Barrios, M., Guilera, G., Nuño, L., Gómez-Benito, J. Consensus in the delphi method: What makes a decision change?, *Technological Forecasting and Social Change*, 2020, https://doi.org/10.1016/j.techfore.2020.120484.

Preprint version

Consensus in the Delphi method: What makes a decision change?

Maite Barrios*[1,2], Georgina Guilera[1,2], Laura Nuño[2,3], Juana Gómez-Benito[1,2]

1. Department of Social Psychology and Quantitative Psychology, University of Barcelona, Barcelona, Spain. 2. Group on Measurement Invariance and Analysis of Change (GEIMAC), Institute of Neurosciences, University of Barcelona, Barcelona, Spain, 3. Clinical Institute of Neurosciences (ICN), Hospital Clinic, Barcelona, Spain.

***Corresponding author**: Maite Barrios, mbarrios@ub.edu

Abstract

This study aimed to examine the factors in a Delphi study that provoke a shift of opinion between rounds, favoring consensus among participants. To this end, the influence that feedback about the level of group agreement has on individual opinion change was first compared across different levels of group agreement. We then assessed the effect of sociodemographic and professional variables related to participants. Five three-Delphi round studies were conducted independently, in which a total of 777 mental health experts participated. In each study, the percentage of group agreement obtained in the second round was shown as controlled feedback in the third round, and responses in the second and third rounds were considered in order to analyze opinion change. Results showed that when the feedback given in round 3 indicated group agreement of 75% or higher, there was a further shift in opinion towards the group opinion (i.e., greater consensus among participants), whereas if the feedback indicated less than 75% group agreement, individual opinions tended to shift against the group opinion (i.e., less consensus among participants). Neither sociodemographic nor professional variables had a significant effect in explaining the shift of opinion. These results show that in the context of a Delphi study, controlled feedback has an influence on individual responses and the achievement of consensus.

Keywords: Delphi method, opinion change, consensus threshold, group agreement, feedback effect.

Highlights

- Delphi consensus is assessed by the shift of opinions between rounds.
- Feedback influences participants' responses and the level of agreement reached.
- Consensus increases after feedback indicating at least 75% group agreement.
- Consensus decreases after feedback indicating less than 75% group agreement.
- Achieving consensus depends on the effect of controlled feedback.

1. Introduction

The Delphi method has been widely used in recent decades and is now regarded in academic research as a valuable technique for reaching consensus about specific issues when empirical evidence is scarce or contentious. The method aims to achieve consensus about a specific topic by using several rounds of questionnaires to collect data from a panel of selected experts on the topic of interest (Keeney et al., 2006), and it has been employed across numerous disciplines, including medicine (Sinha et al., 2011; Prinsen et al., 2014), nursing (Keeney et al., 2006), psychology (van der Vaart et al. 2014), education (Zawacki-Richter, 2009), and business (Jiang et al., 2017; El-Gazzar et al., 2016). Although a number of variants of the Delphi method have been proposed, inherent characteristics of the technique include anonymity among respondents and a controlled feedback process provided in a series of rounds (Linstone & Turoff, 1975). Anonymity is guaranteed since the process is coordinated by a research team using, in most cases, an online platform or e-mail, thus avoiding any interaction between participants. After the first round the research team analyzes and summarizes the responses of the panel of experts in order to provide feedback to participants for the following round.

Whereas anonymity reduces the effect of dominant individuals on participants' responses, the use of controlled feedback encourages experts to reassess their initial judgments based on the information provided by the research team in each round. Feedback thus allows each participant to generate additional insights about the specific questions or items and, consequently, change his or her responses in light of the group's opinion. This means that the results regarding specific items can vary across rounds, thereby favoring the convergence of opinions (Linstone & Turoff, 1975). In this respect, the Delphi method is well suited as a consensus-building technique.

Consensus, however, is one of the most controversial methodological issues in the Delphi process. Although achieving consensus among participants is a key feature in Delphi studies, what is accepted as consensus, or how it is reached, remains unclear (Von der Gracht, 2012; Boulkedid et al., 2011). While it may be argued that the Delphi process can be iterated until consensus is reached, Delphi studies tend to conduct a specific number of rounds with the aim of eliciting consensus among participants, without, however, a formal definition of what is going to be considered consensus or even without specifying a threshold value that determines when consensus has been achieved (Humphrey-Murto et al., 2017; Boulkedid et al., 2011; Diamond et al., 2014; Foth et al., 2016).

Although there are several approaches to reporting consensus (Keeney et al., 2006; Von der Gracht, 2012; Boulkedid et al., 2011), the most common is to use a certain percentage of group agreement (Diamond et al., 2014; Foth et al., 2016). However, a threshold percentage is not always provided a priori in most Delphi studies, and the range reported as an accepted consensus is very wide (50 – 97%); the median threshold accepted as consensus is 75% agreement among participants (Diamond et al., 2014; Foth et al., 2016).

In addition to the uncertainty about what should be considered consensus, little attention has been paid to what factors may favor the achievement of a consensus response, for instance, the specific characteristics of the panel of experts (e.g., perceived expertise, years of experience, gender, age) or the influence of controlled feedback on individual opinion change (Makkonen et al., 2016; Meijering and Tobi, 2016; Meijering and Tobi, 2018; Rowe et al., 2005). Thus, although anonymity among participants, which is a key feature of the method, can reduce the effect of dominant individuals on participants' responses, several studies have found that feedback may influence opinion

change among participants, who tend to follow the view of the majority (Scheibe et al., 1975; Rowe et al., 2005; Makkonen et al., 2016; Meijering and Tobi, 2018). Moreover, the kind of feedback provided to participants (i.e., argumentative or statistical) has also been associated with the tendency to shift opinion between Delphi rounds. Specifically, it has been found that argumentative feedback (i.e., reasons and justifications), as opposed to statistical feedback (e.g., median and range of group responses), provokes less opinion change among participants (Rowe & Wright, 1996).

