delphi involves a series of anonymous questionnaires that a group of hand-picked experts (often referred to as Delphi panelists) answers, initially based on their expertise and personal experience and then based on seeing (and sometimes discussing) how their results compare with those of other panelists. The researchers designing and running Delphi studies (often referred to as panel conveners) may also provide reviews of existing research evidence on the topic under consideration before or in conjunction with completion of the first Delphi questionnaire. If little published evidence exists, panelists may be asked to weigh in on the issues to be included in subsequent rounds.
Once the panelists answer the first round of questions and explain their responses, they see how their answers compare with the other panelists’ responses and are asked to reconfirm or revise their original responses as needed. This process continues through additional rounds, until the responses are stable, a consensus is reached, or a certain number of rounds is completed. The Delphi method may involve paper-based, electronic, or in-person modalities for the data collection and feedback rounds. Panel conveners summarize final panel results, often generating a set of final consensus statements.

Delphi-based expert consensus is considered to be more reliable than the opinion of a single expert because it harnesses the wisdom of several knowledgeable individuals while mitigating the negative effects of dominant personalities and groupthink that are common in group settings (Dalal et al., 2011). The method empowers panelists who may represent different groups and have different viewpoints to have an equal voice in creating a consensus, for example, about the probability of an event happening in the future or the appropriateness of care for a certain type of patient.

Many disciplines use consensus from Delphi studies as a form of empirical evidence, especially when no other sources of evidence exist or evidence is inconclusive. Although Delphi was originally used in military research to make more reliable forecasts of future political developments (such as predicting the impact of technology on warfare), in the 1970s, social sciences, business, education, and psychology started using the Delphi method as an alternative, or complement, to regression-based models to forecast future events and to assess the feasibility and likely impact of various policies under consideration (Winkler and Moser, 2016). The method has also been used to help guide the real-world application of scientific research or technology by accounting for the nuances of the context within which it is being used.

Starting in the 1980s, the medical field became the biggest user of the Delphi method (Khodyakov et al., 2023). The method has been widely used to assess care quality (Fitch et al., 2001; Peter et al., 2016), develop clinical practice guidelines (Grant, Armstrong, and Khodyakov, 2021; Negrini et al., 2012), and identify end-points for clinical trials (Bellera et al., 2015), among other topics. Health researchers also introduced multiple modifications to the method itself, including adding in-person discussion meetings and developing a formal consensus measure (Fitch et al., 2001), often referring to the method itself as modified Delphi.

Delphi was originally developed as an expert elicitation methodology, but it is now used for stakeholder engagement, as well. This application of the method expanded the definition of expertise to include both technical and experiential knowledge (Dalal et al., 2011).

Although the Delphi method has been used by multiple disciplines for more than 70 years, we still lack methodological guidance for how to conduct rigorous Delphi studies that cuts across a wide range of use cases; that is not specific to a particular field, such as palliative care (Jünger et al., 2017); and that is not specific to a single Delphi study (Beiderbeck et al., 2021; Trevelyan and Robinson, 2015). While several ongoing initiatives aim to develop reporting guidelines for studies that use the Delphi method, these guidelines are discipline-specific and
Introduction

While useful, these guidelines focus on how to report a completed Delphi panel, rather than how to prospectively plan and conduct a future Delphi study or how to critically appraise the quality of a published Delphi study. Guidelines for designing, conducting, and assessing Delphi studies can increase methodological rigor and provide standards for evaluating methodological quality of research that can be useful for quality assurance purposes (Chandler and Hopewell, 2013).

Purpose

Therefore, the main goal of this manual is to provide practical advice for what to consider when designing, implementing, and reporting the results of a Delphi study that is applicable to different disciplines and use cases. The guidance is based on our team’s 15 years of experience designing, conducting, and analyzing the results of more than 40 Delphi studies, as well as a review of the published methodological work on the Delphi method that we identified as part of a recent bibliometric analysis of the Delphi literature (Khodyakov et al., 2023). While the manual is not designed to tell the reader how best to design or conduct a Delphi study—because of the variation in the type and uses of the Delphi method—we explain key methodological considerations applicable to the four commonly used types of Delphi studies that researchers should think about to ensure scientific rigor in their use of the Delphi method. Appendix A summarizes all the recommendations we make throughout this document.

This manual is structured around the four Delphi steps: study design, data collection, data analysis, and result reporting. It also includes the Delphi Critical Appraisal Tool (DCAT), which researchers can use both to judge the quality of published stud-

Although the Delphi method has been used by multiple disciplines for more than 70 years, we still lack methodological guidance for how to conduct rigorous Delphi studies that cuts across a wide range of use cases, that is not specific to a particular field, and that is not specific to a single Delphi study.

**Appendix A**

- **STEP 1** Design a study
- **STEP 2** Collect data
- **STEP 3** Analyze data
- **STEP 4** Report results
ies that use Delphi and to guide the rigorous design of their own studies by making quality standards explicit (Shea et al., 2017). Appendix B contains a DCAT checklist that can be used in the process of assessing a Delphi study’s quality.
CHAPTER 2

What Delphi Is

The Delphi method—with its four key characteristics of anonymity, iterative data collection, feedback to panelists, and statistical group response—is considered a best practice for objectively quantifying the results of group decisionmaking (Khodyakov and Chen, 2020; Rowe and Wright, 1999). Delphi research teams, which we also refer to as panel conveners in this manual, start designing a Delphi study by identifying the key research questions it addresses, developing a conceptual framework, and clarifying how the Delphi method fits with the other methods used in the study.

Classical Delphi

Although Delphi studies vary widely in their designs, Figure 2.1 illustrates how classical Delphi panels are commonly designed. Classical Delphi often begins with a series of open-ended questions that ask panelists to nominate issues to be

FIGURE 2.1
Classical Delphi Process

- **Round 0**: Panelists generate ideas.
- **Round 1**: Panelists answer questions and explain their responses.
- **Round 2**: Panelists review summary and can change their answer.
- **Scoring**: Final judgments are aggregated.

What was the consensus on this question?
considered in future rounds. This is often referred to as Round 0 (Fitch et al., 2001; Thorlacius et al., 2018). In the next round, panelists independently provide their judgments on these issues by answering close-ended questions developed by panel conveners and explaining their responses. Panel conveners collate responses and send an anonymized group summary back to panelists. Panelists review the summary, which includes their own original responses to all questions and provides an overview of panelists’ explanations, and then revise their original responses if needed. The process is repeated at least one more time. The final individual judgments are aggregated to determine whether consensus has been reached; sometimes, responses are also weighted based on the individual’s expertise or self-reported confidence (Murphy et al., 1998).

Figure 2.2 briefly describes the four key characteristics of the Delphi method as it was originally designed: anonymity, iteration, feedback to participants, and statistical summary. The degree to which researchers adhere to these key characteristics may vary by type of Delphi study (see below), which makes the creation of methodological guidance for how
What Delphi Is

to design and run a Delphi study challenging. Nonetheless, we believe that all researchers should consider these characteristics as fundamental principles that are at the heart of the Delphi method itself.

Delphi Panelists

The Delphi method was originally created to help researchers explore the existence of consensus among a small group of hand-picked experts. The assumption underlying the Delphi method is that asking experts the same questions more than once and allowing them to see how others answered will lead to the development of consensus. Therefore, it allows for the gradual integration of independent expert judgment into a group-based decision (Gupta and Clarke, 1996). Panelists can revise their answers after considering the perspectives of other participants whose identity they do not know. Because panelists participate iteratively and anonymously, the outcomes of a Delphi study are more likely to be objective than the perspective of one expert would be. Delphi studies have a potential to provide more reliable predictions than individual experts because the method integrates opinions of several experts with relevant scientific information in uncertain situations stemming from rapidly changing social environments or complex problems (Winkler and Moser, 2016).

Although Delphi was developed with experts in mind, Olaf Helmer, one of its creators, argued that “extreme specialists”—experts with narrow areas of expertise—may not be the best Delphi panelists (Helmer, 1949). Instead, he advocated for engaging those who have broad expertise and can take a bigger picture into account. More recently, researchers started using the Delphi method to engage with stakeholders (Dalal et al., 2011). Such panels may include both substantive experts on an issue, such as researchers or clinicians, and laypeople with lived experience, such as regular citizens or patients.

Reasons for Using Delphi

Delphi was envisioned as a forward-looking methodology to help deal with the unknown. The method’s name comes from Greek mythology. Pythia, the most famous oracle of Delphi, was known for her (often ambiguous) predictions about the future. As an expert elicitation methodology, Delphi is often used to estimate the probability of an event happening within a certain period of time or to forecast when an event is likely to occur (Murphy et al., 1998). It is also used to make group-based judgments (Rowe and Wright, 1999), such as identifying and prioritizing key problems that need to be addressed (Wicklein, 1993).

When Delphi is used for stakeholder engagement, it can be viewed as an “educational intervention” that helps panelists learn about the opinion of other participants and allows researchers to explore whether and how exposure to different opinions and opposing perspectives affects an individual panelist’s own perspectives (Khodyakov and Chen, 2020). For example, Delphi is commonly used in prioritization exercises in which stakeholders with dif-
different and often opposing preferences need to choose one or two options from a longer list (Byrne et al., 2020; Grant et al., 2020).

The iteration, anonymity, group feedback, and statistical summary that characterize the Delphi method can help identify areas of agreement and disagreement among panelists and, as a result, can either facilitate the development of consensus or confirm dissensus in diverse groups.

**Question Items and Response Scales**

Although different disciplines ask different types of Delphi questions, there is general agreement that the question items and response scales should be bias-free and readily understandable by all panelists. In forecasting studies, Delphi questions focus on the probabilities of certain events happening within a given time frame (Weaver, 1971). In health studies, researchers often use Likert-type scales with 5 to 9 points to rate treatment options on several criteria. One of the main benefits of using 9-point scales is that they can be combined with a validated measure of consensus described in the RAND/UCLA Appropriateness Method (RAM) manual (Fitch et al., 2001). However, for some studies, 9-point scales may be too granular. It is also important to note that the choice of the rating scale may affect final consensus (Lange et al., 2020). Rating the same item on multiple criteria can become too burdensome for panelists. Therefore, some researchers prefer to use ranking questions. After generating options and eliminating duplicative options, panelists rank-order them relative to each other (Okoli and Pawlowski, 2004).

Because the Delphi method is typically used to address complex questions, some panelists may have less confidence in their responses to some questions. Therefore, some studies ask panelists to rate their confidence in their responses, which is then used to weight their answers (Rowe and Wright, 1999). Nonetheless, research shows that a weighting process based on self-confidence produces less accurate results than unweighted expert input (Dietz, 1987).

**Delphi Types**

Delphi was created as a part of a classified RAND project in the late 1940s and was first described in a peer-reviewed journal article in 1963 (Dalkey and Helmer, 1963). Around the same time, researchers started modifying the method to better suit their study topics and needs (Strasser, 2017). While classical Delphi has four key characteristics (anonymity, iterative data collection, feedback to panelists, and statistical group response), common modifications to classical Delphi include the distribution of pre-reading materials, such as evidence dossiers that summarize existing evidence; in-person or virtual meetings before, during, or after the ratings that make participation not anonymous; provision of group-only (or aggregate) versus individualized statistical reports to panelists; incorporation of focus groups and
other qualitative methods into the data collection; the use of online survey platforms and discussion forums to facilitate data collection; and changes to the wording of items and rating criteria that panelists use in different rounds (Sinha, Smyth, and Williamson, 2011). The development of online technologies also helped create new Delphi types.

