ANZJP

Australian & New Zealand Journal of Psychiatry
2015, Vol. 49(10) 887–897
DOI: 10.1177/0004867415600891

Using the Delphi expert consensus method in mental health research

Anthony F Jorm

Editor's Choice

Abstract

Objective: The article gives an introductory overview of the use of the Delphi expert consensus method in mental health research. It explains the rationale for using the method, examines the range of uses to which it has been put in mental health research, and describes the stages of carrying out a Delphi study using examples from the literature.

Method: To ascertain the range of uses, a systematic search was carried out in PubMed. The article also examines the implications of 'wisdom of crowds' research for how to conduct Delphi studies.

Results: The Delphi method is a systematic way of determining expert consensus that is useful for answering questions that are not amenable to experimental and epidemiological methods. The validity of the approach is supported by 'wisdom of crowds' research showing that groups can make good judgements under certain conditions. In mental health research, the Delphi method has been used for making estimations where there is incomplete evidence (e.g. What is the global prevalence of dementia?), making predictions (e.g. What types of interactions with a person who is suicidal will reduce their chance of suicide?), determining collective values (e.g. What areas of research should be given greatest priority?) and defining foundational concepts (e.g. How should we define 'relapse'?). A range of experts have been used in Delphi research, including clinicians, researchers, consumers and caregivers.

Conclusion: The Delphi method has a wide range of potential uses in mental health research.

Keywords

Delphi method, expert consensus, research methods

The purpose of this article is to give an introductory overview of the use of the Delphi expert consensus method in mental health research. The article explains the rationale for using this method, examines the range of uses to which it has been put in mental health research and gives examples of how the method is used in practice. However, I begin by first discussing the role of consensus in medical science and the misunderstandings that have sometimes led to expert consensus methods having a negative image.

The role of consensus in medical science

The last two decades have seen a strong movement towards evidence-based medicine, with an emphasis on informing clinical decisions by the findings from randomized controlled trials (Sackett and Rosenberg, 1995). A major tool of the evidence-based-medicine movement has been statements of Levels of Evidence. According to these tools, a systematic review of randomized controlled trials is the strongest form of evidence, with various weaker forms of evidence arranged under it in a hierarchy. In some of these schemes, expert consensus is at the bottom of the hierarchy. For example, in the Joanna Briggs Institute Levels of Evidence for Effectiveness, the lowest level is 'Expert Opinion and Bench Research' (Joanna Briggs Institute and University of Adelaide, 2013).

However, expert consensus should not automatically be classed as an inferior method. The strength of an expert consensus method depends in large part on what evidence the consensus is based on. The evidence on which experts make their judgements may include, for example, systematic reviews, individual experiments, qualitative studies and personal experience. When expert consensus is

Centre for Mental Health, Melbourne School of Population and Global Health, The University of Melbourne, Carlton, VIC, Australia

Corresponding author:
Anthony F Jorm, Centre for Mental Health, Melbourne School of Population and Global Health, The University of Melbourne, Carlton, VIC 3010, Australia.
Email: ajorm@unimelb.edu.au

considered weak in Levels of Evidence hierarchies, this may be because the consensus is based purely on clinical or other personal experience – what might be called 'practice-based evidence'.

It is also difficult to maintain that expert consensus is weak given that it has an important role in establishing all the tools of evidence-based medicine. Not only are Levels of Evidence statements themselves based on expert consensus, but so are other tools like the *Cochrane Handbook for Systematic Reviews of Interventions* (Higgins and Green, 2011), the Consolidated Standards of Reporting Trials (CONSORT) Statement on standards for reporting trials (Begg et al., 1996), and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement on systematic reviews and meta-analyses (Moher et al., 2009). Indeed, the Delphi method has been used to develop criteria for quality assessment of randomized clinical trials (Verhagen et al., 1998), guidelines for clinical trial protocol content (Tetzlaff et al., 2012), standards for reporting interventions used in trials (Hoffmann et al., 2014) and methods used in systematic reviews (Pincus et al., 2011). When used in this way, expert consensus methods are a type of foundational methodology upon which all other methodologies rest.

More broadly, expert consensus is a fundamental underpinning of science. It is used to determine what are appropriate methodologies, to decide which grant applications will get funded, which manuscripts will get published, and who will be admitted to learned societies of experts. Citation metrics can also be thought of as reflecting the consensus of a discipline about the importance of a publication. Science can be thought of as what the community of acknowledged experts in a field consider to be the current truth. This consensus changes over time as knowledge increases. The Delphi method is one of many that have been used to determine expert consensus. Sometimes consensus builds rapidly and spontaneously in science, based on a critical piece of evidence. This is more often the case in the physical sciences, where a single piece of evidence may be sufficient to change expert beliefs. By contrast, in sciences dealing with highly complex systems, which include the disciplines relevant to mental health, consensus changes more slowly and formal mechanisms may need to be used to ascertain it.

My conclusion is that it is overly simplistic to characterize expert consensus methods as invariably low on the Levels of Evidence. The quality of the evidence they produce depends on the inputs available to the experts (e.g. systematic reviews, experiments, qualitative studies, personal experience) and on the methods used to ascertain consensus, as described below.