In line with previous studies, our hypothesis here was that when the feedback given indicates that consensus is above a certain threshold, participants who have previously selected an option contrary to the majority will be more likely to revise their initial opinion. Thus, given that any consensus threshold can be questioned, even if it is used as a common threshold, we sought to explore the extent to which feedback may have an influence on the Delphi process, since a greater understanding of its impact could help to better determine the most suitable consensus threshold. More specifically, our aim was to examine the factors that provoke opinion change between Delphi rounds and which favor consensus among participants, focusing specifically on the influence of feedback and certain characteristics of participants. To this end, the influence that feedback about the level of group agreement has on individual opinion change was first compared across different levels of group agreement. We then assessed the effect of sociodemographic and professional variables related to participants.

Method

Participants

Experts were defined as health professionals (i.e., psychiatrists, psychologists, nurses, occupational therapists, and social workers) with at least one year of experience in the

direct treatment of individuals with schizophrenia. Several strategies were used to recruit experts from around the world: International professional associations, universities with healthcare professional training programs, and hospitals were contacted, and we also made use of literature searches, LinkedIn contacts, and personal recommendations. In order to avoid language barriers and to increase the representativeness and participation of experts from around the world, all study materials (i.e., contact letters, questionnaires, etc.) were available in five languages (i.e., Chinese, English, French, Russian, and Spanish) and participants could choose the language in which they wished to respond. All the study materials were translated and supervised by at least two independent native speakers. The initial contact included an invitation to take part and a detailed description of the project targets, the Delphi process, and the timeline. Demographic and professional data were also requested. Of the 1,555 health professionals who agreed to participate, 777 experts finally participated in the first round.

Delphi process

Five independent worldwide Delphi studies were carried out, following the same research design so as to ensure a high level of comparability (Nuño et al., 2018, Nuño et al., 2019a; Nuño et al., 2019b). Each Delphi study comprised three rounds, was conducted through an online survey system (www.qualtrics.com), and lasted around three months. For each round, participants always had two weeks to respond, and three reminders were sent per round: the first, one week before the deadline; the second, two days before the deadline; and the third, on the deadline day itself. Participants were able to answer parts of the survey at different times, and the expected completion time for each survey round was about 15 minutes.

Each Delphi study began with six open-ended questions about issues related to functioning in schizophrenia. All the responses collected in this first round were linked to categories in the International Classification of Functioning, Disability, and Health (ICF) by two health professionals trained in the use of the ICF and with experience of providing care to individuals with schizophrenia (further details about this process can be consulted in Nuño et al. (2018, 2019a, 2019b). Those categories reported by at least 5% of the experts were selected for inclusion in the second Delphi round. In this second round, the experts received a list of the selected ICF categories, along with their respective definitions. Participants were then asked to judge, for each category, whether they thought the category was relevant from their professional perspective to the assessment and/or treatment of individuals with schizophrenia, taking into account that the final list should be as short as possible to be practical but as comprehensive as necessary to capture the most relevant needs of this population. Each participant judged between 160 and 184 categories. In the third round, participants were once again asked to judge each category, but this time they were given feedback (for each category) about the responses of the expert panel as a whole (i.e., the percentage of participants who had considered the category relevant in the second round), as well as a reminder of their own previous response. Thus, in this third round, participants had the opportunity to consider the panel's opinion, to revise their previous responses, and to respond again to the list of categories. The whole Delphi process is summarized in Figure 1.

Insert Figure 1 about here

Data analysis

Descriptive statistics were used to describe the sociodemographic and professional characteristics of participants. In order to examine the effect of feedback on responses in

the third round, we calculated the percentage of categories that were rated as "*relevant*" in the third Delphi round for different intervals of group agreement feedback (e.g., less than 50% of group agreement, 50-54.9% of group agreement, etc.). Group agreement feedback was also considered in order to study the shift in individual opinions between the second and third rounds. Specifically, for each of the intervals of group agreement considered in the previous analysis we calculated the percentage of categories for which the degree of consensus changed (either increasing or decreasing, based on participants' individual responses) between rounds 2 and 3. We also analyzed what we termed "congruent" and "incongruent" opinion change in round 3 following feedback about the level of group agreement achieved in round 2. An opinion change was considered congruent when a participant changed his/her response in order to make it congruent with the group opinion (e.g., a participant rates a category as "*not relevant*" in the second round, but after receiving the group agreement feedback indicating that the category is mainly considered "*relevant*", the participant decides to change his/her response in the third round and rates the category as "*relevant*"). By contrast, an incongruent opinion change occurred when a participant changed his/her response in the opposite direction to the group opinion (e.g., a participant rates a category as "*relevant*" in the second round and the group agreement feedback received in the third round shows high group agreement, indicating that the category is also considered "*relevant*" by most of the experts; however, in the third round the participant decides to change his/her response and rates the category as "*not relevant*"). Similarly, a congruent non-change of opinion occurred when a participant did not change his/her response in the third round because it was already congruent with the group opinion (e.g., a participant rates the category as "*relevant*" in the second round and the group agreement feedback received in the third round shows high group agreement, indicating that the category is

also considered "*relevant*" by most of the experts; thus, congruently, the participant decides to maintain his/her response and again rates the category as "*relevant*"). Finally, an incongruent non-change of opinion occurred when a participant did not change his/her response in the third round despite it being incongruent with the group opinion (e.g., a participant rates the category as "*not relevant*" in the second round, but the group agreement feedback received in the third round shows high group agreement, indicating that the category is considered "*relevant*" by most of the experts; however, the participant, incongruently, decides to maintain his/her response and rates the category as "*not relevant*").

The effect on shifts in opinion of variables including age, gender, profession, professional experience, perceived expertise (rated using a 5-point Likert scale), and the participant's geographical region of origin was also assessed using multiple linear regression.