Below, we briefly describe four commonly used Delphi types: (1) policy Delphi and its methodological cousin, dissensus Delphi, (2) modified Delphi that was originally developed to identify appropriateness of health care, (3) real-time Delphi that does not have sequential rounds, and (4) ExpertLens that allows for implementing modified Delphi online while preserving panelist anonymity.

Policy Delphi

Classical Delphi is focused on establishing consensus and reducing variance in perspectives over multiple rounds. In contrast, policy Delphi focuses deliberately on exploring a range of opinions on a given policy topic or issue, with the objective of identifying divergent views and potential solutions to pressing policy problems that stakeholders with different perspectives can agree on (Mukherjee et al., 2015; Turoff, 1970). The goals of policy Delphi are to consider a wide range of possible policy options and to estimate the likely impact, acceptability, and consequences of each option under consideration (McGeoch, Brunetto, and Brown, 2014; Rowe and Wright, 1999). Policy Delphi studies deliberately select a large (up to 50 panelists) and diverse pool of subject-matter experts and stakeholders with potentially “extreme” perspectives (Makkonen, Hujala, and Uusivuori, 2016) and often explicitly structure panels and design questions to maximize heterogeneity in backgrounds, opinions, and responses (de Loë et al., 2016; Linstone and Turoff, 1975).

Policy Delphi often starts with a series of open-ended questions about potential policy options and their consequences. Then panelists rate importance, desirability, feasibility, and confidence in their judgment of each policy option that researchers generated based on the analysis of comments (Manley, 2013; Turoff, Linstone, and Gibran, 2002). In the next round, panelists review the rating results from the previous round and decide whether they want to change their responses. Panelists are encouraged to explain their responses and suggest changes to the way the rating items have been worded, which can result in new rating items (policy options) to be considered in subsequent rounds. Policy Delphi suggests that the Delphi method can not only develop consensus and minimize variance in perspectives but also help explore opposing perspectives that different stakeholders may have.

It is not surprising that policy Delphi led to the creation of dissensus Delphi, which helps explore variation in perspectives and identify areas of disagreement (Steinert, 2009). If a panel is large and diverse, or includes stakeholders with different perspectives, learning about things they disagree on can be as valuable as knowing where they agree.
Modified Delphi (or RAND/UCLA Appropriateness Method)

Early mentions of modified Delphi in the 1970s were synonymous with the RAND-developed “modified nominal group technique,” in which an initial anonymous ranking round is used to inform an in-person meeting (Murphy et al., 1998). In the health care field, modified Delphi is often synonymous with the RAM, which is frequently referred to as simply an “expert panel” (Fitch et al., 2001). RAM was designed to identify types of care that are overused because they are not appropriate and necessary and to identify the best alternatives. The RAND/UCLA Appropriateness Method User’s Manual (Fitch et al., 2001) described a specific way of conducting modified Delphi panels. In particular, RAM introduced five key changes to classical Delphi:

- specified and limited the number of rounds
- added an in-person discussion round between two rounds of questions, which required an experienced discussion moderator
- limited the number of panelists to 9–18 experts with different expertise
- promoted the use of 9-point Likert scales, for which it created and validated a measure of consensus
- required a review of evidence to be shared with experts before the first round of questions, which essentially replaced the open-ended round common in classical Delphi.

As originally designed, RAM panels start with a review of existing evidence and the development of clinical indications or clinical scenarios to be rated by panelists. Then a multidisciplinary group of clinical experts, including both specialists and primary care physicians, is asked to independently rate the appropriateness of each indication using a 1-to-9 Likert-type rating scale, which has labeled end points. First-round ratings are summarized in individualized reports that show the distribution of all responses, a panelist’s own responses, and whether consensus has been achieved. RAM has its own definition of consensus that accounts for the variation in the panel’s responses and the median answer. Panelists then meet in person for one or two days to discuss the indications that they failed to agree on, and a skilled moderator facilitates the discussion (Shekelle, 2004). Throughout the discussion, experts rerate the indications that were discussed. Finally, the panelists are asked to use the 9-point scale to rate the necessity of the indications that were deemed appropriate in the previous round. It is worth noting that the rating of necessity is similar to the first round of appropriateness ratings but is not included in all RAM panels.
E-Delphi and ExpertLens

Researchers originally relied on paper forms to collect panelist responses but started using email for Delphi data collection in the 1980s. They switched to online survey platforms once these platforms became more widespread. Delphi studies conducted using online survey platforms are often called e-Delphi or online Delphi panels (Heed et al., 2022). In the past 20 years, several online platforms were developed specifically for administering Delphi panels (Aengenheyster et al., 2017; Broder, Gibbs, and Yermilov, 2022), including ExpertLens, which was developed by researchers at RAND (Dalal et al., 2011).

A typical ExpertLens process includes three or four rounds (See Figure 2.3). The number of rounds is determined ahead of time. In an optional Round 0, panelists express their interest in the topic, suggest ideas, and respond to open-ended questions. Researchers synthesize their input and prepare close-ended questions that panelists answer in Round 1 and Round 3. In Round 2, panelists receive individualized reports that show how their responses to Round 1 questions compare with those of others and whether the panel reached consensus. A unique feature of ExpertLens is an online, asynchronous, (partially) anonymous, moderated discussion board that panelists use during Round 2 to share their perspectives and discuss points of disagreement. ExpertLens also has a number of built-in analytic tools that measure participant engagement, use numeric data to determine what the panel “thinks,” and employ artificial intelligence to identify why panelists agree or disagree with one another by looking at their comments. These features allow researchers to complete data collection and analysis quickly, usually within four weeks.

FIGURE 2.3
ExpertLens as an E-Delphi
Real-Time Delphi

New software and internet-based tools have also allowed for the development of real-time or roundless Delphi methods. Real-time Delphi studies do not have defined rounds. Panelists receive individualized feedback showing how their response compares with all other previously provided responses and can alter their answers as often as they want (Gordon and Pease, 2006). Feedback on the panel’s responses in the form of medians, interquartile range, and comments is continually updated to reflect the most recent responses (Aengenheyster et al., 2017). As a result, panelists receive different feedback depending on when they log into the online system.

Studies of real-time Delphi have shown that the results are similar to those of conventional Delphi methods and that real-time Delphi may reduce participation burden and attrition (Gnatzy et al., 2011). Therefore, real-time Delphi may help researchers expedite data collection and analysis and give panelists the flexibility to provide their input as often as they want. Nonetheless, the first panelists to provide their answers may not appreciate roundless approaches because they either see a simulated group response or have to wait until enough panelists answer a given question before meaningful feedback on the panel’s response can be calculated (Dalal et al., 2011). Finally, because the statistical feedback that panelists see will vary, it may be problematic for Delphi studies that consider the method to be an educational intervention.

Methodological Limitations of the Delphi Method

Although methodologically powerful, the Delphi method is not without limitations. First, Delphi is considered a “small N” method in that the participant sample is limited and not representative of a larger population (Akins, Tolson, and Cole, 2005). Delphi samples are also rarely random, because panelists are typically carefully chosen based on their expertise and experiences. Second, the quality of Delphi panel results depends on the expertise of invited panelists and their willingness and ability to articulate their perspectives using the Delphi approach (Khodyakov, Savitsky, and Dalal, 2016). Third, the Delphi method has been criticized for low replicability of its results and for being sensitive to who the panelists are (Shekelle, 2004). Fourth, the methodological rigor of the panel process, including expertise and impartiality of panel conveners and the extent of preliminary work they have conducted, could affect panel outcomes. Finally, Delphi results are sensitive to the statistical approach chosen to calculate consensus (Grant, Booth, and Khodyakov, 2018).
CHAPTER 3

Design

This chapter provides practical advice for designing a study that uses the Delphi method, and later chapters discuss how to collect data, analyze data, and report results.

The steps involved in designing a Delphi study include providing a justification for using the Delphi method; selecting the type of Delphi; assembling the research team; specifying panel eligibility criteria and a recruitment strategy; anticipating attrition and identifying retention strategies; deciding which panel features to use; conducting preliminary work, such as a literature review; and conducting a pilot test. Panel conveners have discretion at each of these steps to make design choices that (hopefully) are most appropriate for the objective of their Delphi study.

Justification for Using Delphi

The Delphi method is not as commonly used as surveys or expert interviews, often because it is more burdensome for panelists, who have to answer the same or similar set of questions more than once (Mukherjee et al., 2015; Veugelers et al., 2020). Therefore, it is important to determine whether the Delphi method is the right methodological approach for each study.

In general, Delphi is appropriate for forward-looking studies on topics for which expert opinion is the only source of evidence and for studies that explore areas of agreement and disagreement in diverse, multistakeholder panels on complex issues with no right or wrong answers (Rowe and Wright, 2011; Winkler and Moser, 2016). At the same time, Delphi may not be the right methodological
approach when the evidence base is strong; when a large, representative sample is needed to answer a study question; and in backward-looking studies that measure personal attitudes or perspectives on things that have already happened (Mukherjee et al., 2015; Steinert, 2009). For example, if a study wants to describe how stakeholders feel about an already implemented policy, survey approaches may work much better. In contrast, Delphi can be appropriate for determining the likelihood that a given policy can be implemented within the next three years, because it is forward-looking.

Selecting the Type of Delphi

Because there are different types of Delphi, research teams should determine which type is most appropriate for a given study. The research question that a study is designed to answer should dictate the type of Delphi used. The type of Delphi, in turn, is likely to affect a number of other design decisions, including the questions asked and the mode of panel administration. Some types of Delphi may be more useful for certain types of research questions. For example, both policy and dissensus Delphi studies may focus explicitly on surfacing the full breadth of views on a specific policy topic or proposal (Mukherjee et al., 2015). However, dissensus Delphi would be less well suited for research focused on reaching consensus. For example, a study producing a core outcome set may be better served by classical or modified Delphi. Moreover, a study that is interested in the extent to which the exposure to different perspectives affects panelists’ own perspectives should not be using real-time Delphi, because panelists will be exposed to different “educational interventions” if they log on to an online platform at different times.

Assembling the Research Team (Roles, Responsibilities, and Expertise)

It is helpful to clarify the roles and responsibilities of the research team, including researchers running the panel, or panel conveners, in advance in ways that account for their expertise and availability. The research team should include those with expertise in the topic under investigation and those with expertise in the Delphi method. Moreover, if Delphi includes a discussion round, moderators should be carefully chosen based on their ability to encourage sufficient and fruitful participation among panelists while avoiding dominating the discussion (Fitch et al., 2001; Khodyakov,
Grant, et al., 2020). Similarly, researchers developing evidence dossiers for panelists ideally should have expertise in systematic review methods (Fitch et al., 2001; Murphy et al., 1998). For larger projects, consider a dedicated project administrator to facilitate communication with panelists across rounds. The research team also needs to decide whether its members can be panelists (e.g., if there are a limited number of subject-matter experts in the topic area) or pilot testers, or will refrain from participating. In general, it is better for research team members not to serve as panelists to avoid potential bias.