Consensus and the wisdom of crowds

For expert consensus to produce good answers, it needs to be ascertained systematically and using methods that are known to produce accurate outcomes. There has been research on the conditions under which groups of individuals with some expertise make good decisions. James Surowiecki (2004) has summarized this literature in his book *The Wisdom of Crowds*, where the term 'crowd' is used to refer to any collection of individuals with some expertise, including scientists and clinicians. The starting point of this literature was a study by Francis Galton in the early 20th century of a competition held at an English country fair to estimate the weight of an ox after it had been butchered. Galton analysed the distribution of the 787 guesses and found that the median guess was remarkably accurate (within 0.8%). It appeared that by aggregating a large number of imperfect estimates, the group could make a much better estimate than the most skilled individuals. Surowiecki's book summarizes numerous other studies showing that crowds can produce better estimates than the best individual experts. However, crowds are not always wise, as seen in the phenomenon of 'groupthink', where group pressures lead to irrational decisions. Surowiecki proposes that certain conditions must be met for a crowd to be wise:

1. *Diversity of expertise*. A heterogeneous crowd of experts will produce better quality decisions than a homogeneous one.
2. *Independence*. The experts must be able to make their decisions independently, so that they are not influenced by others.
3. *Decentralization*. Expertise is held by autonomous individuals working in a decentralized way.
4. *Aggregation*. There is a mechanism for coordinating and aggregating the crowd's expertise.

While crowds are not always wise, the wisdom-of-crowds effect has been found to be robust under a range of conditions (Davis-Stober et al., 2014; Jönsson et al., 2015). Scott Page (2007) has examined various explanations for how a crowd can be wise. He argues that in complex tasks, like guessing the weight of an ox, individuals have 'predictive models' that they use to produce an estimate. (An example of a predictive model might be: '*This ox is about 5 times my size – I weigh 80 kilograms – therefore the ox must weigh 400 kilograms*'.) Page has demonstrated that to get optimal predictions from a crowd, the individuals in the crowd must have good predictive models and these models must be diverse. A crowd where everyone uses the same model or a small range of models will perform worse than a crowd with a more diverse range. In the extreme case, where all members of a crowd use the same predictive model and come up with the same judgment, the crowd will be no better at predicting than a single individual in it.

The basics of the Delphi method

The Delphi technique was described by one of its originators as 'a method of eliciting and refining group judgments'

(Dalkey, 1969). It was originally developed as a method for forecasting, but has since been widely applied in other areas, including health research. The Delphi method has many variants, but the key elements are as follows:

1. There is a facilitator who organizes the Delphi study.
2. The facilitator recruits a group of individuals with some expertise on the topic.
3. The facilitator compiles a questionnaire with a list of statements that the experts rate for agreement.
4. The facilitator gathers responses from the members of the group using the questionnaire.
5. The facilitator gives anonymous feedback to individuals in the group about how their responses compare to the rest of the group.
6. The members of the group are able to revise their responses to the questionnaire after receiving the feedback.
7. Responses converge across rounds of questionnaires, with some statistical criterion being used to define consensus.

When viewed in the light of the wisdom-of-crowds literature, the Delphi method can be seen to incorporate many of the conditions that lead crowds to be wise. These include independence of decisions (through responses to anonymous questionnaires), decentralization (the members of the group operate autonomously, but share decisions through the facilitator) and aggregation (through the facilitator's organization of the group and statistical summarization of results). One other condition for a wise crowd, diversity of expertise, is not a requirement of the Delphi method. However, as discussed below, diversity needs to be considered when selecting panel members in order to promote optimal decision making.

Uses of the Delphi method in mental health research

The Delphi method has been used for a wide variety of purposes in mental health research. To ascertain the range of uses, a search was carried out with PubMed for articles published between 2000 and 18 March 2015 using the search terms: 'Delphi' AND ('mental disorders' OR 'mental health' OR psychiatr*). After excluding studies that did not use the method, a total of 176 articles was found. Table 1 shows the types of use to which the method has been put and gives an illustrative example of each.

The types of consensus decisions that are made in these Delphi studies can be grouped into broad categories as follows:

1. Making estimations where there is incomplete evidence, e.g. What is the global prevalence of dementia? What antipsychotic drug dose is optimal?
2. Making predictions, e.g. What types of interactions with a person who is suicidal will reduce their chance of suicide? What parenting practices will reduce an adolescent's risk of depression?
3. Determining collective values, e.g. What areas of research should be given greatest priority? What should be the performance indicators for mental health care?
4. Defining foundational concepts, e.g. How should we define 'relapse'? How should we segment the hippocampus?

While these broad categories are helpful for thinking about the range of uses of Delphi studies, in practice a particular Delphi study might have aims that span more than one category.

Carrying out a Delphi study

Carrying out a Delphi study involves a series of steps and choices. These are described below with illustrative examples from the mental health research literature.

Framing a research question

As with all research, the first step is to have a clear question that is answerable by the methodology. Examples of questions that might be answered by a Delphi study include the following:

- How can a member of the public best assist a person who is suicidal?
- What mental health research topics should be prioritized by funders?
- How should we define 'relapse' in schizophrenia?
- How many people are affected by dementia globally?

It is notable that some of these questions would be difficult (and in some cases impossible) to answer by methodologies other than expert consensus.

Selecting the expert panel

The researcher has to choose a group of individuals who have expertise relevant to the question. Ideally, there should be a clear definition of what constitutes expertise and a sampling strategy for locating experts who meet it. Here are two examples:

- In a study of parenting strategies to reduce risk of adolescent depression and anxiety disorders, experts had to have a minimum of 5 years of experience in research or clinical treatment on parenting and adolescent depression or anxiety. Researchers were

Table 1. Uses to which the Delphi method has been put in mental health research.

