Guidelines for Designing and Conducting Delphi Consensus Studies: An Expert Consensus Delphi Study

Erik Hohmann, M.B.B.S., F.R.C.S., F.R.C.S. (Tr. & Orth.), M.D., Ph.D., Philippe Beaufils, M.D., Daniel Beiderbeck, Ph.D., Jorge Chahla, M.D., Ph.D., Andrew Geeslin, M.D., Samer Hasan, M.D., Ph.D., Susan Humphrey-Murto, M.D., F.R.C.P.C., M.Ed., Eoghan Hurley, M.B.B.Ch., Ph.D., M.Sc., Robert F. LaPrade, M.D., Ph.D., Frank Martetschläger, M.D., Bogdan Matache, M.D., F.R.C.S.C., Dip Sports Med, Gilbert Moatshe, M.D., Ph.D., Juan Carlos Monllau, M.D., Ph.D., Iain Murray, M.B.ChB., F.R.C.S. Ed., F.R.C.S. Ed. Orth., Marlen Niederberger, M.A., Ph.D., Urs Rüetschi, M.D., Zhida Shang, M.D., Stephen Weber, M.D., Ivan Wong, M.D., F.R.C.S., Dip Sports Med, M.ArM., F.A.A.N.A., and Nicholas P. J. Perry, M.D.

Purpose: To conduct a Delphi project to develop guidelines for the design and execution of Delphi studies within medical and surgical specialties. **Methods:** Open-ended questions in round 1 and open-ended and semi-open questions in round 2 were answered. The results of the first 2 rounds were used to develop a Likert-style questionnaire for round 3. The level of agreement and consensus was defined as 80%. Consensus was further categorized into specific percentage ranges for clarity: 100% unanimous consensus, 90% to 99% very strong consensus, and 80% to 89% consensus. **Results:** Consensus was achieved for 35 of 63 items (56%). Unanimous agreement was reached for 4 items (6.3%), while very strong consensus was established for 12 items (19%). Consensus was reached for an additional 19 items (30.1%), and the panel remained undecided on 7 items (11.1%). **Conclusions:** Unanimous agreement was reached for iteration, the ability to establish treatment guidelines, a proven track record of panel members, and the requirement for at least 1 steering committee member to be a Delphi expert. Very strong consensus was reached on several key requirements: a clear definition of consensus, controlled feedback between rounds, precise definitions of expert and expertise, and the need for panel members to show experience through publications and clinical practice. Criteria for panel selection should ensure diversity and specialization, with steering committee members being content experts and a minimum of 20 to 30 panel members for broader topics. Regional experts should provide consensus on specific topics only. The steering committee should develop questions, with open-ended questions in round 1 and both types in round 2. Limiting the process to 3 rounds is advisable, aiming for at least 80% consensus in the final round. **Level of Evidence:** Level V, expert opinion.

From the Medical School, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa (E.H.); Department of Orthopaedic Surgery and Sports Medicine, Burjeel Hospital for Advanced Surgery, Dubai, United Arab Emirates (E.H.); ESSKA, Centre Medical/Fondation Norbert Metz, Luxembourg City, Luxembourg (P.B.); Center for Sports and Management (CSM), WHU—Otto Beisheim School of Management, Düsseldorf, Germany (D.B.); Midwest Orthopaedics at Rush, Chicago, Illinois, U.S.A. (J.C.); Larner College of Medicine, University of Vermont, Burlington, Vermont, U.S.A. (A.G.); Mercy Health-Cincinnati Sports Medicine and Orthopaedic Center, Cincinnati, Ohio, U.S.A. (S.H.); Department of Medicine, University of Ottawa, Ottawa, Canada (S.H-M.); Department of Orthopaedic Surgery, Duke University School of Medicine, Durham, North Carolina, U.S.A. (E.H.); Twin Cities Orthopedics, Edina, Minnesota, U.S.A. (R.L.); German Center for Shoulder Surgery, ATOS Clinic, Munich, Germany (F.M.); Department of Orthopaedic Surgery, Ottawa Hospital, University of Ottawa, Ottawa, Canada (B.M.); Division of Orthopaedic Surgery, Oslo University Hospital, University of Oslo, Oslo, Norway (G.M.); Hospital del Mar, Hospital Universitari Dexeus (ICATME), Universitat Autònoma de Barcelona (UAB), Barcelona, Spain (J.C.M.); Edinburg Orthopaedics, Royal Infirmary of Edinburgh, University of Edinburgh, Edinburgh, U.K. (I.M.); Department of Research Methods in Health and Prevention, Institute for Health Sciences, University of Education, Schwäbisch Gmünd, Germany (M.N.); AO Foundation, AO Education Institute, Davos, Switzerland (U.R.); Faculty of Medicine and Health Sciences, McGill University, Montreal, QC, Canada (Z.S.); Department of Orthopedics, The Johns Hopkins School of Medicine, Baltimore, Maryland, U.S.A. (S.W.); Division of Orthopaedic Surgery, Department of Surgery, Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada (I.W.); and Naval Medical Center San Diego, San Diego, California, U.S.A. (N.P.J.P.).