Results

An overview of the five independent Delphi studies is presented in Table 1. A total of 777 health professionals completed the first Delphi round, of whom 628 completed the second and third rounds. This implies a response rate across rounds one to three of 80.8%. Data from those participants who completed both the second and the third round were used for the analysis conducted in this study. The sociodemographic and professional data of these participants are shown in Table 2.

Table 1. Overview of the five Delphi studies.

Delphi study	Professional profile	Participants round 1 n (%)	Participants round 3 n (%)	Number of categories rated
1	Psychiatrists	352 (45.30)	303 (48.25)	166
2	Psychologists	175 (22.52)	137 (21.82)	176
3	Social workers	57 (7.34)	36 (5.73)	160
4	Occupational therapists	92 (11.84)	73 (11.62)	184
5	Nurses	101 (13)	79 (12.38)	177
Total		777	628 (80.82)	863

Table 2. Demographic and professional characteristics of participants in the second and third Delphi rounds.

Age (years) *mean (SD)*	45.59 (15.9)
Gender (Female) *n (%)*	302 (48.09)
World region *n (%)*	
Africa	39 (6.21)
Americas	142 (22.61)
Eastern Mediterranean	38 (6.05)
Europe	200 (31.85)
South-East Asia	102 (16.24)
Western Pacific	106 (16.88)
Professional experience (years) *mean (SD)*	18.8 (10.6)
Perceived expertise *mean (SD)*	4.00 (0.88)

Note: SD = standard deviation. Perceived expertise: Self-rating of schizophrenia expertise from 1 = limited expertise to 5 = extensive expertise.

In terms of participants' responses regarding the relevance of each category, 95,547 category ratings (88.8%) did not change between the second and the third round, whereas for the remaining 12,038 (11.2%) a change was observed. On average, the proportion of opinion changes from the second to the third round per participant was quite low (Median 0.08; min. 0.0, max. 0.82). This means that 50% of participants changed their opinion for less than 8% of the categories they had to rate.

Data showed that feedback in round 3 about the level of group agreement achieved in round 2 had an effect on category ratings in the third round, causing a shift of opinions. Figure 2 plots the percentage of categories selected as "relevant" in round 3 for each interval of group agreement feedback.

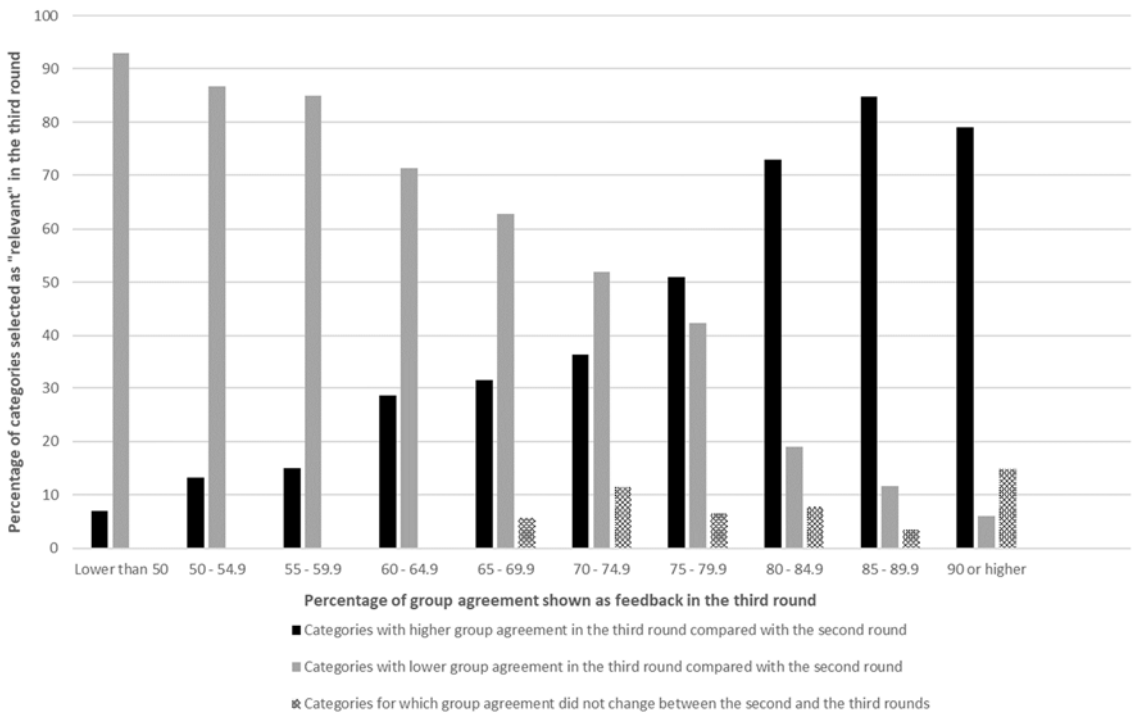

Figure 2. Effect of feedback about the group agreement achieved in round 2 on category ratings in round 3.

It can be seen that as the percentage of group agreement (which is the feedback given to participants in round 3) increases, so too does the percentage of categories yielding greater group agreement in round 3 compared with round 2 (i.e., consensus increases between rounds 2 and 3). Conversely, when the feedback given in round 3 indicated lower levels of group agreement in round 2, the categories in question were more likely to yield even less consensus in round 3. More specifically, the figure shows that once the feedback given to participants in round 3 indicated group agreement of 75% or higher in round 2, then consensus was achieved for an increasing proportion of categories in round 3. The opposite effect can be observed for levels of group agreement feedback below 75%. It can also be seen in Figure 2 that once feedback in round 3 indicated at least 65% group agreement in round 2, a small proportion of categories showed no change in their relevance rating.