Type of use	Example
Improve professional practice	Standards of practice for the adult mental health workforce (Goodyear et al., 2015)
Improve professional training	Content of mental health-care training for practitioners in remote and rural areas (De Mello et al., 2013)
Improve mental health systems	Core set of performance indicators for public mental health care (Lauriks et al., 2014)
Develop content of an intervention	Content of a basic mental health first aid course for adolescents to help their peers (Ross et al., 2012)
Improve medication use	Antipsychotic dosing (Gardner et al., 2010)
Improve caregiving	Caregiving towards a person with bipolar disorder (Berk et al., 2011)
Improve public action on prevention or early intervention	Parenting strategies for reducing the risk of adolescent depression and anxiety disorders (Yap et al., 2014)
Improve cultural competence	Culturally appropriate mental health first aid to an Aboriginal or Torres Strait Islander adolescent (Chalmers et al., 2014)
Develop policy	How tertiary education institutions can support students with a mental illness (Reavley et al., 2013)
Develop a health economic model	Cost-effectiveness of 3 antidepressants in major depressive disorder in the UK (Lenox-Smith et al., 2009)
Define a concept	Definition of relapse in schizophrenia (San et al., 2015)
Develop a scale	Item content for a geriatric depression inventory appropriate to Chinese culture (Xie et al., 2015)
Improve classification or diagnosis	Clinical subtypes of core premenstrual disorders (Ismail et al., 2013)
Determine epidemiology (prevalence, risk factors)	Global prevalence of dementia (Ferri et al., 2005)
Determine research priorities	Research priorities for public mental health in Europe (Forsman et al., 2015)
Improve brain imaging analysis	Protocol for manual hippocampal segmentation on magnetic resonance (Boccardi et al., 2015)

identified through authorship of articles in a systematic review, while clinicians were identified by searching a database on a professional society's website (Yap et al., 2014).

- In a study of the essential evidence-based components of first-episode psychosis services, the experts were identified through a systematic literature search and had to be a first author or lead author on at least one relevant publication in a peer-reviewed journal (Addington et al., 2013).

However, sometimes experts are harder to define and there is no clear sampling frame, in which case snowball sampling may be required. Here is an example:

- A study of mental health first aid for non-suicidal self-injury included a panel of consumers (Ross et al., 2014a). To locate potential panel members, approaches were made to depression and mental disorder advocacy organizations, and to consumers who had written peer support websites. Consumers who responded were asked to nominate others they knew as potential panel members. Consumers who were interested in participating were asked to give an expression of interest and provide an outline of their first-hand experiences of non-suicidal self-injury.

In most Delphi studies in the mental health area, the experts are professionals. However, increasingly other types of expertise are being recognized by the inclusion of consumer and caregiver advocates in Delphi panels. The type of experts to be included depends on the question being asked. Some Delphi studies are specifically focused on determining consumers' consensus on a topic, e.g. service users' preferences for treatment of psychosis (Byrne and Morrison, 2014). Other studies include consumers

along with professionals because of the diverse expertise they provide on the topic, e.g. how a member of the public could best assist a person who is suicidal (Ross et al., 2014b). On the other hand, there are areas where consumer or caregiver expertise would not be appropriate, e.g. how to segment the hippocampus (Boccardi et al., 2015).

The wisdom-of-crowds literature clearly shows that crowds make better decisions when they include diverse expertise (Page, 2007). This literature supports the value of including consumers and caregivers alongside professionals, and also the value of diversity within a professional panel (e.g. including members from a variety of mental health professions). However, diversity should not be seen as an end in itself. It only produces gains when the diverse members have relevant expertise. Adding school children to a panel, for example, would add sociodemographic diversity but not diverse expertise.

A number of Delphi studies have included separate professional and consumer panels and required consensus from both. These studies show that professionals and consumers have a surprisingly high level of agreement, despite the diverse sources of their expertise. In studies covering a range of topics, correlations in endorsement rates across Delphi questionnaire items have been 0.71 (Cairns et al., 2015), 0.75 (Reavley et al., 2013), 0.89 (Reavley et al., 2012), 0.91 (Ross et al., 2014a) and 0.92 (Ross et al., 2014b).

Determining expert panel size

Determining the size of a Delphi panel is an issue where there is little firm guidance. However, findings will be more stable with larger panels. Consider, for example, a panel of 10, where 80% agreement is required for consensus. One person's response represents 10% of the panel and can make a major difference to what items are endorsed. With progressively larger samples, each individual's responses will have less influence and findings will be more stable. One study of health-care quality and safety used bootstrap sampling to investigate the stability of response characteristics and found that a panel of 23 experts produced stable results (Akins et al., 2005). However, whether this can be generalized to other areas is unknown.

Additional evidence on stability of results comes from studies that have replicated findings across Delphi studies. A study on the appropriateness of various mental health treatments in primary care randomly assigned a group of general practitioners (GPs) to one of two Delphi panels, and a mixed group of GPs and mental health professionals were randomly assigned to another two panels (Hutchings et al., 2006). The GP panels (with 42 and 43 members) had a kappa for agreement of 0.88. For the mixed panels (with 42 members each), kappa was 0.90. Another replication was carried out in a Delphi study on suicide first aid which updated an earlier study on the same topic, with 94 items repeated in both studies (Ross et al., 2014b). Despite the use of different panels and a gap of 6 years, there was considerable stability of item endorsement frequencies. The item endorsement frequencies of 22 professionals in the earlier study correlated 0.84 with those of 41 professionals in the later study. Similarly, endorsement frequencies from a panel of 16 consumers and caregivers in the earlier study correlated 0.77 with 35 consumers from the later study.