Received November 30, 2024; accepted March 7, 2025.

Address correspondence to Erik Hohmann, M.B.B.S., F.R.C.S., F.R.C.S. (Tr. & Orth.), M.D., Ph.D., Burjeel Hospital for Advanced Surgery, Sheikh Zayed Road, 46WF+PC Dubai, United Arab Emirates. E-mail: drerik@burjeelspecialty.com

© 2025 by the Arthroscopy Association of North America. Published by Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

0749-8063/241975/$36.00
https://doi.org/10.1016/j.arthro.2025.03.038

The theory of evidence-based medicine (EBM) is defined as the "conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients."[1] It is now a standard component of both undergraduate and graduate medical education curricula, and the principles of EBM have been further adopted across the medical field, leading to the broader implementation of evidence-based practice.[2,3] One of the perceived risks of EBM is that it may be viewed as an unquestionable truth, potentially undermining the individual physician-patient relationship.[4]

While EBM is currently considered the gold standard, its limitations, including research bias, lack of evidence, conflicts of interest, industry-sponsored trials, reliance on surrogate markers, poor data standardization, and selection bias, must be acknowledged and may result in applying unproven treatments.[5-7]

A potential future alternative to EBM could be the integration of artificial intelligence and machine learning technologies. These tools could assist clinicians in their decision-making processes, generating personalized treatment plans and tailoring medical interventions to individual patients' needs.[8-10] However, until artificial intelligence and machine learning tools reach a sufficient level of maturity and reliability, consensus-based evidence techniques remain a viable alternative to EBM. These methods, which rely on expert opinions and collective clinical experience, offer a practical approach for guiding medical decision-making in areas where high-quality data may be limited and facilitate a scientific approach to generate a reliable consensus opinion of a group of experts.[11-13]

Delphi techniques are generally more flexible and adaptable to emerging methodologies, facilitate consensus among experts, and are particularly valuable when high-quality evidence is lacking.[14] Unlike EBM or conventional expert panels, the anonymity in the Delphi process minimizes bias, reducing the influence of dominant individuals and promoting independent, well-considered responses.[11-14]

Although various Delphi techniques have been described by the RAND Corporation, the most widely utilized is the modified Delphi technique, which was developed to assess the appropriateness of health care interventions.[15] While the modified Delphi method is now commonly employed across multiple medical specialties, the lack of a clear definition and standardized methodological guidelines for conducting rigorous Delphi studies raises concerns about how to properly implement the method.[15]

The purpose of this study was to conduct a Delphi project to develop guidelines for the design and execution of Delphi studies within medical and surgical specialties. It was hypothesized that consensus would be achieved for most items, providing clear guidance for medical professionals.

Methods

The Delphi panel technique was applied as previously outlined.[11,12] The steering committee independently developed the first round of questions in accordance with previous recommendations.[11,12] The questions for round 1 were designed to be simple and open-ended, allowing the panel to identify key issues and respond based on their experience and interpretation of the current evidence and practices. On completion of round 1, a written summary was provided to the panel. Using answers from round 1, round 2 open questions were generated and e-mailed to the panel. For round 3, a Likert-style questionnaire was developed and e-mailed to the panel. Although there are no agreed-upon recommendations for consensus, several authors have suggested that consensus should be defined by an agreement level of 80%.[16-21]

Composition of Steering Committee

The steering committee for this Delphi project was composed of 2 orthopaedic surgeons. The senior orthopaedic surgeon (E.H.) is an experienced clinical researcher with expertise in Delphi consensus studies. The junior orthopaedic surgeon (N.P.J.P.) is a fellowship-trained surgeon with experience in study design, research methodology, and critical appraisal of the literature and manuscripts.