Figure 3 shows how the difference in the percentage of group agreement between rounds 2 and 3 increases as we move away from the threshold of 75% consensus. Above and below this threshold, group agreement becomes, respectively, progressively stronger and weaker between rounds 2 and 3.

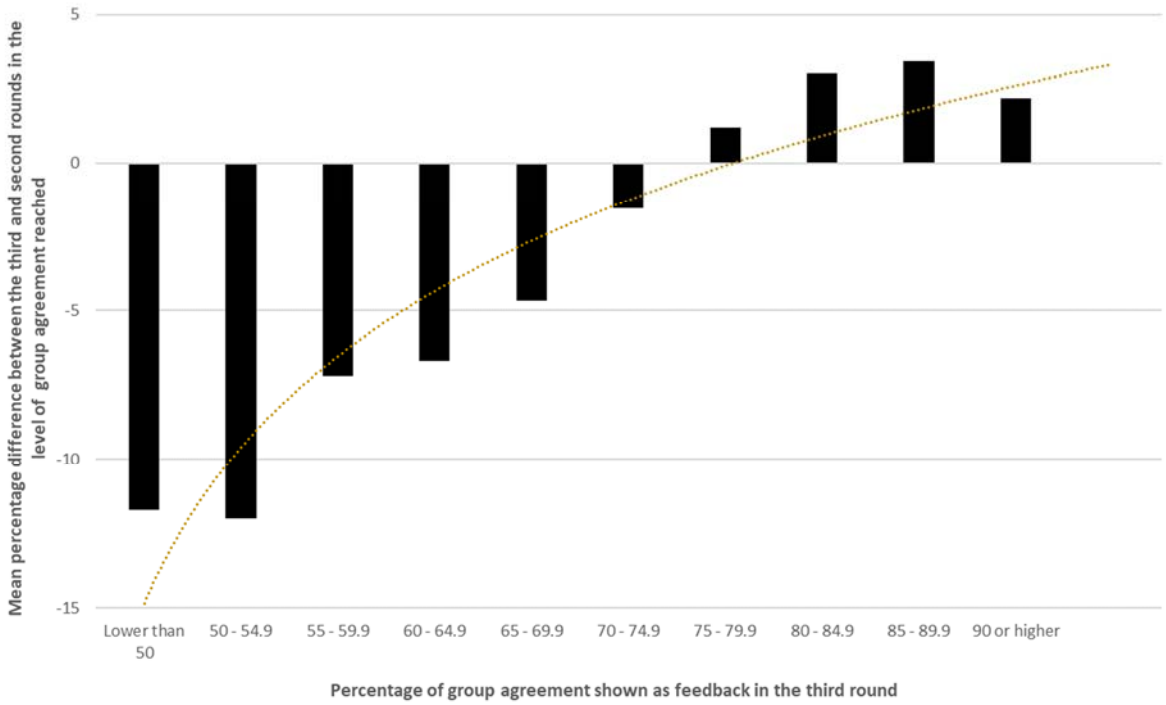

Figure 3. Mean percentage difference between the third and second rounds in the level of consensus achieved, according to each interval of group agreement feedback.

Based on the 75% threshold of group agreement, calculation of an odds ratio showed that when a category achieved group agreement of at least 75% in round 2 the probability of its group agreement increasing in round 3 was higher than that of a category with group agreement below 75% in round 2 (*OR* = 15.327; 95% CI = 10.198 – 23.036; $p < .0001$). We also calculated the percentage of congruent and incongruent shifts of opinion for each participant. Data showed that when the feedback given in round 3 indicated group agreement below 75% in round 2, the categories in question were associated with a significantly higher percentage of opinion change in the third round (both congruent and incongruent). Conversely, when the feedback in round 3 indicated group agreement of at least 75%, participants were significantly more likely to maintain the same response in the third round (i.e., congruent with the group agreement

feedback they had received). Table 3 shows the percentage shift of opinion in each condition based on this 75% threshold.

Table 3. Mean percentage of change of opinion based on the 75% threshold of group agreement.

Change of opinion	Group agreement of 75% or higher	Group agreement lower than 75%
	Mean (SD) [95% CI]	Mean (SD) [95% CI]
Non-change (congruent)	84.4 (15.0) [83.3 – 85.6]	40.2 (22.4) [38.5 – 42.0]
Non-change (incongruent)	6.4 (9.2) [5.7 – 7.2]	39.9 (22.2) [38.2 – 41.7]
Change (congruent)	5.5 (8.3) [4.8 – 6.1]	13.7 (16.7) [12.4 – 15.0]
Change (incongruent)	3.7 (8.5) [3.0 – 4.3]	6.2 (9.6) [5.4 – 6.9]

Note: SD = Standard deviation. CI = Confidence interval.

Multiple regression showed the presence of collinear predictors, and hence perceived expertise was removed from the analysis. Neither sociodemographic nor professional variables had a significant effect in terms of explaining the shift of opinion. Table 4 summarizes the results of the regression analysis.

Discussion

This paper explores the process through which consensus is reached in a Delphi study, focusing specifically on the influence of controlled feedback on participants' responses across rounds. Overall, our results indicate that providing participants with feedback about the level of group agreement reached in the previous round has an effect on the level of consensus that is achieved subsequently.