**Specifying Panel Eligibility Criteria**

Panel composition can significantly influence Delphi study results and their validity (Shekelle et al., 1998). High-quality Delphi panels should strive for diverse perspectives and experiences. Panel conveners should consider the type and level of expertise, demographic and geographic diversity, and potential conflicts of interest of potential panelists (Kahan et al., 1996). Depending on the panel topic and goals, *expertise* may be defined broadly to include lived experience. For example, clinical panels commonly include specialists (including those who perform a certain procedure and those from related specialties who do not perform such procedures) and primary care providers but not patient representatives (Kahan et al., 1996), whereas public health–focused panels commonly can engage researchers, clinicians, and patients (Bodnar et al., 2021; Kim et al., 2018). The panel composition should reflect the purpose of the Delphi study. For example, studies aiming to examine areas of (dis)agreement and (un)certainty should recruit panelists representing a wide range of perspectives (Concannon et al., 2012). Panel conveners should identify key characteristics they want to ensure diversity on before assembling the panel.

**Identifying Recruitment Strategies**

Delphi panelists are generally selected purposefully rather than randomly to ensure that they have the required expertise and experience. Random selection may be helpful if more potential panelists have expressed interest than are needed for the panel. Most common selection criteria include confirmed or self-identified expertise in the topic, organizational or institutional affiliation, nominations from colleagues, and previous publications on the topic (Spranger et al., 2022).

For recruitment purposes, panel conveners should use a multipronged approach, including direct outreach to subject-matter experts whom the team has identified (through, e.g., a literature review, colleague recommen-
dations) and snowball sampling, in which experts who are contacted nominate other experts to be approached (Rowe and Wright, 2011). Personalized outreach by a study principal investigator can facilitate recruitment and retention (Keeney, Hasson, and McKenna, 2006; Landeta, 2006). Outreach to professional societies and the use of social media platforms, such as LinkedIn and Facebook, may help cast a wider net and recruit a more diverse group of panelists (Belton et al., 2019). However, it is important to ensure that potential panelists have the required experience and expertise by reviewing their CVs and publication records or asking them to describe their relevant experience and expertise.

Using Participant Retention Strategies

Panelist nonresponse and attrition are common in Delphi studies; in fact, it is not rare for half or more panelists to drop out (Elwyn et al., 2006; Hsu and Sandford, 2019). Although common, substantial participant drop-out may negatively affect the quality of Delphi study results and therefore should be minimized. Research teams may consider offering incentives to help retain panelists, such as participation honoraria, acknowledgment as panel members, and co-authorship of manuscripts based on the data collected. However, a 2021 review of Delphi studies found that just 8 percent described using incentives of any kind (Barrington, Young, and Williamson, 2021).

If offered, financial incentives should be consistent with the time panelists will be required to spend; all panelists should receive the same amount to signal that everyone’s expertise is equally valued. Projects that cannot afford to pay all panelists may consider using lotteries to randomly pick a small number of panelists who completed all Delphi rounds to receive a prize (Khodyakov et al., 2016).

Professional recognition may also be useful for retention. Being nominated by a professional society or peers has also been shown to increase participation and retention rates because people often view it as an honor to be asked (Fitch et al., 2001). Panelists may also appreciate official recognition in public documents (Keeney, Hasson, and McKenna, 2001; Rowe and Wright, 2011) or an opportunity to serve as manuscript co-authors if there are no conflicts of interest and they meet co-authorship criteria.

Researchers should think carefully about the right incentives and keep in mind that people who have a strong interest in the study topic might be more motivated to participate than those who are just interested in financial rewards. Regardless of the chosen approach, researchers should document their use of incentives and other efforts to encourage participation and retention.
Below, we summarize recommendations that the Delphi literature and our experience suggest can help minimize participant attrition (Gargon et al., 2019; Hall et al., 2018; Hsu and Sandford, 2019; Khodyakov, Grant, et al., 2020; Khodyakov et al., 2011):

**RECOMMENDATIONS**

- Recruit more panelists than needed.
- Address panelists by name in all study communications.
- Explain what the Delphi method is and consider using instructional videos because the Delphi method is often confused with surveys, which are generally not iterative.
- If emails are used for communication, ensure that panelists actually receive them before the start of data collection.
- Choose an easy-to-use e-Delphi platform for data collection if feasible.
- Limit the number of questions asked and ensure that they are clear to panelists.
- Do not conduct too many Delphi rounds.
- Select the right number of panelists; too many can be burdensome for panelists, and small panels, especially those conducted online, may not be engaging enough.
- Keep panelists informed about any changes in the study timelines and send reminders to those who have not provided their responses yet.
- Decide whether to allow those who missed a previous round to participate in any subsequent rounds.
- Use financial and nonfinancial incentives.

While these recommendations are useful, the key is two-fold: (1) invite people who truly care about the topic and have the right expertise and experience and (2) design the least burdensome study possible. If the questions do not make sense or the online platform used for data collection is not intuitive, panelists will drop out, even if they are offered financial incentives for participation.

**Conducting Preliminary Work**

Before starting the first Delphi round, panel conveners should decide whether they will share a summary of existing evidence with panelists to help inform their answers. Literature reviews (or evidence syntheses) are very common in health and medical research fields (Fitch et al., 2001), but they are less common in Delphi studies conducted by other disciplines. Although systematic literature reviews can be time-consuming, summarizing what is known about the topic is particularly helpful in panels that involve
a mix of diverse stakeholders to promote a shared, baseline understanding of the issue at hand. Using infographics is an effective way to communicate complex information to diverse audiences (Radomski et al., 2022).

Moreover, it is not uncommon to ask future panelists for their thoughts on the study topic in an open-ended manner before the first round of structured questions (Claassen et al., 2014; Murphy et al., 1998). Doing so not only helps engage panelists and ensure their input shapes the wording of questions in subsequent rounds but could also serve as a way to recruit panelists, collect their demographic data, and ensure that they have the right type of experience and expertise (see appendix of Miller et al., 2023).

Conducting a Pilot Test

Conducting a pilot test with a small group of panelists who are similar to those who will be invited is important to ensure that the data collection process, instructions, and questions are clear (Landeta, 2006). This process can also help mitigate issues that would otherwise hinder retention, such as reducing the number of items to be rated. In the pilot test, panelists should be provided with instructions, the list of items to be assessed and questions asked (e.g., rating criteria to be used), and/or a link to the programmed questionnaire so they can test the proposed process and suggest revisions. Interviewing pilot testers to solicit their feedback directly may also be useful (Khodyakov, Kinnett, et al., 2020). Panel conveners should decide how pilot testers’ suggestions will be addressed and how conflicting recommendations will be resolved. If pilot testers are participating in an e-Delphi, it is important to ensure that they use different Internet browsers. Administering the pilot several weeks before the Delphi study starts should provide enough time to make necessary changes.

Deciding Which Study Features to Use

Panel conveners must make several decisions about study features: whether the panel should take place in person or online, the number of rounds, how long to keep each round open, the number of participants and panels needed, how to ensure a diversity of expertise and experience among panelists, how much feedback to provide to the panelists, the level of anonymity the study should involve, and criteria to determine when the rounds are complete.

Format

Delphi studies can be conducted using different modes of administration or formats, including pen and paper, telephone, email, and internet (de Loë et al., 2016). The specific format
typically depends on the researchers’ needs and study design. For example, pen-and-paper Delphi studies are most convenient when all data collection activities take place during an in-person meeting. This mode of data collection generally assumes a small panel and in-person discussion. Researchers should account for the time needed to analyze data and show the panel results before asking the panelists to revise their original answers.

Although Delphi was created in the late 1940s–early 1950s, its creators relied on virtual modes of communication from the very beginning. For example, in the late 1940s, Olaf Helmer used teletype to send Delphi questions from RAND’s Santa Monica office to its Washington, D.C., office to be distributed to participants located there on paper (Helmer, 1949). In the 1980s, he used email to collect data from Delphi panelists. By now, online panels have become the norm. They provide a variety of benefits, including making it more practical to engage larger and more diverse groups, speeding up the elicitation process, and minimizing burdens for panelists (Boel et al., 2021; Grant, Armstrong, and Khodyakov, 2021). One concern with e-Delphi, however, is that all communication with panelists takes place via email, so undeliverable emails, spam filtering, and inattention to unexpected emails may negatively affect response rates (Boulkedid et al., 2011; Gargon et al., 2019). E-Delphi may also have upfront costs, such as licensing, time to learn how to use the platform, and staff time to program and test the Delphi study (Belton et al., 2019).

Nevertheless, the pervasiveness of email and a variety of relatively cheap and easy-to-use online survey and video conferencing platforms make online Delphi implementation a compelling option (Broder, Gibbs, and Yermilov, 2022). Moreover, certain types of Delphi studies (e.g., real-time Delphi) can only be conducted online (Gordon and Pease, 2006). Some online Delphi platforms, such as ExpertLens, can accommodate discussion between panelists while still maintaining anonymity and can automate data analysis and preparation of individualized panelist reports, which expedite data collection and reduce the costs of convening large-scale national and international panels (Dalal et al., 2011). Moderated discussions can also take place virtually using online conferencing platforms, such as Zoom.

**Number of Rounds**

Panel conveners should either specify the number of rounds or identify clear stopping criteria (consensus, stability, etc.) in advance and be clear about what is driving that decision. If the panel does not use a stopping criterion, the number of proposed rounds could be motivated
by budgetary constraints, cost-benefit analysis, concerns about psychological factors, a desire not to force consensus, panelist attrition, or other reasons (von der Gracht, 2012).

Although a few published studies have high response rates into a fourth or fifth round, researchers have noted panelist fatigue and an unwillingness to continue Delphi processes as rounds progress (Belton et al., 2021; Keeney, Hasson, and McKenna, 2006). Research also demonstrates that additional rounds lead to attrition and could encourage the development of forced consensus (that is, a situation in which panelists whose perspectives differ from the consensus will change their responses just to complete the Delphi process; Humphrey-Murto and Wit, 2019).

Even though strict guidelines on the number of rounds do not exist, Delphi studies often consist of two or three rounds (Landeta, 2006; Nowack, Endrikat, and Guenther, 2011). The number may also depend on whether the wording of questions or items remains the same throughout the study and whether the items for which consensus has been achieved are excluded from subsequent rounds. While fewer rounds may lead to higher overall response rates and thus lower attrition, several studies suggest that motivated panelists, easy-to-complete surveys, and ongoing study reminders can help maintain high response rates even with a higher number of rounds (Veugelers et al., 2020).

**RECOMMENDATION**

Ask the same questions at least twice to allow panelists to know what others think.

**Whether to Include a Discussion Round**

It is important for panel conveners to determine whether the Delphi study will include a discussion round. Discussions are useful for identifying the cause for lingering disagreements, highlighting any definitional or interpretation differences, improving information exchange between panelists, and ultimately resolving differences in perspectives (Murphy et al., 1998; Rowe and Wright, 2011). Some researchers have advocated for using “hybrid” methods that combine Delphi questionnaires with focus groups and the nominal group technique (Landeta, Barrutia, and Lertxundi, 2011), which have their own rules for managing group dynamics and provide more structure to discussions.