While these studies show stability with panels of around 20 or more members, there are Delphi studies in the mental health research literature with much smaller panels. Such studies may produce unstable findings. In planning panel size, allowance also needs to be made for panel attrition, which is likely to occur across survey rounds. Panel attrition is likely to be larger in studies that have a long questionnaire and involve substantial time commitment.

Constructing the questionnaire

Delphi panel members need to be given a series of questionnaires, which have to be populated with items related to the research question. The aim is to include items which cover the complete domain of possibilities in the area, so a systematic approach is required. There are a number of potential sources of items for the initial questionnaire. The first is a systematic literature search, which may cover academic literature or grey literature on the Internet. Here is an example:

- A study to develop guidelines for tertiary education institutions on how to support students with a mental illness carried out a systematic review of websites, books and journal articles (Reavley et al., 2013). Each search was described in sufficient detail to allow replication. For example, the website search was described as follows:

  This involved a comprehensive search in Google search engines (www.google.com.au, www.google.co.uk, www.google.ca, www.google.com). The following search terms were entered into each: 'tertiary education OR higher education OR vocational education OR college OR campus AND depression OR anxiety OR mental disorders OR psychiatric disability'. The first 50 sites for each set of search terms were examined for statements about how institutions could support students with a mental illness. Any links that appeared on these web pages that the authors thought may contain useful information were followed.

Another common method is to source the initial set of questionnaire items from the expert panel or other stakeholders using qualitative methods, such as focus groups. Here is an example:

- A study to develop guidelines for adults on how to communicate with adolescents about mental health problems and other sensitive topics derived items

from a literature search and focus groups (Fischer et al., 2013). Two focus groups were carried out, one with clinicians and one with consumers from a youth mental health service. The clinician group was asked to reflect on what works and what does not in communicating with young people, while the consumer group was asked to think of a time when they were a teenager talking to an adult and they felt the adult communicated effectively with them.

The literature searches or qualitative methods will not directly result in questionnaire items. The content has to be analysed to derive concepts that are written into items. Thematic or content analysis methods may be useful for this purpose (Crowe et al., 2015). Here is an example from the study cited above about how to communicate with an adolescent:

- The first author

  transcribed the audio recordings from each focus group and extracted identified patterns of meaning to create potential themes ... Data around each potential theme were grouped into paragraphs. This process was replicated for text found in the literature search and combined with focus group data. Ideas within each paragraph were written as statements to create draft questionnaire items. This involved writing one idea per statement, with no ambiguity, written as an action, with minimal overlap with other items. A working party of the authors met to discuss and refine the draft items to ensure uniformity while trying to remain as faithful as possible to the original wording or source. (Fischer et al., 2013)

While systematic literature searches and focus groups are often used for the initial Delphi questionnaire, expert panel members are given the opportunity to suggest additional items when they are completing the questionnaire. These suggestions need to be evaluated by the research team to ensure that they are not already covered, that they are within the scope of the study and that they are clearly worded. These additional questions are then rated in subsequent survey rounds.

Delphi questionnaires can vary greatly in number of items. For longer questionnaires, the items can be grouped into themes to make it easier for panel members to make judgments and to spot any omissions. Here is an example:

- A study on antidepressant use in bipolar disorders derived items from a literature search on PubMed (Pacchiarotti et al., 2013). Statements on antidepressant use in bipolar disorder that could be useful to clinicians were derived from the content of the literature search and classified into six themes: acute treatment; maintenance treatment; monotherapy; switch to mania, hypomania, or mixed states and rapid cycling; use in mixed states; and drug class.

The questionnaire items need to be rated on a scale to show the extent of each expert's agreement. Here are some examples of rating scales:

- In a study of parenting strategies to prevent body dissatisfaction and unhealthy eating patterns in preschool children, experts were asked to rate statements describing parenting strategies as *Essential, Important, Don't Know/Depends, Unimportant* or *Should not be included* (Hart et al., 2014).
- A study of the development of post-disaster psychosocial care guidelines asked experts to rate statements on a 9-point scale, where 1 = *completely disagree*, 9 = *completely agree* and 5 = *neither* (Bisson et al., 2010).

Information provided to panel members to aid their judgments

When the questionnaire is administered to the expert panel members, some Delphi studies provide them with additional information to consider when making their ratings, whereas others leave it to panel members to draw on whatever sources of expertise are available to them. Typically, the additional information consists of reviews of the available evidence, but might also involve definitions of key concepts.

Whether or not evidence reviews are included depends both on the availability of evidence and on the nature of the judgments the panel is being asked to make. Here are three common situations:

1. The Delphi method is used because of the lack of other research evidence on the topic. Panel members are being asked to draw on their professional or personal experience, i.e. practice-based evidence. Examples are the development of guidelines on how a member of the public should assist a person who has been self-injuring (Ross et al., 2014a) and how to communicate with an adolescent about a mental health problem (Fischer et al., 2013).
2. There is evidence available, but it is incomplete or not suitable for translation into practice. Panel members are being asked to make decisions on the basis of imperfect evidence or to make judgments about practical implementation of the evidence. Examples are asking experts to make judgments about the global prevalence of dementia when prevalence data are variable across regions of the world (Ferri et al., 2005) and rating the helpfulness of self-help strategies for sub-threshold depression when only some strategies have been studied in trials and the conditions of those trials are different from real-life self-help (Morgan and Jorm, 2009).
3. Sometimes the Delphi method is used to develop a consensus of value judgments, in which case a

review of evidence may not be appropriate. Examples are determining research priorities (Forsman et al., 2015) and establishing the meaning of 'recovery' among individuals with experience of psychosis (Law and Morrison, 2014).