The power of feedback has been demonstrated by several studies (Scheibe et al., 1975; Makkonen et al., 2016; Bolger et al., 2011; Rowe et al., 2005), and it has also

been shown that Delphi participants who change their opinion are more likely to shift towards the majority group opinion (Makkonen et al., 2016; Bolger et al., 2011; Rowe et al., 2005). However, our findings indicate that the effect of controlled feedback about group agreement depends on the level of agreement that is shared with participants as feedback. Specifically, and based on the large number of category ratings analyzed, our study has been able to identify a specific threshold on either side of which the response trend differed. When the feedback given in round 3 indicated that at least 75% group agreement had been achieved in round 2, those participants who had not regarded a category as relevant in the previous round tended to shift their view towards the majority opinion. By contrast, when feedback indicated less than 75% group agreement in round 2, participants who had previously rated a category as relevant were more likely to change their response in round 3. This finding adds to previous evidence that when the percentage of group agreement is shared with experts as controlled feedback, their views tend to shift towards the majority opinion only when support is strong among the group as a whole. Conversely, if the level of group agreement is not perceived as indicating strong group support, participants are more likely to change their opinion and stop supporting the majority view, thus hindering the achievement of consensus. Our results suggest that feedback indicating group agreement of at least 75% tends to favor even greater consensus, whereas below the 75% threshold, feedback may make consensus less likely. Interestingly, systematic reviews by Diamond et al. (2014) and Foth et al. (2016) both found that 75% agreement was commonly accepted as a consensus threshold in Delphi studies.

Importantly, our data also suggest that feedback about the level of group agreement may have a progressively greater effect the further away agreement is from the 75% threshold. Thus, in the Delphi studies analyzed here, the highest levels of group

agreement feedback led to even greater consensus among the panel of experts. By contrast, as the level of group agreement decreased further below the 75% threshold, experts were increasingly likely to change their mind regarding the relevance of a category. A possible explanation for this finding is that the experts were specifically told that the final list of categories should be as short as possible but comprehensive enough to capture the needs of individuals with schizophrenia. This instruction may have encouraged some experts to change their mind when the level of group agreement appeared to fall short of a consensus.

In line with some previous studies (Bolger et al., 2011; Kauko & Palmroos, 2014; Makkonen et al., 2016), we found no evidence of a relationship between opinion change and sociodemographic and professional characteristics of participants, including for variables related to professional experience that have been regarded, theoretically, as being associated with shifts of opinion among Delphi participants. This finding suggests that sociodemographic and professional characteristics of panelists are of no relevance in a consensus-building process. However, given that our results are contrary to those of other authors (Mulgrave & Ducanis, 1975; Rowe & Wright, 1996; Rowe & Wright, 1999; Yaniv & Milyavsky, 2007) who reported a relationship between high relative expertise and low propensity to change opinion over rounds, further experimental studies are needed to shed light on this issue. An important point to note in this context is that even if the characteristics of the panel are not important in terms of reaching a consensus, the quality of the Delphi process and its results do depend on the appropriate recruitment and selection of qualified experts who are able to provide a truly representative view of the issue under investigation (Keeney et al., 2006; Donohoe & Needham, 2009; Sobaih et al., 2012).

Regarding opinion change, it is worth noting that the percentage of opinion shift between rounds in the Delphi studies analyzed here is low compared with the number of responses that did not change between the second and third rounds. It should be noted that across the five Delphi studies a large number of categories reached the agreed consensus threshold (i.e., ≥75% group agreement) in the second round, and hence participants did not need subsequently to change their opinion insofar as their initial opinion was already congruent with the majority view. Moreover, these were real-world Delphi studies involving participants with considerable expertise in relation to the issue they were asked to rate, and this, as Hussler et al. (2011) point out, makes it less likely that experts will change their original opinions. A further point to consider is that each of the five Delphi studies followed the recommendation to provide experts, in the third round, with information about their own ratings in the previous round (Boulkedid et al., 2011; Keeney et al., 2006; Murphy et al., 1998), an approach which, according to Meijering and Tobi (2018), is associated with less opinion change.

A key strength of the present analysis is that it is based on five real-world Delphi studies involving a large sample of experts from around the world, each with extensive experience concerning the issue they were asked to appraise. This ecological validity compensates to some extent for the lack of experimental control, since as several authors have pointed out (Rowe & Wright 1999; Rowe et al., 1991; Meijering & Tobi, 2018), some experimental Delphi studies derive their findings from samples of students who are asked to make judgments about topics on which they cannot be considered experts.

However, several other potential weaknesses should also be mentioned. First, our analysis is based on the shift of opinion between two rounds, and further studies involving more rounds are needed to confirm the feedback effect we observed here.

Second, although controlled feedback in the form of information about the percentage of group agreement is frequently used in Delphi studies (Diamond et al., 2014), our findings need to be corroborated by studies that examine the effect of other kinds of feedback, for example, measures of central tendency and dispersion (e.g., mean, median, range, interquartile range), or argumentative feedback (e.g., justifications, reasons). Third, the present study provides no insight into why experts changed their opinion, although our findings may be useful for predicting, based on the feedback given to the participants, whether a high level of group agreement might be achieved in the next round or, on the contrary, whether consensus is unlikely.

The results of this study have a number of implications. First, researchers in any Delphi study need to be aware that controlled feedback in a given round can, depending on the information provided, have different effects on participants' subsequent responses. Thus, although the feedback effects observed in this study strongly support the use of the Delphi technique for consensus building, since high group agreement favors even greater consensus, decision makers who use the technique for the purposes of forecasting should also consider, when drawing conclusions and making recommendations based on their results, that giving controlled feedback has the potential to introduce desirability bias (Ecken et al., 2011). Consequently, using the Delphi method in this context does not necessarily mean that forecasting accuracy will be improved, even if greater consensus is achieved. Further research is required to assess the effect of controlled feedback when the Delphi technique is used as a forecasting method.

Another implication of our analysis is that by studying opinion change across rounds it is possible to determine whether consensus was present from the outset or was only achieved as a result of feedback being given. Consideration of these issues is

important with regard to the validity and reliability of the panel's final decision, and it brings greater transparency to the decision-making process. This is why it is highly advisable, when conducting a Delphi study, to establish a priori a threshold above which consensus is considered to have been reached. The recommended threshold based on our results would be 75% agreement, since the pattern of responses differs on either side of this level of consensus. When feedback indicates group agreement of at least 75% in the previous round, it is likely that even greater consensus will be achieved in the next round, whereas consensus over a given item will weaken if participants are told that agreement is below 75%.At all events, further studies are needed to confirm this pattern.