If the panel includes a discussion round, panel conveners should decide when the discussion will take place and how it will be conducted. The timing of the discussion round can vary, and there can be more than one discussion round, depending on the study needs. Some Delphi studies schedule discussion rounds before the rounds of questions to engage with the panelists, explain the panel goals, and fine-tune questions to be included in subsequent rounds (Fitch et al. [2001] describe the use of a Round 0 panel meeting). ExpertLens panels include discussion rounds between the rounds of questions to discuss points of disagreement (Dalal et al., 2011; Khodyakov et al., 2011). Finally, some studies have discussion rounds at the very end to share the panel results and resolve any remaining uncertainties or outstanding issues (Boulkedid et al., 2011).
Delphi discussion mode can also vary. Panel conveners could conduct discussions in-person, by telephone, using video conferencing platforms, or using an online discussion board. It is worth noting that in-person and virtual meetings with video capabilities cannot preserve panelist anonymity (a key characteristic of the Delphi method) and may be subject to the negative effects of dominant personalities and groupthink (Dalal et al., 2011). In-person meetings are also costly, hard to schedule, and typically limited to no more than 18 panelists. Discussion rounds also require careful moderation to create a comfortable and safe sharing environment and ensure productive use of the meeting time (Fitch et al., 2001; Murphy et al., 1998). Ideal discussion moderators should have a broad understanding of the discussion topic and know how to manage group dynamics. The RAM manual recommends that a subject-matter expert (typically a physician in RAM panels) moderate the discussion and that another person with methodological expertise run the Delphi panel if possible.

Online modified Delphi platforms, such as ExpertLens, that offer online discussion boards can help address two key downsides of in-person discussion rounds: lack of anonymity and scheduling challenges (Dalal et al., 2011). For example, ExpertLens panelists use a threaded discussion forum that is accessible at their convenience. They can start their own discussion threads, respond to comments posted by other panelists or discussion moderators, and react to comments with thumbs-up or thumbs-down buttons at any time the discussion round is open. Such online discussions still benefit from having an experienced moderator who can ask clarifying and unbiased discussion questions, point out divergent perspectives, and facilitate asynchronous engagement and productive exchange of ideas (Khodyakov, Grant, et al., 2020).

**RECOMMENDATION**

- Recruit discussion moderators who have a broad understanding of the discussion topic and know how to manage group dynamics.

**Length and Timing of Each Round**

Although very little specific guidance exists on the length and timing of a Delphi round, past research has suggested that panelists are generally given from one to eight weeks to complete each round (Belton et al., 2021; Dalal et al., 2011). Panel conveners should be careful not to rush panelists. High time pressure affects decisionmaking and may lead to a group or individual focusing just on completing the task and may result in forced consensus. More moderate time pressure may allow for more focus on output quality and on careful review of the information (Murphy et al., 1998). Aspects of the Delphi study should help determine the right timing for each round. For example, if panelists are expected to spend hours reading literature review documents, directions, and other resources, researchers should provide additional time for these tasks.

Round length also depends on the complexity of the task, number of panelists and questions, and format. Empirical research on the Delphi method shows that factors related to complexity, including the number of questions and the panel size, are significantly linked to
higher participation burden and can lead to lower response rates (Gargon et al., 2019). Too much time between rounds and rounds that last too long may cause panelists to lose motivation and interest as the rounds drag out—just as a greater number of rounds causes panelist fatigue (Veugelers et al., 2020).

Online panels may require less time to complete each round. For example, ExpertLens rounds are usually open for about 7–14 days, which includes situations in which rounds have been extended to give panelists more time (Dalal et al., 2011). Panelists being asked to review and respond with detailed comments may appreciate having extra time to complete the task. When providing pre-reading materials and evidence dossiers, panelists should be encouraged to take their time and carefully review the documents and should be given additional time to complete the required tasks (Fitch et al., 2001).

The time it takes to complete a Delphi study also depends on the time between rounds. Research shows that three to six weeks should be enough to analyze data and prepare individualized reports for each panelist (Belton et al., 2019; Fitch et al., 2001). Researchers should consider using online Delphi platforms that can automatically analyze the data to reduce or even eliminate the time between rounds.

**Number of Panels**

The purpose of the study should determine the number of panels. Different stakeholder groups may have different opinions on a given topic, and those can either be analyzed separately or together. If the goal is to highlight commonality within groups and conflicting views between groups, having one panel for each group of subject-matter experts or stakeholders can accomplish this. If the goal is to examine the existence of consensus between different groups, then using one mixed panel may be beneficial. It should be noted, however, that ensuring equal representation of different stakeholder groups in one mixed panel is important (Sinha, Smyth, and Williamson, 2011).

Nonetheless, a key criticism of Delphi is that its results are hard to reproduce, depending on who is on the panel (Shekelle et al., 1998). To increase validity of Delphi panel results, panel conveners should consider conducting at least two concurrent Delphi panels that use the same data collection protocol and randomly assign panelists to one of the two panels to increase the validity of Delphi results.

**RECOMMENDATION**

- Conduct at least two concurrent Delphi panels that use the same data collection protocol and randomly assign panelists to one of the two panels to increase the validity of Delphi results.
Criteria for Panel Homogeneity and Heterogeneity

Assembling diverse panels is important because results may vary substantially between groups, diversity of perspectives can help fill blind spots a given panelist may have, and representation of all relevant perspectives can help with future dissemination, relevance, and acceptance of the research results. Researchers often suggest assembling diverse panels along one of two dimensions—diversity of specialty or expertise and diversity of represented stakeholder groups—to solicit a wide range of opinions and perspectives on a given topic (Belton et al., 2021). Some panels may have only academic experts but be diverse in that the experts come from different disciplines (Fitch et al., 2001). Alternatively, panel diversity can be achieved by including a broad range of stakeholder groups that would be affected by the study’s results. For studies that focus on patient-relevant outcomes, Humphrey-Murto and Wit (2019) recommend engaging with at least two or three stakeholder groups. For example, a Delphi study that focuses on creating a core outcome set for a given disease could include clinicians, researchers, patients, payers, and policymakers.

Panelist characteristics such as age, gender, education, and employment may also be important when the study question or objective is expected to vary on these characteristics (Boulkedid et al., 2011).

Although panel diversity is helpful for ensuring generalizability of panel results (Nasa, Jain, and Juneja, 2021), consensus or a meaningful middle ground may not be achieved when panelists have opposing views (Belton et al., 2019). Moreover, the impact of panel diversity on its outcomes is uncertain. Some studies on panel diversity seem to suggest that laypeople in mixed panels are more likely to be persuaded by the perspectives of experts and change their original responses, whereas experts are more likely to change their responses in expert-only panels (Hussler, Muller, and Rondé, 2011; Khodyakov et al., 2022). However, response changes are more likely to occur on preference-sensitive topics for which there are several viable alternatives, and the degree of association between panel composition and response changes depends on the topic (Khodyakov et al., 2022).

Therefore, the appropriate level of panel heterogeneity depends on the study objectives. Studies interested in ensuring the validity of panel conclusions should ensure that all relevant stakeholders have been engaged and diverse perspectives have been considered (Murphy et al., 1998). One concern when assembling heterogeneous panels is that group or individual status can influence the panel dynamics. Although participant anonymity is designed to help address this concern, steps should be taken to try to mitigate

RECOMMENDATION

Include all relevant stakeholders to ensure the validity of panel findings.
the potential influence of more powerful or vocal panelists during in-person meetings or other situations in which anonymity cannot be achieved.

Panel Size
The number of panelists should be determined by the mode of data collection, the level of panel heterogeneity, and how panelists will interact. Delphi studies that include a round of in-person discussions are typically limited to 9–18 people to ensure productive dialogue (Fitch et al., 2001). In contrast, online panels that do not have a discussion component can include hundreds of panelists (Niederberger and Spranger, 2020). Virtual panels with an online discussion component, such as those administered using ExpertLens, typically include 40–60 panelists (Khodyakov et al., 2011). Conducting larger panels online allows for expanding the type and number of panelists. This means that, for example, for a panel focused on health care, panel conveners can invite clinicians and patients to share their perspectives because their participation can remain strictly (or partially) anonymous. Given that attrition is common in Delphi panels, the final decisions of larger panels are more likely than small panels to be informed by a large-enough number of panelists and reflect diverse perspectives.

Controlled Feedback
Individualized feedback reports that show how a given panelist’s responses to each question compare with those of other panelists are an important characteristic of the Delphi method. While high-quality Delphi studies provide each panelist with such reports, panelists often get to see only the group’s responses from a previous round. This is particularly true when researchers use traditional survey platforms for data collection that lack a built-in analysis feature that can create individualized reports (Donohoe, Stellefson, and Tennant, 2012). Because it is hard for panelists to recall all their answers, it is crucial to remind panelists of their earlier responses before asking them whether their perspectives changed after seeing how other panelists answered the same questions.

Individualized reports should provide a statistical summary of the panel’s responses to the questions asked in the previous round and include an appropriate measure of central tendency and dispersion. It is also important to indicate whether consensus has been achieved for each question asked. High-quality reports also describe why panelists chose a particular response. In some studies, it might be helpful to highlight perspectives that differ from the consensus.
Level of Anonymity

Panel conveners should determine the level of panelist anonymity before enrolling Delphi panelists. While anonymity is one of the Delphi method’s defining characteristics, many Delphi modifications eliminate or reduce the extent of anonymity guaranteed to panelists. In classical Delphi studies, anonymity is guaranteed to all panelists but not to the study staff and researchers. Some online Delphi software can keep responses completely anonymous even from the study staff (Gnatzy et al., 2011). Researchers should decide whether this degree of anonymity is advantageous to their project.

Anonymity in modified Delphi studies is often limited to responses to survey questions if panelists engage with each other during in-person meetings or workshops to discuss points of agreement or disagreement. Moreover, the list of panelists or the organizations they represent is often included in publications detailing Delphi findings, and panelists sometimes are invited to become co-authors of the publications (de Loë et al., 2016). While panelists may not know who the other panelists are during the data collection, these approaches reduce the level of anonymity by the end of the study.

Finally, panel conveners may opt for partial panelist anonymity. For example, ExpertLens panelists receive unique IDs that may include their stakeholder group or other relevant information but not their real name. A 2019 evaluation of ExpertLens panels revealed that panelists in a mixed, partially anonymous panel that included both patients and caregivers wanted to know more about the other panelists. They suggested making partially anonymous profiles with information such as a patient’s age, disease stage, and geographic location visible to other panelists so they can better understand and contextualize the person’s responses (Armstrong et al., 2019). This approach allows for tracking a given panelist’s comments and provides some potentially useful information without revealing their identity.