Administering the questionnaire

Because panel members make independent judgments, they do not have to meet. For this reason, questionnaires are typically administered by post or web survey software, with the latter now the norm. Web surveys make it possible to include experts from across the globe, adding to the potential diversity of expertise that can be drawn on.

Analysing rounds and providing feedback to the panel

Analysis of Delphi data requires a quantitative definition of 'consensus'. There is no single definition of consensus and it is up to the researcher to make a definition and give a rationale. There are a number of factors that might affect the definition used. For example, a study that assesses simultaneous agreement across multiple panels (e.g. professionals, consumers, caregivers) might have a lower cut-off for consensus than a study which involves a single expert panel. Similarly, the definition of consensus might be tighter for a study that aims to determine a small number of key statements of agreement than for one that aims to arrive at comprehensive and detailed guidance.

Here are some examples of how consensus has been defined:

- A study to develop guidelines for caregivers of people with bipolar disorder had separate expert panels of clinicians, caregivers and consumers, and required that each item had to have at least 80% endorsement as 'essential' or 'important' by each of the panels (Berk et al., 2011).
- A study to develop mental health first aid guidelines for Indigenous Australians required that an item had to have at least 90% endorsement as 'essential' or 'important' by a panel of Indigenous mental health experts (Hart et al., 2009).
- A study to develop post-disaster psychosocial care guidelines asked a mixed group of panelists to rate items on a 9-point scale from 'completely disagree' to 'completely agree' and required that an item had to have a mean score of >7 and 70% of panel members scoring 7 or above (Bisson et al., 2010).

Panel members are given feedback on how their responses compare to the rest of the panel and are asked to re-rate the items after considering the feedback. The feedback might consist of percentage endorsement of each item or mean score for each item on a Likert rating scale. The feedback can continue over several rounds of questionnaires, but the most common approach is to allow each item to be reconsidered only once. If the initial questionnaire has very many items, it may not be feasible to ask the panel members to rate each one again after feedback. Instead, items which are way off from the consensus criterion might be immediately eliminated and only those that come close to consensus are re-rated in a second round. Here is an example:

- In a study to determine parenting strategies for reducing the risk of adolescent depression and anxiety, there were 402 items in the Round 1 questionnaire (Yap et al., 2014). Consensus was defined as 90% or greater endorsement by the panel as 'essential' or 'important'. At Round 1, 168 items met this criterion, leaving 234 which did not. Rather than ask panelists to re-rate all 234 after receiving feedback, only the 116 items that reached 80–89% endorsement were presented for re-rating in Round 2, and the other 118 were immediately excluded.

While feedback is a traditional component of the Delphi method, it is not clear from the 'wisdom of crowds' literature that it improves judgments. There is some evidence that feedback on the group average when making estimates can undermine the 'wisdom of crowds' effect (Lorenz et al., 2011). However, this undermining occurs in situations where estimates have to be made of values with open-ended possibilities, like the border length of a country. With such estimates, extreme values are possible and outliers can have a distorting influence on the mean. Such distortions are unlikely in most Delphi studies, where the range of values is constrained by a Likert rating scale. Nevertheless, the role of feedback on judgments in Delphi studies is an area meriting further investigation.