Panel Selection

To select suitable panel members, the current literature was searched on PubMed using the search terms "Delphi," "modified Delphi," and "Delphi guidelines." Senior and primary authors of relevant studies were contacted via e-mail and invited to participate. They were also asked to recommend colleagues with extensive experience in Delphi methodology, and these recommended experts were subsequently approached via e-mail.

Based on previous recommendations, the panel size should consist of a minimum of 15 to 18 members.[11,12] Following these guidelines and utilizing the search strategy outlined above to identify experts, the steering committee identified 25 potential candidates, who were subsequently contacted via e-mail or LinkedIn. Two potential panel members were not able to be contacted. One declined due to retirement, 1 declined due to feeling inadequately qualified, and 3 declined due to other professional obligations or lack of time. Therefore, the final panel was composed of 18 experts. Based on the recommendations, the number of panel members was deemed sufficient, and the steering committee ceased further solicitation.[11,12]

Question Development

The clinical guideline published in *Arthroscopy* served as the basis for question development.[12] Furthermore, the most recent literature on conducting Delphi studies in medicine was utilized.[15,19,22] The existing literature and established guidelines were reviewed to identify the principal steps in the design of the Delphi technique. Key themes were then determined, forming the basis for the development of open-ended questions for round 1. The responses from round 1 were analyzed to inform the formulation of questions for round 2. Any disagreements among the 2 steering committee members were resolved through discussion.

Rounds 1 and 2

In round 1, the panel was asked to respond to 21 open-ended questions covering various aspects of a Delphi project, including the philosophy of Delphi, panel formation, question development, structure and flow, final round design and objectives, consensus achievement, and statistical analysis (Table 1). These questions were sent individually via e-mail to each panel member by a steering committee member, ensuring the anonymity of the panel. The panel was requested to respond to the questions in a detailed narrative format based on their experience and published Delphi guidelines. After summarizing the results of round 1, the agreements and controversies from round 1 were highlighted and e-mailed to the panel (Appendix 1, available at www.arthroscopyjournal.org).

Based on the panel's responses, 2 semi-closed questions and 3 additional open questions were formulated (Table 2). The semi-closed questions asked panel members to rank the importance of essential Delphi features and to assess the limitations of the Delphi method. For the question regarding Delphi limitations, the ranking system was defined as 1 (very important) to 9 (least important). The questions were again delivered to the panel via e-mail. After summarizing the results of round 2, the agreements and controversies from round 2 were highlighted and e-mailed to the panel (Appendix 2, available at www.arthroscopyjournal.org).

Round 3

Based on the responses from round 1 and round 2, 55 Likert-style items were developed for round 3 (Table 3). The items were ordered in a random fashion and e-mailed to the panel for final voting. Consistent with a prior Delphi study,[23] panel members were asked to respond to each item using a scale that included "strongly agree," "agree," "neither agree nor disagree," "disagree," or "strongly disagree."

Consensus

Consensus was defined as achieving a minimum level of 80% agreement. Consensus was further categorized into specific percentage ranges for clarity: 100% indicates "unanimous consensus," 90% to 99% represents "very strong consensus," and 80% to 89% signifies "consensus."[24] If the panel reached a consensus to disagree on a particular item, the results were reported accordingly. To develop recommendations for the design of a Delphi study, items achieving consensus thresholds were categorized based on the level of agreement. Items with a consensus of 90% or higher were deemed critical, items with a consensus between 80% and 89% were considered important, and items that did not meet the 80% consensus threshold were regarded as less important.

Statistical Analysis

The results of round 3 were presented as calculated percentiles.