Stopping Criteria

Stopping criteria determine when to conduct the final Delphi round and present final results. Besides completing a predetermined number of rounds, which is the approach that we recommend, the most common stopping criterion is consensus or agreement. As noted above, a predetermined number of rounds can be dictated by study timeline, budget, or other consid-
erations. The advantage of relying on this approach is that it allows for identification of both consensus and disagreement, assuming that there was no change in the data collection protocol throughout the Delphi process and all questions were asked in all rounds. Studies that predetermine the number of rounds, however, should develop procedures for dealing with items that do not reach consensus after multiple rounds (Jünger et al., 2017).

We generally do not recommend using consensus as a stopping criterion because it can lead to the development of a forced consensus (Mukherjee et al., 2015), which can hide disagreements within a group that may be relevant to the study (Jandhyala, 2020). However, if panel conveners use consensus as a stopping criterion, they should specify the definition of consensus before the start of the study (Boulkedid et al., 2011; Grant, Booth, and Khodyakov, 2018; Jünger et al., 2017). The advantage of using consensus as a stopping criterion is that it ensures there is agreement on all topics at the end of the study.

Stability—when the results of two subsequent Delphi rounds are not significantly different—is also sometimes used as a Delphi stopping criterion because it does not force consensus and focuses on changes to responses across rounds (Holey et al., 2007). Therefore, it is directly linked to iteration, which is a key characteristic of the Delphi method. If stability is used as a stopping criterion, panel conveners should prespecify the operational definition of stability as a metric for determining when no additional rounds of Delphi will be conducted. Research suggests that three rounds is generally sufficient for a clear pattern to emerge from panelists, though measuring stability explicitly is more reliable and may allow for fewer rounds if opinions do not change much between rounds one and two (Belton et al., 2019).

RECOMMENDATION

- Decide on the number of Delphi rounds before the start of data collection.
CHAPTER 4

Data Collection

Collecting data for a Delphi study involves making a series of decisions. What items will panelists be asked to rate? How should questions be worded, and how many should there be? What rating scale will be used? Should questions be modified between rounds? Should there be a discussion moderator? For e-Delphi panels, what software will be used? This chapter offers some practical considerations for making these decisions, based on our experience and the research literature.

Items

It is common for Delphi studies to ask panelists to react to a series of statements or scenarios (collectively referred to as items) by answering one or more questions about each one. In policy Delphi studies, panelists may be asked to consider a series of candidate policies (de Loë et al., 2016). In RAM studies, panelists rate a series of clinical vignettes or scenarios describing different patients to determine which of the proposed treatment options might be most appropriate and necessary (Fitch et al., 2001). In forecasting studies, panelists are asked to rate the likelihood of different hypothetical events occurring, the severity of their impact, and the time frame within which they may occur, with the eventual goal of clustering results to produce different scenarios (Di Zio, Bolzan, and Marozzi, 2021; Rowe and Wright, 1999).

The items used in Delphi questionnaires vary as much as the topics being studied with the method. In general, the items are simple statements or vignettes that provide some context, which panelists are asked to rate, rank-order, or answer close-ended questions about. The response options should be numeric so that the data can be analyzed quantitatively. For example, in a study on patient-centeredness of clinical guidelines for Duchenne muscular dystrophy, we asked patients and their caregivers to use 9-point Likert-type scales to rate a series of recommendations (items), such as “Recombinant human growth hormone should be reserved for treating children 7 years of age or younger only if they had abnormal growth
hormone stimulation tests due to the inconclusive evidence about its effectiveness” on two rating criteria (questions): importance and acceptability for a typical individual/family with Duchenne (Khodyakov et al., 2019).

Panel conveners should consider asking the same questions about all items considered by the panelists to allow for comparing and contrasting the items on one or more dimensions. Moreover, asking the same questions can simplify the process of data collection because panelists are not forced to switch between different types of rating scales or answer different types of questions about different items (Fitch et al., 2001).

Questions

Although Delphi studies include self-administered questionnaires, panels should not be called surveys because surveys generally are not iterative and do not include discussion rounds. Our experience suggests that calling Delphi studies surveys increases panelist confusion and negatively affects participation rates. It is true that each round of questions is essentially a survey, Delphi studies are complex interventions that contain multiple parts, including literature reviews, questionnaires, group feedback, and discussion sessions, among others. Nonetheless, a lot of best practices for survey design apply to Delphi studies (Krosnick, 2018).

Delphi questions should be specific, clear, unambiguous, and unbiased; response options should be exhaustive and relevant to the questions asked. As discussed earlier, panel conveners should limit the number of questions to minimize attrition (Landeta, 2006) and pilot test the questionnaire to estimate panelist burden and identify unclear or repetitive questions. Such questions are often good candidates for deletion, given that people tend to skip questions that they do not understand, consider repetitive, or think they do not have enough expertise to answer. Therefore, it is important to ensure that the questions are not duplicative, that they measure unique constructs, and that the right panelists have been recruited.

As mentioned in prior chapters, Delphi studies often ask open-ended questions in the initial round (Keeney, Hasson, and McKenna, 2006; Veugelers et al., 2020) and include space for panelists to explain their answers (Belton et al., 2021), provide counter-arguments, expand on their answers, or otherwise clarify their responses (Nowack, Endrikat, and Guenther, 2011; Veugelers et al., 2020; Winkler and Moser, 2016). In some Delphi studies, panelists in subsequent rounds are asked to explain their answers only if they fall outside of a certain range of responses; this is particularly the case in real-time Delphi, where the software can imme-

RECOMMENDATION

▲Ask the same questions about all items considered to compare and contrast panelists’ views along several dimensions.

▲Ask specific, clear, unambiguous, and unbiased questions; provide response options that are exhaustive and relevant to the question.
Data Collection

Diately flag these responses (Gordon and Pease, 2006). Therefore, panel conveners should decide whether and how to ask open-ended questions.

Handling “I Do Not Know” Answers

Debates about including “do not know” or “do not have sufficient expertise” response options commonly arise during the Delphi design stage, and some researchers support including a neutral option (Belton et al., 2019). While it may be a best practice in survey design to ensure that response options provide panelists with a way to accurately respond to all questions, it may not always be a good approach for Delphi studies that are conducted when there are no right or wrong answers. Delphi panelists are experts on the topic or stakeholders with lived experience and should be able to answer all questions. If panelists are not comfortable answering some questions in the initial round out of concern that they might provide an incorrect response, they should be reminded that they can revise their responses in subsequent rounds. Therefore, the research team should emphasize that panelists should provide their “best guess” in the initial round and then revise their responses later if needed. Another option is to allow for nonresponse to items and then communicate the level of response to each item in the results (Jünger et al., 2017).

Sometimes, a middle point on a rating scale with an odd number of points can be used as a neutral response category or as a way to provide a “do not know” response. However, this approach is contested: Some research recommends a scale with an even number of points to “force” panelists to provide a positive or negative answer (Akins, Tolson, and Cole, 2005), whereas other studies caution that this approach removes the option to convey genuine uncertainty or a lack of expertise on that specific item (Drumm, Bradley, and Moriarty, 2022).

Response Scales

The number of points on a response scale is also an important consideration. All else being equal, having fewer response options increases the likelihood of consensus development. However, consensus in this case may be an artifact of a limited number of response options rather than a reliable assessment of panelists’ views on the question. Therefore, panel conveners should carefully think about what questions and response scales to use (Lange et al., 2020). In health care panels, researchers often prefer to use 9-point Likert scales; other disciplines have also used these scales in part because RAND’s RAM manual provided a validated approach (Fitch et al., 2001; Niederberger and Spranger, 2020). However, 5-point and 7-point Likert scales are also commonly used in Delphi studies.

RECOMMENDATION

Do not include “I do not know” as a response option.

RECOMMENDATION

Choose what response scales to include as they may affect consensus development.
Modifications to Items and Questions

In classical Delphi studies, questionnaires generally remain unchanged between rounds after items, questions, and response options have been adjusted based on pilot testers’ feedback. However, in modified Delphi studies, panel conveners sometimes edit items, questions, or response options between rounds (Moher et al., 2010). Most commonly, they add or subtract from the list of items considered in each round.

When adding items, panel conveners review panelist suggestions about items that they might have overlooked, consolidate similar responses, and then add items that received sufficient support and are not too similar to an existing item. Alternatively, panel conveners sometimes decide to remove items between rounds if consensus has already been reached on them. While doing so can reduce panelist burden in subsequent rounds and is a common practice in RAM panels, it is important to remember that such an approach violates the main characteristic of the Delphi method: iteration. Some argue that asking a question only once prevents panelists from learning about other perspectives, which can introduce “a structural bias to the consensus generating process” (Jandhyala, 2020, p. 1875).

The decision to modify the data collection protocol between Delphi rounds may depend on the type of Delphi conducted and the study needs. Nonetheless, panel conveners should decide whether they will be making any changes before the start of data collection to avoid potential bias.

Discussion Moderation

One common modification to classical Delphi is the addition of moderated discussion sessions. By facilitating the exchange of ideas and perspectives, discussion moderators play a crucial role in a modified Delphi process (Murphy et al., 1998). One of their main tasks is to prevent the development of a forced consensus by actively communicating dissenting opinions and allowing dissenters to share their views, acknowledging that dissensus is a valid outcome of the process, and clamping down on groupthink (Mukherjee et al., 2015). According to the RAM manual (Fitch et al., 2001), discussion moderators should understand how the Delphi process works, have relevant subject-matter expertise, review the results of the previous Delphi rounds, have access to the answers provided by each panelist, and prepare discussion questions that help explore areas of agreement and disagreement.

Until recently, Delphi discussions were typically conducted in person, but the moderator’s role is simi-
lar in online Delphi platforms such as ExpertLens. Whether online or in person, moderators encourage conversations and panelist engagement, intervene when conversations drift off topic or become unproductive, and ask neutral probing questions to encourage the sharing of all opinions within a group (Khodyakov, Grant, et al., 2020).

Software Used

Conducting Delphi panels online is convenient to both panel conveners and panelists themselves and can expedite data collection and analysis. Nowadays, online survey platforms, such as Qualtrics, SurveyMonkey, REDCap, and SelectSurvey, allow researchers to easily field Delphi questionnaires. However, panel conveners should keep in mind that these online survey tools are not Delphi platforms. They do not necessarily have an ability to remind panelists about their responses to questions asked in previous rounds or generate individualized panelist reports.

The choice of the software used may also depend on whether and how researchers want Delphi panelists to interact with each other. Online survey platforms typically do not have the discussion functionality, but hybrid Delphi studies can be convened using a combination of online surveys and virtual video meetings (Broder, Gibbs, and Yermilov, 2022).

Nonetheless, several software tools exist specifically for conducting e-Delphi studies that may include the discussion functionality (Aengenheyster et al., 2017; Dalal et al., 2011). Because these tools have different functionalities, panel conveners should determine the pros, cons, and relative value of different platforms when choosing the software to use. Some factors to consider include research team preferences, level of methodological and technical support provided, available question types, ability to output data in formats that can be easily analyzed, built-in analytic features, overall user-friendliness, platform accessibility, ability to amend items or feedback, availability of a discussion functionality, ability to produce individualized reports, levels of panelist anonymity, data privacy policies, and costs.

RECOMMENDATION

Review available Delphi software packages and their functionalities before choosing one that meets the study’s needs.
Data Analysis

To avoid the appearance of bias, panel conveners should prespecify their data analysis plan. The plan should include simple descriptive statistics to communicate the results; it should also detail how the research team will examine the stability of the responses for each item, measure consensus and disagreement, and analyze panelist comments.