Reporting the results

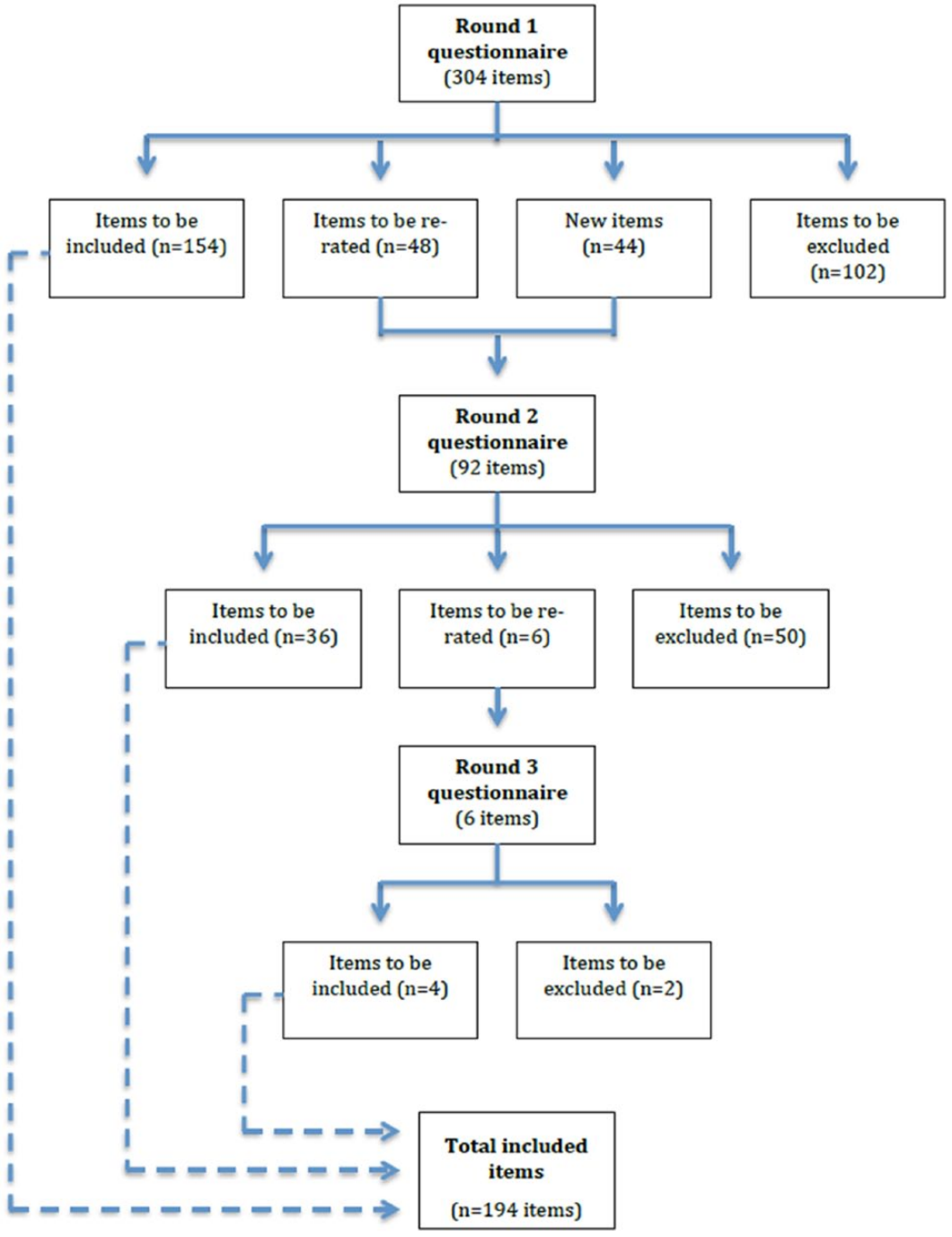
The results of the various rounds of a Delphi study can be complex to communicate. There is value in using a flowchart showing the fate of items at each round. An example is shown in Figure 1.

The simplest output from a Delphi study is a list of accepted items. In some cases, particularly where the number of accepted items is small, simply reporting the list may be sufficient. In other cases, where there are many accepted items, they may be reported under thematic headings or woven into a piece of guidelines text to aid communication.

Planning implementation

As with most research, publication of a Delphi study is unlikely in itself to produce any practical benefits. However, translation of research into practice is potentially easier

Figure 1. Example of a flowchart showing the number and outcomes of items in each Delphi round (Chalmers et al., 2014).



with Delphi studies than with many other types of research. This is because Delphi studies typically deal with questions that are closely related to practice needs. Furthermore, the expert panelists are potential stakeholders in any implementation and have been involved at an early stage.

In planning a Delphi study, it is useful to develop a plan for implementing the findings. The over-arching question needs to be: What will we do with the findings? If the researcher has no plans or capability to implement them, then a partnership with a potential implementer needs to be considered at an early stage.

Table 2 shows four examples of Delphi studies that were used to develop an intervention that was then evaluated for impact.

Table 2. Examples of successful implementation of Delphi findings.

Problem addressed	Delphi study	How implemented	Implementation outcomes
How to care for a person with bipolar disorder?	A Delphi study was carried out with panels of caregivers, consumers and clinicians (Berk et al., 2011)	Content of bipolarcaregivers website (www.bipolarcaregivers.org)	Surveys of website users found that most found it useful and many used the information to benefit caregiving (Berk et al., 2013)
What self-help strategies are helpful for sub-threshold depression?	A Delphi study was carried out with panels of depression consumers and professionals (Morgan and Jorm, 2009)	Messages used in email-based promotion of self-help for sub-threshold depression (MoodMemos) (Morgan et al., 2012)	In a randomized controlled trial, emails with self-help messages were found to produce a small improvement in depressive symptoms in people with sub-threshold depression (Morgan et al., 2012)
How to provide mental health first aid to people developing mental disorders or in mental health crises?	A series of Delphi studies was carried out with panels of professionals, consumers and caregivers on how to assist a person with a variety of mental health problems or crises (Jorm and Kitchener, 2011)	Curriculum of Mental Health First Aid training courses (www.mhfa.com.au)	Over 1% of Australian adults have received Mental Health First Aid training (Jorm and Kitchener, 2011). A meta-analysis of trials showed that training increases knowledge, reduces stigma and increases helping behaviours (Hadlaczky et al., 2014)
How to carry out internationally consistent manual segmentation of the hippocampus in studies of Alzheimer's disease?	A Delphi study was carried out with the developers of the 12 most frequently used protocols for hippocampal segmentation (Boccardi et al., 2015)	A web-based environment where human tracers can be trained and algorithm developers can test the performance of their products (Frisoni and Jack, 2015)	The Alzheimer's Association is planning the development of a certification procedure for automated hippocampal segmentation algorithms (Frisoni and Jack, 2015)

Conclusion

The Delphi method has been used widely in mental health research, often to answer questions that may not be possible or feasible with alternative methodologies. These include situations where experimental or epidemiological data are not available, where the data are incomplete or not directly applicable to the problem of interest, where the data are extensive but complex to draw conclusions from, or where a consensus of values is needed. Implementation of Delphi study outcomes has led to important advances in a range of practices in the mental health field.