Results

Panel Composition

All 18 panel members participated in and completed all 3 rounds. There was no dropout. All 18 were well versed in the Delphi method, having participated in multiple Delphi projects and published extensively in this field. Fourteen of the panel members were practicing academic orthopaedic surgeons, with 6 based in Europe and 8 in North America. One panel member is the former director of education at the AO Foundation and currently serves as the head of Educational Special Projects at the AO Foundation. Another panel member is a Schwarzman Scholar and the author of "The Use of Delphi in Health Sciences Research." One panel member, a graduate of the Otto Beisheim School of Management, has published Delphi studies and authored a guideline titled "Preparing, Conducting, and Analyzing Delphi Surveys: Cross-Disciplinary Practices, New Directions, and Advancements." Additionally, like the other panel members, 1 panelist is a faculty member at the University of Ottawa's medical school, bringing extensive expertise in Delphi methodology and having authored several Delphi guidelines and projects.

Round 1

The responses for the first round are summarized in Appendix 1 (available at www.arthroscopyjournal.org). Briefly, the main responses were that Delphi features include blinding, an independent steering panel not involved in voting, a heterogeneous panel of leading experts with a track record both academically and clinically, iteration, controlled feedback between rounds, and a clear definition of consensus as the nonnegotiable features of Delphi. Delphi methodology is suitable when strong evidence is lacking or controversy exists. Panels should be diverse, except for highly specialized topics requiring specific expertise. Members

Table 1. Delphi Round 1 Questions

	Philosophy of Delphi
[1]	What are the essential features of the Delphi method? What are the nonnegotiable features of the Delphi method?
[2]	When is the Delphi method the most appropriate method and, conversely, when is it not a suitable tool?
[3]	What are the limitations of Delphi? How would you mitigate or overcome these limitations?
[4]	Can you comment on the CREDES criteria and how would you perceive their relevance and applicability?
	Panel Creation
[5]	Which criteria would you apply when selecting a Delphi panel?
[6]	What qualifications and experience would you deem essential for individuals to have in order to be eligible to serve on a Delphi panel? We are interested in understanding what you believe are the minimum necessary qualifications versus the desired or ideal qualifications.
[7]	When selecting a panel member, do you believe that choosing a friend or colleague, who is perceived as an expert, may introduce bias or interfere with the Delphi process? If so, how would you mitigate this potential influence?
[8]	When creating a Delphi panel, should you prioritize a panel with a diversity of experiences and backgrounds or a more focused, specialist panel? What factors ensure the effectiveness of Delphi within this context?
	Question Creation
[9]	For first-round, open-ended questions, what are the characteristics or qualities of a well-written, effective question?
[10.1]	Do you believe it is more effective for the research team/steering committee to formulate questions for round 1 or should the panel be given the opportunity to submit topics of interest?
[10.2]	Can you envision where utilizing both techniques would be beneficial?
	Structure and Flow of Delphi
[11]	How would you determine the optimal number of rounds? What factors dictate the minimum number of rounds to achieve consensus and substantive valid outcomes?
[12]	For the second round, should open- or closed-ended questions be employed? When should the panel transition from open-ended to closed-ended questions?
[13]	How should feedback be organized between rounds? Does this vary between rounds with open-ended questions and closed-ended questions?
[14]	Would you consider employing Likert-style questions in early rounds, followed by eliminating items that fail to achieve consensus as a viable strategy, or do you perceive it as potentially deviating from Delphi principles?
[15]	What is the optimal maximum time frame between rounds and for the overall completion of a Delphi project?
[16]	What constitutes an acceptable attrition rate of panel members?
	Final Round
[17]	What are your thoughts on the optimal structures and objectives for the final round?
	Achieving Consensus
[18]	What percentage of agreement between panel members do you consider to determine and ensure robust and meaningful consensus?
[19]	Is there a scenario where achieving 100% agreement among panel members is deemed necessary?
[20]	Which aspect is more important: consensus or stability of responses?
	Consensus Statement
[21]	What considerations should guide the selection of statistical methods? Are there any standard approaches that should be used?

must have relevant expertise and a proven track record. Recommendations from friends or colleagues are acceptable if they do not compromise selection criteria or anonymity. Round 1 should use neutral, open-ended questions developed by the steering committee. An attrition rate of up to 10% is acceptable, with the final round reserved for voting only.