Descriptive Statistics

Descriptive statistics are the most common way of summarizing Delphi results that are shared with panelists and included in the final study reports. Researchers typically report a measure of central tendency and dispersion along with a frequency distribution (e.g., histogram) for each item in each round. At a minimum, panel conveners should run simple descriptive statistics to communicate the results. The choice of the measures used should be determined by the type of questions asked. For example, for questions asking panelists to state the likelihood of an event happening in the next 20 years using a 0–100 percent scale, means and standard deviations may be appropriate measures of central tendency and dispersion, respectively. For 9-point Likert-type scales commonly used by RAM panels, medians and interquartile ranges may be more appropriate. Moreover, panel conveners should ensure that all panelists understand the meaning behind the descriptive statistics they are shown. Explanations are particularly important in diverse panels because they can help build panelists’ research and engagement capacity (Khodyakov, Grant, et al., 2020).

RECOMMENDATION

Show measures of central tendency and dispersion along with a frequency distribution of responses to each question to panelists.
Stability

Panel conveners should investigate the stability of responses for each question between rounds. Some Delphi methodologists even advocate for the use of group stability (i.e., when the results of two subsequent Delphi rounds are not significantly different) as the best criterion for stopping a Delphi study and then examining consensus. For example, identifying stable disagreement across rounds can provide informative insights and highlight differences in expert perspectives. Numerous analytic approaches exist to examine this construct, including both descriptive and inferential statistics. Common options include nonparametric tests (e.g., binominal, chi-square, Wilcoxon matched-pairs signed-ranks) for independence of responses and examining changes in measures of dispersion (e.g., percent change in rating distributions) from round to round (Dajani, Sincoff, and Talley, 1979; von der Gracht, 2012). The research team should also consider examining whether any variables are associated with (lack of) stability across items; for example, the research team could examine whether one stakeholder group has stable ratings and another does not.

Consensus

Because Delphi is considered a consensus methodology, panel conveners should choose the right measure of consensus, especially in studies where it is used to determine whether additional Delphi rounds are needed. However, there is no agreement in the research literature on how best to measure consensus. Von der Gracht (2012) provides a comprehensive overview of consensus measurement in Delphi studies, including descriptive statistics (e.g., certain percentage of agreement, average percentage of majority opinions, cutoffs for measures of central tendency and dispersion), inferential statistics (e.g., chi-square test for independence, McNemar change test, Wilcoxon matched-pairs signed-ranks test, intraclass correlation coefficient, kappa statistics, Spearman’s rank-order correlation coefficient, Kendall’s W coefficient of concordance, t-statistics, F-tests), and nonquantitative measures (e.g., stipulated number of rounds, subjective analysis, postgroup consensus).

Our team tested five broad types of consensus measures typically used in Delphi studies: median alone, mean alone, median with a measure of dispersion, mean with a measure of dispersion, and level of agreement. We found that different measures lead to drastically different findings, ranging from saying that consensus was reached on almost all questions to saying that consensus was not reached on any of them (Grant, Booth, and Khodyakov, 2018).

RECOMMENDATION

Investigate the stability of responses for each question between rounds using nonparametric tests or measures of dispersion.

RECOMMENDATION

Choose and prespecify a measure of consensus that is appropriate for the type and goal of the study’s Delphi panel.
To avoid any appearance of data fishing, prespecifying the types of consensus measures to be used before data collection begins is very important.

Comment Analysis

Conveners of Delphi panels that encourage panelists to explain their responses should group the comments according to their numeric responses and identify major themes to reduce the amount of information panelists are asked to review and help them better understand why other responses differ from their own answers (Khodyakov et al., 2019; Nowack, Endrikat, and Guenther, 2011; Veugelers et al., 2020).

Qualitative analysis should be conducted if any open-ended questions are asked, explanations for responses are requested, or a discussion round has been conducted. Qualitative analysis can be done using several methods, including content analysis, thematic analysis, or qualitative cluster analysis (de Loë et al., 2016). Comment analysis can help describe the most and least common perspectives, identify potential reasons that panelists drop out, track ongoing disagreements among panelists, and list caveats at the end of the Delphi process.

Because Delphi is a good example of a mixed-methods methodology, qualitative and quantitative analyses should be aligned and well-integrated (Khodyakov et al., 2019). The analyses should be carefully documented so that readers understand how they were carried out, who performed them, and how the analyses informed the final results. Best practices for conducting qualitative data analysis and reporting qualitative research should be followed in analyzing Delphi comments (Malterud, 2001).

Missing Data

Given that panelist attrition is common over Delphi rounds, panel conveners should prespecify a plan for handling missing data. Options include not performing any data imputation, imputing the value from a previous round (Last Observation Carried Forward; Dechartrres et al., 2011), or using multiple imputations from values of other data in the same round (see Khodyakov and Chen, 2020). Panel conveners should also try to identify the reasons for nonresponse, which may include competing commitments, lack of relevant expertise, poorly worded instructions or questions, and high perceptions of burden (Landeta, 2006). It is also important to note any patterns in

---

**RECOMMENDATION**

- Use content analysis, thematic analysis, or qualitative cluster analysis to summarize panelists’ comments and link them with the quantitative results.

- Develop a plan for handling missing data, explore the reasons for nonresponse, and look for patterns in the demographic characteristics of nonresponders.
the demographic characteristics of panelists who do not complete all rounds. Such patterns should be carefully considered during result interpretation and reported in the limitations section.

**Weighting Responses**

Delphi methodologists have varying views on the utility of weighting panelist responses in data analysis. Those in favor argue that more weight should be given to responses from panelists with greater expertise in the topic area or confidence in their ratings (Kawamoto et al. [2019] and Winkler and Moser [2016] highlight some solutions for how this is done). Those against argue that measures of the degree of expertise can be unreliable and that (over)confidence is not associated with more-accurate ratings. In general, researchers suggest avoiding differential weighting without a clear empirical justification and evidence of successful performance in similar tasks (Murphy et al., 1998).

Studies have explored using self-rated measures of expertise and/or confidence to either select panelists or differentially weight their responses. Expertise is critical for the accuracy of Delphi results, but evidence on the utility of self-rated measures is equivocal, suggesting that some of these measures may be better than others depending on the specific context of the study (Rowe and Wright, 1999).

**RECOMMENDATION**

Avoid differential weighting of responses without a clear empirical justification and evidence that doing so improves the quality of Delphi results in similar studies.
CHAPTER 6

Reporting

Result reporting is the final step. This chapter offers practical considerations for reporting on panelist attrition across rounds, clearly conveying findings and interpreting the results, publishing in a way that promotes transparency, accounting for ethical considerations, and returning the results to the panelists. Additionally, in preparing their results for publication, panel conveners should consult Delphi reporting guidelines, such as the ACCORD and DELPHISTAR standards, to ensure that all applicable elements are covered appropriately.

Panelist Flow and Attrition Across Rounds

Delphi research teams should describe their panelists by reporting their demographic characteristics and areas of expertise, state the number of panelists in each round, and calculate participation rates. Clearly stating what denominator was used to calculate participation rates in each round helps explain attrition (Miller et al., 2023): Is it all invited panelists, those who agreed to participate, or those who completed a previous round who drop out? Explaining procedures for handling missing data (e.g., unanswered questions) is important. If no data imputation was performed, reporting Ns for each question presented is helpful if participants were allowed to skip questions.

RECOMMENDATION

Specify the panel's intended size, describe panelists’ demographic characteristics and areas of expertise, and present participation rates in each round.
Findings

Even small Delphi panels typically generate a lot of data. Deciding how to present the results is not an easy task. Panel conveners need to think about their audience and determine what the key findings and takeaways are. While some audiences may want to see intermediary results, such as how experts’ perspectives changed between the rounds, these findings may be better suited for an appendix or a methodological paper. For papers that present substantive results, presenting the final list of items with their associated ratings is important (Khodyakov et al., 2019). Finding a way to visually display the results is also helpful. For example, the use of forest plots to show results of trial results in a meta-analysis has been adapted to visually display the median and dispersion of ratings (see Figure 6.1) among policy experts in Delphi studies (Smart and Grant, 2021). While research teams typically focus on measures of central tendency, seeing the responses outside of a range of common answers may be helpful. Showing anonymized quotes illustrating dissenting views is a powerful strategy, as well (Khodyakov et al., 2017).

FIGURE 6.1
Forest Plot Example

Recommendation 1
Recommendation 2
Recommendation 3
Recommendation 4
Recommendation 5
Recommendation 6

Consider the study’s audience in determining what results and takeaways should be highlighted.
Interpretation of Results in Context

Events that occur during a Delphi study have the potential to affect the results. For example, the lockdown protocols for the coronavirus disease 2019 (COVID-19) pandemic—and subsequent changes to telehealth policy—influenced expert ratings of opioid policies with telehealth components (Grant and Smart, 2022; Smart and Grant, 2021). At a minimum, providing the dates of the overall panel and each round can help readers to understand how recent the findings are and consider how contextual factors might have affected them. In addition, discussing how the study’s results relate to other relevant evidence should help readers interpret the findings (Page et al., 2021). For example, authors might compare their results with results of similar studies and explore possible reasons for replicated versus discordant results. Similarly, the research team might summarize additional information relevant to decisionmakers in a policy area that was not explored in the Delphi study, such as findings of studies summarized in the evidence dossiers provided to Delphi panelists.

Addressing Limitations of the Delphi Process

When discussing study findings, research teams should carefully consider Delphi limitations and how these limitations might have affected their conclusions. Was the panel large enough to represent a range of perspectives on the topic? Were all relevant stakeholder groups represented on the panel? Were some types of panelists more likely than others to drop out? How confident is the research team in the panel findings? In general, we recommend conducting two or more concurrently run panels that use the same data collection protocol and have similar types and numbers of panelists randomly assigned to a given panel as a way to address the criticism that Delphi panels results are not replicable. Comparing panel results and moving forward with the items that both panels agreed on increases confidence in Delphi results (Khodyakov et al., 2019).

Research Transparency

There is a growing “open science” movement in health and social science research, which calls for researchers to prospectively register their research protocols and analysis plans to indicate that a study is ongoing; share data, code, and materials to address publication bias (i.e., selectively reporting results based on the nature of findings); and publish using the open
access model to allow free access to study results (Nosek et al., 2018). Study registration is a time-stamped entry of a minimum set of information in a publicly accessible, independently controlled registry, such as the Open Science Framework (OSF) Registry, which provides an identification number that links all study products and outputs, including its protocol and manuscripts (De Angelis et al., 2005). For Delphi studies, prospective registration and protocol-sharing is particularly important for guarding against post hoc data analyses that can threaten the credibility of consensus analyses.

Ethical Considerations

Panel conveners should also account for ethics considerations, including human subjects protections and Institutional Review Board (IRB) approval, conflicts of interest, and the role of funding and endorsing organizations. All Delphi panelists, including experts, are considered human subjects. Although expert panels are often determined to be exempt from IRB review, the research team should submit a protocol to their IRB, and panelists should provide informed consent prior to the start of data collection. Although written consent may not be required, panelists should be informed about the study purpose, data collection procedures used, and risks and benefits associated with their participation. This can be accomplished by providing a study information sheet, programming a consent form in the e-Delphi software used, or sharing this information during a discussion meeting.