While consensus methods are often thought of as producing a weak type of evidence in evidence-based medicine, this is an over-simplification. Delphi and other consensus methods underpin the methodologies of evidence-based medicine. Furthermore, expert consensus can be informed by a wide range of data, ranging from personal experience to systematic reviews of epidemiological and experimental studies. The strength of the evidence they provide depends on the sources of expertise available to the experts. Even where expertise is imperfect, the 'wisdom of crowds' literature supports the validity of group consensus judgments, provided certain conditions are met. Delphi methodology provides a systematic way of meeting these conditions that can be widely applied in mental health research.

Acknowledgements

The author thanks the following for suggestions on how to improve the paper: Amy Morgan, Laura Hart and Kathy Bond.

Declaration of interest

The author reports no conflicts of interest. The author alone is responsible for the content and writing of the paper.

Funding

The author is supported by a National Health and Medical Research Council (NHMRC) Fellowship.

References

Addington DE, McKenzie E, Norman R, et al. (2013) Essential evidence-based components of first-episode psychosis services. *Psychiatric Services* 64: 452–457.

Akins RB, Tolson H and Cole BR (2005) Stability of response characteristics of a Delphi panel: Application of bootstrap data expansion. *BMC Medical Research Methodology* 5: 37.

Begg C, Cho M, Eastwood S, et al. (1996) Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 276: 637–639.

Berk L, Berk M, Dodd S, et al. (2013) Evaluation of the acceptability and usefulness of an information website for caregivers of people with bipolar disorder. *BMC Medicine* 11: 162.

Berk L, Jorm AF, Kelly CM, et al. (2011) Development of guidelines for caregivers of people with bipolar disorder: A Delphi expert consensus study. *Bipolar Disorders* 13: 556–570.

Bisson JI, Tavakoly B, Witteveen AB, et al. (2010) TENTS guidelines: Development of post-disaster psychosocial care guidelines through a Delphi process. *British Journal of Psychiatry* 196: 69–74.

Boccardi M, Bocchetta M, Apostolova LG, et al. (2015) Delphi definition of the EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance. *Alzheimer's & Dementia* 11: 126–138.

Byrne R and Morrison AP (2014) Service users' priorities and preferences for treatment of psychosis: A user-led Delphi study. *Psychiatric Services* 65: 1167–1169.

Cairns KE, Yap MBH, Reavley NJ, et al. (2015) Identifying prevention strategies for adolescents to reduce their risk of depression: A Delphi consensus study. *Journal of Affective Disorders* 183: 229–238.

Chalmers KJ, Bond KS, Jorm AF, et al. (2014) Providing culturally appropriate mental health first aid to an Aboriginal or Torres Strait Islander adolescent: Development of expert consensus guidelines. *International Journal of Mental Health Systems* 8: 6.

Crowe M, Inder M and Porter R (2015) Conducting qualitative research in mental health: Thematic and content analyses. *Australian and New Zealand Journal of Psychiatry* 49: 616–623.

Dalkey NC (1969) *The Delphi Method: An Experimental Study of Group Opinion*. Santa Monica, CA: Rand.

Davis-Stober CP, Budescu DV, Dana J, et al. (2014) When is a crowd wise? *Decision* 1: 79–101.

De Mello G, Fraser F, Nicoll P, et al. (2013) Mental health care training for practitioners in remote and rural areas. *Clinical Teacher* 10: 384–388.

Ferri CP, Prince M, Brayne C, et al. (2005) Global prevalence of dementia: A Delphi consensus study. *The Lancet* 366: 2112–2117.

Fischer JA, Kelly CM, Kitchener BA, et al. (2013) Development of guidelines for adults on how to communicate with adolescents about mental health problems and other sensitive topics a Delphi study. *SAGE Open*. Epub ahead of print 15 December. DOI: 2158244013516769.

Forsman AK, Wahlbeck K, Aaro LE, et al. (2015) Research priorities for public mental health in Europe: Recommendations of the ROAMER project. *European Journal of Public Health* 25: 249–254.

Frisoni GB and Jack CR (2015) HarP: The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation. A standard of reference from a global working group. *Alzheimer's & Dementia* 11: 107–110.

Gardner DM, Murphy AL, O'Donnell H, et al. (2010) International consensus study of antipsychotic dosing. *American Journal of Psychiatry* 167: 686–693.

Goodyear M, Hill TL, Allchin B, et al. (2015) Standards of practice for the adult mental health workforce: Meeting the needs of families where a parent has a mental illness. *International Journal of Mental Health Nursing* 24: 169–180.

Hadlaczky G, Hökby S, Mkrtchian A, et al. (2014) Mental Health First Aid is an effective public health intervention for improving knowledge, attitudes, and behaviour: A meta-analysis. *International Review of Psychiatry* 26: 467–475.

Hart LM, Damiano SR, Chittleborough P, et al. (2014) Parenting to prevent body dissatisfaction and unhealthy eating patterns in preschool children: A Delphi consensus study. *Body Image* 11: 418–425.

Hart LM, Jorm AF, Kanowski LG, et al. (2009) Mental health first aid for Indigenous Australians: Using Delphi consensus studies to develop guidelines for culturally appropriate responses to mental health problems. *BMC Psychiatry* 9: 47.