Conclusions

Based on our findings in this study we conclude that the likelihood of opinion change among participants in a Delphi study is influenced by the controlled feedback they receive. By contrast, the sociodemographic and professional characteristics of the panel of experts appear to be of no relevance in this respect. Importantly, the effect of controlled feedback depends on the level of agreement that is shared as feedback, and thus it may facilitate or hinder the consensus-building process. When the feedback given indicates strong agreement among the group as a whole, participants tend to shift towards the majority opinion, whereas when the feedback is not perceived as indicating strong group support, opinions are more likely to change in a way that hinders consensus. Our data indicate that group agreement of 75% acts as a threshold, since the pattern of responses observed differs on either side of this level of consensus. More specifically, consensus increases when feedback indicates group agreement of at least 75% and decreases when it is less than 75%. This finding highlights the importance of

looking at the consensus-building process across Delphi rounds in order to ensure that the decisions made are valid and reliable.

Funding

This work was supported by Spain's Ministry of Economy and Competitiveness [grant PSI2015–67984-R], and by the Agency for the Management of University and Research Grants of the Government of Catalonia [grant 2017SGR1681].

References

Bolger, F., Stranieri, A., Wright, G., & Yearwood, J. (2011). Does the Delphi process lead to increased accuracy in group-based judgmental forecasts or does it simply induce consensus amongst judgmental forecasters? *Technological Forecasting and Social Change*, *78*(9), 1671–1680. https://doi.org/10.1016/j.techfore.2011.06.002

Boulkedid, R., Abdoul, H., Loustau, M., Sibony, O., & Alberti, C. (2011). Using and reporting the Delphi method for selecting healthcare quality indicators: A systematic review. *PLoS ONE*, *6*(6), e20476. https://doi.org/10.1371/journal.pone.0020476

Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology*, *67*(4), 401–409. https://doi.org/10.1016/j.jclinepi.2013.12.002

Donohoe, H., & Needham, R. D. (2009). Moving best practice forward: Delphi characteristics, advantages, potential problems, and solutions. *International Journal of Tourism Research*,11(5), 415–437.

Ecken, P., Gnatzy, T., & von der Gracht, H. A. (2011). Desirability bias in foresight: Consequences for decision quality based on Delphi results. *Technological Forecasting and Social Change*, *78*(9), 1654–1670. https://doi.org/10.1016/j.techfore.2011.05.006

El-Gazzar, R., Hustad, E., & Olsen, D. H. (2016). Understanding cloud computing adoption issues: A Delphi study approach. *Journal of Systems and Software*, *118*, 64–84. https://doi.org/10.1016/J.JSS.2016.04.061

Foth, T., Efstathiou, N., Vanderspank-Wright, B., Ufholz, L.-A., Dütthorn, N., Zimansky, M., & Humphrey-Murto, S. (2016). The use of Delphi and Nominal

Group Technique in nursing education: A review. *International Journal of Nursing Studies*, *60*, 112–120. https://doi.org/10.1016/j.ijnurstu.2016.04.015

Humphrey-Murto, S., Varpio, L., Wood, T. J., Gonsalves, C., Ufholz, L.-A., Mascioli, K., … Foth, T. (2017). The use of the Delphi and other consensus group methods in medical education research. *Academic Medicine*, *92*(10), 1491–1498. https://doi.org/10.1097/ACM.0000000000001812

Hussler, C., Muller, P., & Rond, P. (2011). Is diversity in Delphi panelist groups useful? Evidence from a French forecasting exercise on the future of nuclear energy. *Technological Forecasting and Social Change*, *78*(9), 1642–1653. https://doi.org/10.1016/j.techfore.2011.07.008

Jiang, R., Kleer, R., & Piller, F. T. (2017). Predicting the future of additive manufacturing: A Delphi study on economic and societal implications of 3D printing for 2030. *Technological Forecasting and Social Change*, *117*, 84–97. https://doi.org/10.1016/J.TECHFORE.2017.01.006

Kauko, K., & Palmroos, P. (2014). The Delphi method in forecasting financial markets: An experimental study. *International Journal of Forecasting*, *30*(2), 313–327. https://doi.org/10.1016/j.ijforecast.2013.09.007

Keeney, S., Hasson, F., & McKenna, H. (2006). Consulting the oracle: Ten lessons from using the Delphi technique in nursing research. *Journal of Advanced Nursing*, *53*(2), 205–212. https://doi.org/10.1111/j.1365-2648.2006.03716.x

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

Makkonen, M., Hujala, T., & Uusivuori, J. (2016). Policy experts' propensity to change their opinion along Delphi rounds. *Technological Forecasting and Social Change*, *109*, 61–68. https://doi.org/10.1016/j.techfore.2016.05.020

Meijering, J. V., & Tobi, H. (2018). The effects of feeding back experts' own initial ratings in Delphi studies: A randomized trial. *International Journal of Forecasting*, *34*(2), 216–224. https://doi.org/10.1016/j.ijforecast.2017.11.010

Meijering, J. V., & Tobi, H. (2016). The effect of controlled opinion feedback on Delphi features: Mixed messages from a real-world Delphi experiment. *Technological Forecasting and Social Change*, *103*, 166–173. https://doi.org/10.1016/j.techfore.2015.11.008

Mulgrave N. W. & Ducanis A. J. (1975) Propensity to change responses in a Delphi round as a function of dogmatism. In *The Delphi Method: Techniques and Applications* (Linstone H.A. & Turoff M. eds.), Addison-Wesley, London, pp. 288-290.

Murphy M. K., Black N. A., Lamping D. L., McKee C. M., Sanderson CF, et al. (1998) Consensus development methods, and their use in clinical guideline development. *Health Technol Assess* 2: i–iv, 1–88.