Panelists should also be informed about whether or not other panelists would know their identities during the data collection process and be asked whether they want their names and affiliations to be shared in final reports. Panel conveners should also consider and report the measures they used to identify panelists and mitigate potential conflicts of interest among them, the research team itself, and any organizations funding or endorsing the panel. Furthermore, panel conveners should disclose the funder and any organizations endorsing the panel or its findings, as well as their role in the process (Krimsky, 2013). Finally, panel conveners should delineate the criteria for deciding whether a study team member made sufficient contributions to warrant an authorship credit (versus an acknowledgement).

RECOMMENDATION

Register the study prospectively and share its protocol as a way to increase the credibility of panel results.

RECOMMENDATION

Account for human subjects’ protection, consider potential conflicts of interest, and state the role of funding and endorsing organizations.
Return of Results to Panelists

As discussed, returning the results to panelists is a key characteristic of Delphi. While showing intermediary panel results to panelists is part of the methodology, sharing the results of the final Delphi round is not. Nonetheless, doing so can be useful and is consistent with best practices of stakeholder-engaged research (Esposito et al., 2015). Sharing of final results could take place via a webinar in which panelists can finally see who the other panelists were (assuming that their panel was anonymous), discuss any remaining disagreement, and help the panel conveners interpret the final results. Other approaches include sending a summary of findings to panelists or sharing copies of the manuscripts before they are published.

**RECOMMENDATION**

Share both intermediary and final Delphi results with panelists.
CHAPTER 7

DCAT: Delphi Critical Appraisal Tool

To increase the methodological rigor of Delphi studies, this chapter presents the Delphi Critical Appraisal Tool (DCAT). The DCAT makes explicit the standards by which others might evaluate the quality of a Delphi study. We envision that both those designing and those reading the results of a Delphi study can use the DCAT to assess the quality of its design, conduct, and analysis. Researchers can use the tool prospectively to design and implement high-quality Delphi studies and thoroughly report their panel results. The tool is a checklist with 16 items (four “core” and 12 “additional”) that inform assessments of a Delphi study’s overall quality (see Appendix B).

DCAT items are self-explanatory; the tool includes brief descriptions of all items using the information from earlier report chapters and should be accessible to people who are not trained in the Delphi method. However, because the issues underlying each DCAT item can be complex and subject to interpretation, readers should refer to earlier chapters for more details.

When using this tool to critically assess a Delphi study, appraisers should first answer each question by selecting “yes” if the Delphi study met quality standards for that item, “no” if it did not meet standards, and “unsure” if insufficient information is provided to make an assessment. In the latter case, the appraiser may consider contacting the study authors for further information. Then, appraisers should make an informed quality assessment of a given Delphi study, focusing on its overall design, implementation, and analysis.

Core Items

The DCAT begins with four core items that reflect the four key characteristics of the Delphi method (see Chapter 2): anonymity of collected data and reporting, iteration, controlled feedback, and statistical summary. While all steps in the design, conduct, and analysis are important, we believe that these four key characteristics critically affect a Delphi study’s quality. The way that some core items are implemented can differ by the type of Delphi used (e.g., modified Delphi studies following RAM may have a face-to-face discussion that cannot be anonymous). Therefore, it is important that appraisers assess the core item in a manner appropriate to the specific type of Delphi that they are assessing.
ITEM 1. Did the research team appropriately employ anonymity in the Delphi study?

Anonymity is a key Delphi characteristic. It is important for increasing objectivity of Delphi results because it helps reduce groupthink, the “halo effect,” the negative impact of the dominant personalities, and the pressure on individuals to change their opinions based on the status of other panelists or to move closer to consensus. Anonymity also allows panelists to share opinions that they think would otherwise be unpopular or unconventional and makes it easier for them to change their responses without feeling as if they are losing face. Delphi studies should report the level of anonymity provided to panelists. Partial anonymity (when panelists know the stakeholder group of other panelists) may be acceptable if panelist names are not revealed. Delphi studies with synchronous discussion components (e.g., modified Delphi following RAM) may still employ anonymity appropriately if individual responses to rating questions are not attributable to a given panelist.

To score “Yes,” appraisers should be confident that the research team appropriately maintained the (partial) anonymity of panelists.

ITEM 2. Did the research team appropriately employ iteration in the Delphi study?

Because iteration is another key characteristic of the Delphi method, Delphi studies should have a minimum of two rounds of questions. While minor modifications of questions and response items may be permissible, a Delphi study should involve at least two rounds that include the same rating protocol, with panelists having at least one opportunity to revisit their responses in a previous round. Although dropping items that have already reached consensus may be customary for some types of Delphi, such as RAM, it is not appropriate to score all items only once. Measuring response stability also requires iteration. Published Delphi studies typically include two to seven rounds, with most on the lower end of that range. Research has found that two or three rounds are generally sufficient for most Delphi studies and that additional rounds lead to attrition and could encourage forced consensus.

To score “Yes,” appraisers should be confident that the Delphi study included at least two rounds that included the same rating protocol, with panelists having at least one opportunity to revise their original responses.
**ITEM 3. Did the research team appropriately employ statistical summaries of group responses in the Delphi study?**

A third key characteristic of the Delphi method is the use of statistical summaries of group responses to summarize the findings of the panel. Delphi studies should use statistical analysis to determine what the panel “thinks” to ensure objectivity. Responses from the previous round may be summarized by reporting one or more measures of central tendency (e.g., mode, median, or mean) and dispersion (e.g., range, interquartile range, and standard deviation), depending on the questions asked. Consensus should be also determined statistically by calculating the percent of responses above a certain point (for example, 75 percent of responses greater than 7 on a 1–9 scale) or by presenting various measures of agreement, consensus, and/or stability.

To score “Yes,” appraisers should be confident that appropriate statistical analyses of the group responses were used as the basis for determining the panel results (e.g., consensus among panelists).

---

**ITEM 4. Did the research team appropriately employ controlled feedback in the Delphi study?**

The final key characteristic of a Delphi method is controlled feedback, which is important for helping panel members reach consensus. Seeing how their answers compare with those of others helps panelists reevaluate their perspectives and change their answers if needed. Delphi studies should describe the type of feedback the panel received, including quantitative and qualitative summaries that include panelists’ own responses.

To score “Yes,” appraisers should be confident that the research team showed panelists the group response to each question in the previous round and personalized feedback that listed panelists’ own original answers and how their answers compare with those provided by others.
Additional Items

This tool also has 12 “additional” items that cover various steps of Delphi design, implementation, and analysis. While these additional items do not pertain to key Delphi characteristics, they cover important methodological considerations that may affect Delphi study quality and therefore should be critically appraised.

**ITEM 5.** Did the research team provide an appropriate justification for the use of the Delphi method in this study?

The Delphi method can be used to achieve different goals, including consensus-building and identifying areas of disagreement. Consequently, the research team should explain and justify the use of this method. In general, Delphi is appropriate for forward-looking studies on topics for which other sources of evidence are not available or existing evidence is inconclusive, as well as for studies that seek to explore areas of agreement and disagreement in diverse, multistakeholder panels on complex issues for which there are no right or wrong answers.

To score “Yes,” appraisers should be confident that the research team provided appropriate justification for using the Delphi method for the study topic.

**ITEM 6.** Did the research team provide an appropriate justification for the type of Delphi used in this study?

Because there are different types of Delphi (e.g., classical, RAM, real-time), studies should use the one that is most appropriate. The research question that a study is designed to answer should dictate the type of Delphi used. In turn, the type of Delphi is likely to affect a number of other design decisions, including the questions asked and the mode of panel administration.

To score “Yes,” appraisers should be confident that the type of Delphi was appropriate for the research question and study objectives.

**ITEM 7.** Did the research team seek the appropriate expertise for the Delphi study?

Panel composition, including the type and level expertise, demographic and geographic diversity, and potential conflicts of interest of panelists, can significantly influence Delphi study results and their validity. Panelists are generally selected purposefully to ensure that
they have the required expertise and experience. Depending on the panel topic and goals, *expertise* may be defined broadly to include lived experience. Studies should use a multi-pronged approach to recruitment, such as direct outreach to individuals and professional societies, personalized contact, and social media. All panelists should be screened for potential conflicts of interest if needed.

To score “Yes,” appraisers should be confident that the research team clearly described panel eligibility criteria and recruitment strategy and that both were appropriate, with the composition reflecting the purpose of the study.

**ITEM 8.** Did the research team sufficiently retain panelists throughout the Delphi study?

Panelist nonresponse and attrition are common in Delphi panels, especially those that are conducted online or do not include a face-to-face discussion round. At a minimum, the research team should report attrition for each round, types of panelists lost to attrition, reasons for attrition (if known), and how missing data were handled (e.g., complete case analysis, carrying forward responses from the previous round). Although attrition may be unavoidable, researchers should use strategies to minimize it, such as inviting panelists who truly care about the topic and have the right type of expertise, incentivizing their participation in all rounds, and minimizing participation burden.

To score “Yes,” appraisers should be confident that the research team reported participation rates, used appropriate techniques to mitigate attrition, and ensured that the final sample size was sufficient for a reliable summary of expert views.

**ITEM 9.** Did the research team conduct preliminary work prior to the first round of the Delphi study?

Before starting the first round, the research team should consider including a summary of existing evidence to be shared with panelists to help inform their answers. Literature reviews or evidence syntheses are very common in health and medical research fields, but they are not as typically used in Delphi studies conducted by other disciplines. Summarizing what is known about the topic using infographics may be particularly helpful in panels that involve stakeholders with diverse areas of expertise to facilitate a shared, baseline understanding of an issue at hand. Chosen panelists may also be asked about their thoughts on the study topic.
in an open-ended manner before the first round of structured questions to ensure that their input shapes the wording of the questions in subsequent rounds.

To score “Yes,” appraisers should be confident that the research team adequately conducted preliminary work prior to the first survey round.

ITEM 10. Did the research team make appropriate design choices for the Delphi study?

Research teams should design Delphi studies in ways that do not force the development of consensus but allow for exploring its existence after exposing panelists to different perspectives in an unbiased way. For example, Delphi studies can be conducted using different modes of administration, with online panels becoming the norm. Research teams also need to consider and prespecify the number of panels to assemble, the length and timing of each round, and the number of rounds or the stopping criteria. Studies using modified Delphi also need to consider how to conduct and moderate the discussion round.

To score “Yes,” appraisers should be confident that the research team selected appropriate design features (e.g., format, panels, rounds, stopping criteria, discussion round) for the study.

ITEM 11. Did the research team ask appropriate questions in the Delphi study?

Delphi studies ask panelists to answer questions about a series of items (e.g., statements or scenarios). Researchers should consider asking the same questions about all items to allow them to compare and contrast the items on multiple dimensions and to simplify and expedite data collection. They should follow best practices for survey design in developing Delphi questions by ensuring that questions are specific, clear, unambiguous, unbiased, and not double-barreled.