Higgins JPT and Green S (2011) *Cochrane Handbook for Systematic Reviews of Interventions* (Version 5.1.0). The Cochrane Collaboration. Available at: www.cochrane-handbook.org (accessed 18 March 2015).

Hoffmann TC, Glasziou PP, Boutron I, et al. (2014) Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *British Medical Journal* 348: g1687.

Hutchings A, Raine R, Sanderson C, et al. (2006) A comparison of formal consensus methods used for developing clinical guidelines. *Journal of Health Services Research & Policy* 11: 218–224.

Ismail KM, Nevatte T, O'Brien S, et al. (2013) Clinical subtypes of core premenstrual disorders: A Delphi survey. *Archives of Women's Mental Health* 16: 197–201.

Joanna Briggs Institute and University of Adelaide (2013) New JBI levels of evidence. Available at: http://joannabriggs.org/assets/docs/approach/JBI-Levels-of-evidence_2014.pdf (accessed 18 May 2015).

Jönsson ML, Hahn U and Olsson EJ (2015) The kind of group you want to belong to: Effects of group structure on group accuracy. *Cognition* 142: 191–204.

Jorm AF and Kitchener BA (2011) Noting a landmark achievement: Mental Health First Aid training reaches 1% of Australian adults. *Australian and New Zealand Journal of Psychiatry* 45: 808–813.

Lauriks S, de Wit MA, Buster MC, et al. (2014) Composing a core set of performance indicators for public mental health care: A modified Delphi procedure. *Administration and Policy in Mental Health* 41: 625–635.

Law H and Morrison AP (2014) Recovery in psychosis: A Delphi study with experts by experience. *Schizophrenia Bulletin* 40: 1347–1355.

Lenox-Smith A, Greenstreet L, Burslem K, et al. (2009) Cost effectiveness of venlafaxine compared with generic fluoxetine or generic amitriptyline in major depressive disorder in the UK. *Clinical Drug Investigation* 29: 173–184.

Lorenz J, Rauhut H, Schweitzer F, et al. (2011) How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences USA* 108: 9020–9025.

Moher D, Liberati A, Tetzlaff J, et al. (2009) Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Journal of Clinical Epidemiology* 62: 1006–1012.

Morgan AJ and Jorm AF (2009) Self-help strategies that are helpful for sub-threshold depression: A Delphi consensus study. *Journal of Affective Disorders* 115: 196–200.

Morgan AJ, Jorm AF and Mackinnon AJ (2012) Email-based promotion of self-help for subthreshold depression: Mood Memos randomised controlled trial. *British Journal of Psychiatry* 200: 412–418.

Pacchiarotti I, Bond DJ, Baldessarini RJ, et al. (2013) The International Society for Bipolar Disorders (ISBD) task force report on antidepressant use in bipolar disorders. *American Journal of Psychiatry* 170: 1249–1262.

Page SE (2007) *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton, NJ and Oxford: Princeton University Press.

Pincus T, Miles C, Froud R, et al. (2011) Methodological criteria for the assessment of moderators in systematic reviews of randomised controlled trials: A consensus study. *BMC Medical Research Methodology* 11: 14.

Reavley NJ, Ross A, Killackey EJ, et al. (2012) Development of guidelines to assist organisations to support employees returning to work after an episode of anxiety, depression or a related disorder: A Delphi consensus study with Australian professionals and consumers. *BMC Psychiatry* 12: 135.

Reavley NJ, Ross AM, Killackey E, et al. (2013) Development of guidelines for tertiary education institutions to assist them in supporting students with a mental illness: A Delphi consensus study with Australian professionals and consumers. *PeerJ* 1: e43.

Ross AM, Hart LM, Jorm AF, et al. (2012) Development of key messages for adolescents on providing basic mental health first aid to peers: A Delphi consensus study. *Early Intervention in Psychiatry* 6: 229–238.

Ross AM, Kelly CM and Jorm AF (2014a) Re-development of mental health first aid guidelines for non-suicidal self-injury: A Delphi study. *BMC Psychiatry* 14: 236.

Ross AM, Kelly CM and Jorm AF (2014b) Re-development of mental health first aid guidelines for suicidal ideation and behaviour: A Delphi study. *BMC Psychiatry* 14: 241.

Sackett DL and Rosenberg WM (1995) The need for evidence-based medicine. *Journal of the Royal Society of Medicine* 88: 620–624.

San L, Serrano M, Canas F, et al. (2015) Towards a pragmatic and operational definition of relapse in schizophrenia: A Delphi consensus approach. *International Journal of Psychiatry in Clinical Practice* 19: 90–98.

Surowiecki J (2004) *The Wisdom of Crowds: Why the Many Are Smarter Than the Few*. London: Abacus.

Tetzlaff JM, Moher D and Chan AW (2012) Developing a guideline for clinical trial protocol content: Delphi consensus survey. *Trials* 13: 176.

Verhagen AP, de Vet HC, de Bie RA, et al. (1998) The Delphi list: A criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *Journal of Clinical Epidemiology* 51: 1235–1241.

Xie Z, Lv X, Hu Y, et al. (2015) Development and validation of the geriatric depression inventory in Chinese culture. *International Psychogeriatrics* 27: 1505–1511.

Yap MB, Pilkington PD, Ryan SM, et al. (2014) Parenting strategies for reducing the risk of adolescent depression and anxiety disorders: A Delphi consensus study. *Journal of Affective Disorders* 156: 67–75.