Nuño, L., Barrios, M., Rojo, E., Gómez-Benito, J., & Guilera, G. (2018). Validation of the ICF Core Sets for schizophrenia from the perspective of psychiatrists: An international Delphi study. *Journal of Psychiatric Research*, *103*(May), 134–141. https://doi.org/10.1016/j.jpsychires.2018.05.012

Nuño, L., Barrios, M., Moller, M. D., Calderón, C., Rojo, E., Gómez-Benito, J., & Guilera, G. (2019). An international survey of Psychiatric-Mental-Health Nurses on the content validity of the International Classification of Functioning, Disability and Health Core Sets for Schizophrenia. *International Journal of Mental Health Nursing*. https://doi.org/10.1111/inm.12586

Nuño, L., Guilera, G., Coenen, M., Rojo, E., Gómez-Benito, J., & Barrios, M. (2019). Functioning in schizophrenia from the perspective of psychologists: A worldwide study. *PLOS ONE*, *14*(6), e0217936. https://doi.org/10.1371/journal.pone.0217936

Prinsen, C. A. C., Vohra, S., Rose, M. R., King-Jones, S., Ishaque, S., Bhaloo, Z., … Terwee, C. B. (2014). Core Outcome Measures in Effectiveness Trials (COMET) initiative: protocol for an international Delphi study to achieve consensus on how to select outcome measurement instruments for outcomes included in a 'core outcome set.' *Trials*, *15*(1), 247. https://doi.org/10.1186/1745-6215-15-247

Rowe, G., & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting*, 12, 73–89.

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

Rowe, G., Wright, G., & McColl, A. (2005). Judgment change during Delphi-like procedures: The role of majority influence, expertise, and confidence. *Technological Forecasting and Social Change*, *72*(4), 377–399. https://doi.org/10.1016/j.techfore.2004.03.004

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

Sinha, I. P., Smyth, R. L., & Williamson, P. R. (2011). Using the Delphi technique to determine which outcomes to measure in clinical trials: Recommendations for the future based on a systematic review of existing studies. *PLoS Medicine*, *8*(1), e1000393. https://doi.org/10.1371/journal.pmed.1000393

Sobaih, A. E., Ritchie, C., & Jones, E. (2012). Consulting the oracle? Applications of modified Delphi technique to qualitative research in the hospitality industry. *International Journal of Contemporary Hospitality Management*, 24(6), 886–906.

van der Vaart, R., Witting, M., Riper, H., Kooistra, L., Bohlmeijer, E. T., & van Gemert-Pijnen, L. J. (2014). Blending online therapy into regular face-to-face therapy for depression: Content, ratio and preconditions according to patients and

therapists using a Delphi study. *BMC Psychiatry*, *14*(1), 355. https://doi.org/10.1186/s12888-014-0355-z

von der Gracht, H. A. (2012). Consensus measurement in Delphi studies. Review and implications for future quality assurance. *Technological Forecasting and Social Change*, *79*(8), 1525–1536. https://doi.org/10.1016/j.techfore.2012.04.013

Yaniv, I., Milyavsky, M., 2007. Using advice from multiple sources to revise and improve judgments. *Organ. Behav. Hum. Decis. Process*. 103 (1), 104–120.

Zawacki-Richter, O. (2009). Research areas in distance education: A Delphi study. *The International Review of Research in Open and Distributed Learning*, *10*(3). https://doi.org/10.19173/irrodl.v10i3.674