To score “Yes,” appraisers should be confident that the research team provided a sufficient rationale for how the questions were selected and created in relation to the study goals and that all questions are clearly worded, easy to interpret, and adequately measure underlying constructs.
**ITEM 12.** Did the research team use appropriate response items in the Delphi study?

Most Delphi studies include self-administered questionnaires. Therefore, response options should be clear, exhaustive, and relevant to the questions asked. Some studies may use a middle point on a rating scale with an odd number of points to indicate a neutral or “do not know” response. Others use scales with an even number of points to encourage panelists to provide either a positive or negative, rather than neutral, answer. The number of points on a response scale is also an important consideration. All else being equal, having fewer response options increases the likelihood of consensus development. Therefore, researchers should carefully select questions and response scales and explain their design rationales. They should also include space for panelists to explain their answers (e.g., in open-text comment boxes after rating items).

To score “Yes,” appraisers should be confident that the research team provided appropriate response options and that the available response options did not negatively affect the validity of results.

Yes □ No □ Unsure □

**ITEM 13.** Did the research team conduct a pilot test before launching the Delphi study?

Research teams should conduct a pilot test with a small group of panelists to ensure that the data collection process, instructions, and questions are clear. This process can also help mitigate issues that would otherwise hinder retention in the panel (e.g., reducing the number of items that panelists are asked to rate). In the pilot test, the panelists should be provided with instructions, the list of items to be assessed and questions asked (e.g., rating criteria to be used), and/or a link to the programmed questionnaire so that they can preview the proposed process and suggest revisions. Conducting interviews with pilot testers to solicit their feedback directly and reporting on the changes made based on pilot testers’ feedback may also be useful.

To score “Yes,” appraisers should be confident that the research team conducted a pilot test and used its results to inform the final study design.

Yes □ No □ Unsure □
ITEM 14. Did the research team appropriately analyze the data from the Delphi study?

The research team should have an appropriate plan for descriptive data analyses, measures of stability and consensus, and qualitative analysis of panelist comments that allows the comments to complement the results of statistical analyses. To avoid the appearance of bias, the research team should determine all analytic procedures in advance; the best way to assure proper pre-specification is to publish or register a prospective protocol and data analysis plan (e.g., using the Open Scientific Framework [OSF] Registry).

To score “Yes,” appraisers should be confident that the research team used an appropriate and comprehensive data analysis approach.

☐ Yes  ☐ No  ☐ Unsure

ITEM 15. Did the research team transparently report their Delphi study?

When reporting final results, the research team should adhere to Delphi reporting guidelines, such as the ACCORD and DELPHISTAR standards, to increase research transparency and support evaluations of the study’s quality.

To score “Yes,” appraisers should be confident that the research team supplied sufficient information to understand how the data have been collected and analyzed and to assess the scientific rigor of the methodological approach.

☐ Yes  ☐ No  ☐ Unsure

ITEM 16. Did the research team appropriately consider ethical issues?

The research team for Delphi studies should adhere to all relevant principles of ethical research conduct, including human subjects’ protections, Institutional Review Board approval, panelist consent, assessment of possible conflicts of interest, and identification of the role of funding and endorsing organizations.

To score “Yes,” appraisers should be confident that the research team gave ethical considerations the appropriate consideration.

☐ Yes  ☐ No  ☐ Unsure
Overall Quality of the Delphi Study

Drawing on the above assessments, appraisers should rate the overall quality of the Delphi study, including its design, implementation, and analysis using the categories in Figure 7.1.

FIGURE 7.1
Categories for Rating the Overall Quality of a Delphi Study

Following other critical appraisal tools (Shea et al., 2017), we propose a scheme for incorporating weaknesses detected in core and additional items into the overall quality assessment. A “No” answer to any of the four core items should downgrade the overall assessment by at least one level (e.g., “High” to “Moderate”). An answer of “No” to one or a few additional items may or may not lead to downgrading, but multiple “No” answers should downgrade the overall assessment by at least one level. This differential treatment of core and additional items is advisory; appraisers should clearly specify and explain which items are most important for their assessment. They should also provide brief statement describing the main reasons for their overall quality assessment.

Explanation of overall quality assessment: ________________
__________________________________________________________________
__________________________________________________________________
APPENDIX A

Summary of Recommendations

FIGURE A.1
Recommendations for Design of Delphi Studies

**RECOMMENDATIONS FOR DESIGN OF DELPHI STUDIES**

- Use Delphi in forward-looking studies on topics for which expert opinion is the only source of evidence and for studies that explore areas of agreement and disagreement on complex issues.
- Use the study research question to determine the most appropriate type of Delphi.
- Include researchers with expertise in the topic under investigation and expertise in the Delphi method on the study team.
- Assemble panels that represent diverse perspectives and experiences.
- Employ a multipronged recruitment approach to assemble diverse panels.
- Recruit more panelists than needed.
- Use a range of panelist retention strategies.
- Summarize what is known about the topic to help develop a shared, baseline understanding of the issue at hand.
- Conduct a small pilot test to ensure that the data collection process, instructions, and questions are clear to panelists.
- Use the study’s design, goals, and needs to determine panel format.
- Ask the same questions at least twice to allow panelists to know what others think.
- Recruit discussion moderators who have a broad understanding of the discussion topic and know how to manage group dynamics.
- Use task complexity, number of questions and panelists, and panel format to decide on the length and timing of each round.
- Conduct at least two concurrent Delphi panels that use the same data collection protocol and randomly assign panelists to one of the two panels to increase validity of Delphi results.
- Include all relevant stakeholders to ensure the validity of panel findings.
- Determine panel size by considering the mode of data collection and panel interaction, the level of panel heterogeneity, and the likely attrition rates.
- Ensure that feedback reports include statistical summaries of panelists’ responses and list participants’ own answers.
- Determine the level of panelist anonymity before enrolling panel members.
- Decide on the number of Delphi rounds before the start of data collection.
FIGURE A.2
Recommendations for Data Collection in Delphi Studies

RECOMMENDATIONS FOR DATA COLLECTION IN DELPHI STUDIES

- Ask the same questions about all items considered to compare and contrast panelists’ views along several dimensions.
- Ask specific, clear, unambiguous, and unbiased questions; provide response options that are exhaustive and relevant to the question.
- Do not include “I do not know” as a response option.
- Choose what response scales to include as they may affect consensus development.
- Decide whether modifications to items and questions will be made between rounds before the start of the data collection.
- Ensure that moderators facilitate the exchange of ideas, ask neutral probing questions, engage all panelists in the discussion, and keep discussion on track.
- Review available Delphi software packages and their functionalities before choosing one that meets the study’s needs.

FIGURE A.3
Recommendations for Data Analysis in Delphi Studies

RECOMMENDATIONS FOR DATA ANALYSIS IN DELPHI STUDIES

- Show measures of central tendency and dispersion along with a frequency distribution of responses to each question to panelists.
- Investigate the stability of responses for each question between rounds using nonparametric tests or measures of dispersion.
- Choose and prespecify a measure of consensus that is appropriate for the type and goal of the study’s Delphi panel.
- Use content analysis, thematic analysis, or qualitative cluster analysis to summarize panelists’ comments and link them with the quantitative results.
- Develop a plan for handling missing data, explore the reasons for nonresponse, and look for patterns in the demographic characteristics of nonresponders.
- Avoid differential weighting of responses without a clear empirical justification and evidence that doing so improves the quality of Delphi results in similar studies.

FIGURE A.4
Recommendations for Reporting in Delphi Studies

RECOMMENDATIONS FOR REPORTING IN DELPHI STUDIES

- Specify the panel’s intended size, describe panelists’ demographic characteristics and areas of expertise, and present participation rates in each round.
- Consider the study’s audience in determining what results and takeaways should be highlighted.
- Provide panel dates, summarize relevant evidence, and compare panel results with existing literature to help readers contextualize Delphi results.
- Interpret Delphi results by carefully considering the method’s limitations.
- Register the study prospectively and share its protocol as a way to increase the credibility of panel results.
- Account for human subjects’ protection, consider potential conflicts of interest, and state the role of funding and endorsing organizations.
- Share both intermediary and final Delphi results with panelists.
APPENDIX B

DCAT Checklist

Figure B.1 on the next page presents the Delphi Critical Appraisal Tool (DCAT) checklist.
## FIGURE B.1
DCAT Checklist

<table>
<thead>
<tr>
<th>ARTICLE ID:</th>
<th>REVIEWER INITIALS:</th>
</tr>
</thead>
</table>

### CORE ITEMS

<table>
<thead>
<tr>
<th>ITEM</th>
<th>Description</th>
<th>Yes</th>
<th>No</th>
<th>Unsure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Did the research team appropriately employ anonymity in the Delphi study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Did the research team appropriately employ iteration in the Delphi study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Did the research team appropriately employ statistical summaries of group responses in the Delphi study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Did the research team appropriately employ controlled feedback in the Delphi study?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### ADDITIONAL ITEMS

<table>
<thead>
<tr>
<th>ITEM</th>
<th>Description</th>
<th>Yes</th>
<th>No</th>
<th>Unsure</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.</td>
<td>Did the research team provide an appropriate justification for the use of the Delphi method in this study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Did the research team provide an appropriate justification for the type of Delphi used in this study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>Did the research team seek the appropriate expertise for the Delphi study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>Did the research team sufficiently retain panelists throughout the Delphi study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>Did the research team conduct preliminary work prior to the first round of the Delphi study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>Did the research team make appropriate design choices for the Delphi study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>Did the research team ask appropriate questions in the Delphi study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>Did the research team use appropriate response items in the Delphi study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>Did the research team conduct a pilot test before launching the Delphi study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>Did the research team appropriately analyze the data from the Delphi study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td>Did the research team transparently report their Delphi study?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16.</td>
<td>Did the research team appropriately consider ethical issues?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### OVERALL QUALITY ASSESSMENT

[ ] [ ] [ ] [ ]

Explanation of overall quality assessment:______________

High  Moderate  Low  Critically low
References


The Delphi method is an iterative, anonymous, structured, group-based communication process and elicitation technique designed to help policymakers make decisions under conditions of uncertainty and incomplete information. It is based on the premise that asking a hand-picked group of anonymous experts the same questions several times and sharing the other experts’ answers will help objectively develop group consensus, which is used as a form of evidence. Although it was originally developed by RAND researchers as a forecasting methodology in military research, Delphi underwent many modifications and is now used by different disciplines, most notably by medicine, as a gold-standard approach for expert elicitation and stakeholder engagement. Researchers often rely on Delphi to estimate the probability of an event happening within a certain period of time, to forecast when an event is likely to occur, and to identify and prioritize key policy issues that need to be addressed.

Although the Delphi method is used by different disciplines and for different purposes, there is still a lack of methodological guidance for how to conduct rigorous Delphi studies. This manual addresses this critical gap by providing practical advice for when to use Delphi, what type of Delphi to choose, and what to consider when designing, implementing, and reporting Delphi panel results. The manual also includes the Delphi Critical Appraisal Tool (DCAT), which researchers can use both retrospectively to appraise the quality of an already conducted Delphi study and prospectively to design and implement a Delphi study.