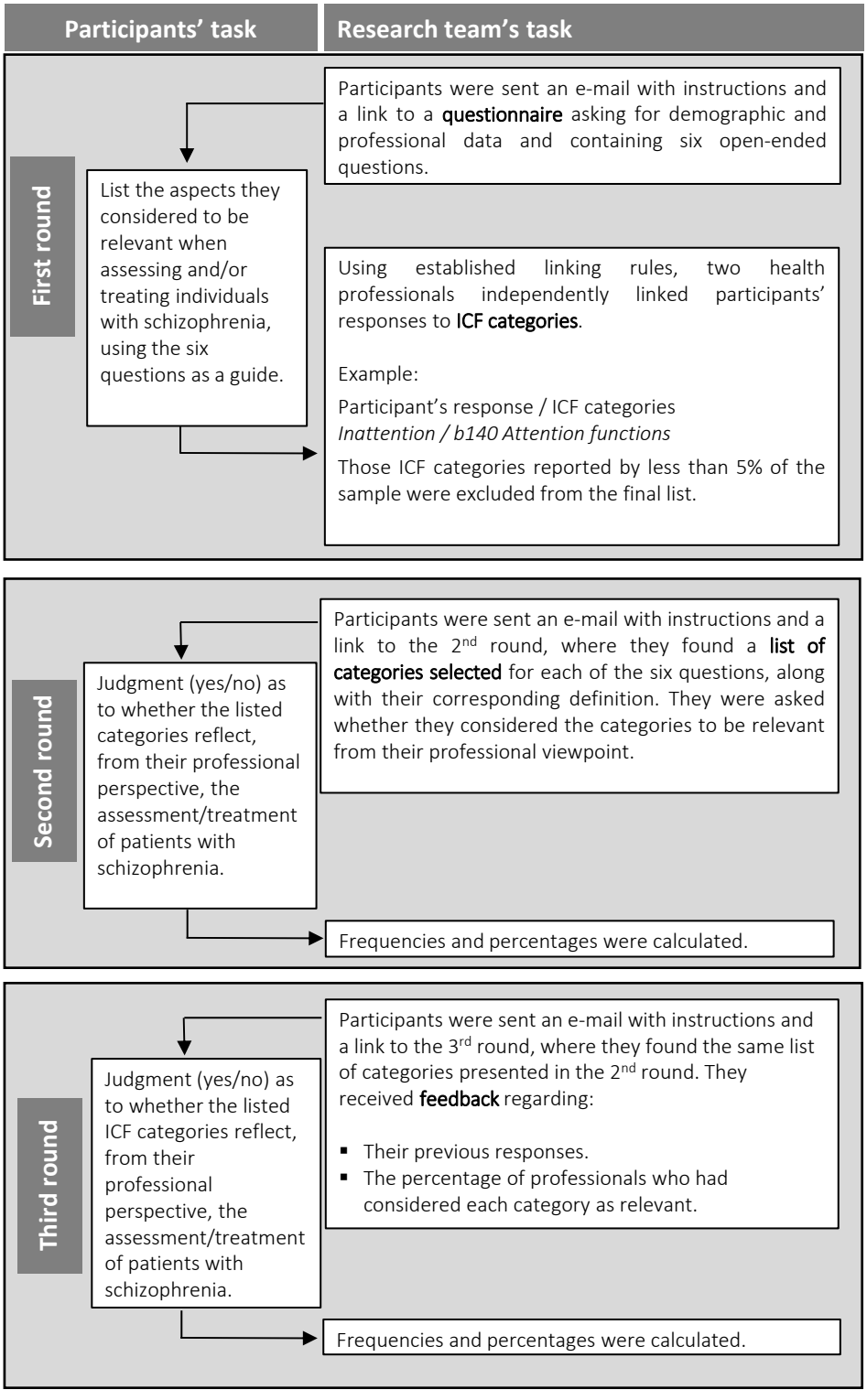
Participants' task
Research team's task
First round
Participants were sent an e-mail with instructions and a link to a questionnaire asking for demographic and professional data and containing six open-ended questions.
List the aspects they considered to be relevant when assessing and/or treating individuals with schizophrenia, using the six questions as a guide.
Using established linking rules, two health professionals independently linked participants' responses to ICF categories.
Example:
Participant's response / ICF categories
Inattention / b140 Attention functions
Those ICF categories reported by less than 5% of the sample were excluded from the final list.
Second round
Participants were sent an e-mail with instructions and a link to the 2nd round, where they found a list of categories selected for each of the six questions, along with their corresponding definition. They were asked whether they considered the categories to be relevant from their professional viewpoint.
Judgment (yes/no) as to whether the listed categories reflect, from their professional perspective, the assessment/treatment of patients with schizophrenia.
Frequencies and percentages were calculated.
Third round
Participants were sent an e-mail with instructions and a link to the 3rd round, where they found the same list of categories presented in the 2nd round. They received feedback regarding:
Their previous responses.
The percentage of professionals who had considered each category as relevant.
Judgment (yes/no) as to whether the listed ICF categories reflect, from their professional perspective, the assessment/treatment of patients with schizophrenia.
Frequencies and percentages were calculated.

Table 4. Sociodemographic and professional variables and their relevance in explaining the shift of opinion based on the 75% threshold of group agreement.

	≥75% threshold								**<75% threshold**							
	Non-change (congruent)		Non-change (incongruent)		Change (congruent)		Change (incongruent)		Non-change (congruent)		Non-change (incongruent)		Change (congruent)		Change (incongruent)	
Variable	**B**	**SE B**	**B**	**SE B**	**B**	**SE B**	**B**	**SE B**	**B**	**SE B**	**B**	**SE B**	**B**	**SE B**	**B**	**SE B**
Age	0.173	0.146	-0.058	0.087	-0.054	0.082	-0.061	0.082	0.079	0.214	-0.215	0.211	0.031	0.155	0.104	0.093
Professional experience	-0.139	0.150	-0.007	0.089	0.039	0.085	0.108	0.084	0.052	0.220	0.031	0.217	0.021	0.159	-0.104	0.096
Gender	1.453	1.551	-1.554	0.923	-1.414	0.873	1.524	0.868	-6.255	2.268	3.862	2.240	2.756	1.639	-0.363	0.987
Profession[a]																
Nurses	0.369	2.023	1.443	1.204	1.110	1.138	-2.926	1.131	-0.827	2.958	0.632	2.922	-2.267	2.137	2.462	1.287
Occup. therapists	-6.951	2.302	-2.382	1.370	-0.935	1.295	-3.668	1.288	3.669	3.367	-3,999	3.326	0.952	2.433	-0.623	1.465
Psychologists	-0.476	3.295	1.070	1.962	1.385	1.854	-0.982	1.843	-3.997	4.818	8.732	4.760	-3.609	3.482	-1.126	2.097
Social workers	0.610	2.760	2.135	1.643	0.329	1.553	-3.078	1.544	-9.163	4.036	15.622	3.986	-9.930	2.916	3.471	1.756
Region[b]																
Africa	1.716	3.071	-3.128	1.828	3.165	1.728	-1.744	1.718	0.490	4.491	2.258	4.436	-3.722	3.246	0.974	1.955
America	2.365	1.952	-2.403	1.162	-0.222	1.098	0.267	1.092	-0.493	2.855	1.034	2.820	1.345	2.063	-1.886	1.242
Eastern Mediterranean	5.381	3.065	-1.61	-4.738	0.896	1.729	-1.528	1.715	-2.404	4.482	5.338	4.427	-4.830	3.239	1.896	1.951
South-East Asia	5.434	2.131	-2.20	-5.324	0.909	1.199	-1.013	1.192	-7.636	3.116	6.871	3.079	-0.905	2.252	1.669	1.356
Western Pacific	2.327	2.071	0.16	-2.971	0.060	1.165	0.584	1.159	-1.631	3.029	2.921	2.992	0.683	2.189	-1.973	1.318
R^2	0.036		0.032		0.023		0.041		0.052		0.072		0.043		0.040	
p-value	0.102		0.005		0.457		0.046		0.008		< 0.001		0.034		0.058	

Note: [a] Psychiatrists was used as the reference category. [b] Europe was used as the reference category.