Round 2

The importance of nonnegotiable features was ranked as follows: iterative process, anonymity, heterogeneous panel of experts, clear definition of consensus, controlled feedback between rounds, and an independent steering committee. The limitations of the Delphi method were ranked in terms of importance as follows: potential bias in panel selection, inherent bias within the Delphi process, lack of standardization in question development, low level of evidence, potential omission of experts, lack of panel objectivity, consensus based on the lowest common denominator, geographic bias, and limited generalizability. The steering committee should consist of 3 to 5 independent members, including at least 1 Delphi expert. While the ideal panel size was not fixed, most recommended 10 to 15 members for specialized topics and more for broader ones. Friend referrals were debated; some opposed them unless anonymity was guaranteed, while others supported them under those conditions. Nonrespondents should be reminded 2 to 3 times before exclusion, with attrition rates up to 20% considered acceptable. To reduce responder fatigue, limit questions, minimize open-ended queries, avoid redundancy, prioritize multiple-choice questions, provide updates, send thank-you notes, inform panelists of workload upfront, allow feedback, share results, and

Table 2. Round 2 Delphi Questions

[1]	Please rate the importance of nonnegotiable features, with 1 being the most important and 6 being the least important. Additionally, feel free to include any other items you consider mandatory and nonnegotiable.
	Anonymity and blinding
	An independent steering panel not involved in voting
	An iterative process
	A heterogeneous panel of experts with both clinical and academic expertise
	Controlled feedback between rounds
	A clear definition of consensus
[2]	Please rate the limitations of the Delphi method according to their importance in introducing bias and error, thereby reducing validity, with 1 being the most important and 9 being the least important. Additionally, feel free to include any other items you consider mandatory and nonnegotiable.
	Level of Evidence V (LOE V)
	Inherent bias
	Lack of standardization for generating questions
	Agreement based on the lowest common denominator
	Dependence on the panel and potential selection bias
	Lack of panel objectivity
	Geographic bias
	Limited generalizability
	Potential omission of relevant experts, which may influence the direction of the results
[3]	Could you please comment on the composition of the steering panel?
[a]	Is there a minimum number of panel members required, and should they all be experts in the field and the Delphi method?
[b]	Is it acceptable to have a steering panel member who is not an expert in the field but is familiar with the Delphi method?
[c]	Or would you consider it mandatory for at least 1 member of the steering committee to be a Delphi expert?
[4]	The minimum and ideal size of the panel remains unclear. Could you please specify the minimum and ideal number of panel members?
[5]	The use of friend or colleague referrals requires further exploration. Could you please specify the conditions under which you consider such referrals appropriate and when they are not? For this item, please use bullet points and a yes/no structure.
[6]	What are the optimal methods for handling nonresponders after sending out a survey? How should researchers remind panelists to participate, and how many reminders should be sent? If a panelist is dropped from 1 round, should they be allowed to participate in the next round or dismissed from the study?
[7]	Does exceeding the acceptable attrition rate of panelists in a Delphi study invalidate its findings? Can the study still be considered publishable despite this? How should researchers proceed in such a situation?
[8]	How can researchers designing a Delphi project best prevent burnout or responder fatigue among participants? What methods can be used to incentivize or encourage panelists before or during the Delphi Project?

mention coauthorship opportunities (Appendix 2, available at www.arthroscopyjournal.org).

Round 3

Of the 63 items and subitems evaluated, 4 achieved unanimous consensus, 12 attained very strong consensus, and 19 reached consensus. Overall, 56% of the items met the consensus criteria. Additionally, 6 items approached consensus, while the panel was unable to reach any consensus on 7 items. Unanimous consensus was reached on the following features: iterative process essential for Delphi, its applicability in establishing treatment guidelines, the requirement for panel members to have a proven track record in the relevant topic, and the necessity for at least 1 member of the steering committee to be a Delphi expert. For further details, please refer to Table 3.

Discussion

This Delphi expert consensus project showed that unanimous agreement was reached on several key features that should be regarded as critical elements of any Delphi process: the iterative nature of the Delphi method, its ability to establish treatment guidelines, the requirement for panel members to have a proven track record in the relevant area, and the necessity for at least 1 member of the steering committee to be a Delphi expert. Other features that achieved very strong consensus and should be regarded as critical elements of any Delphi process include the requirement for a clear definition of consensus, controlled feedback between rounds, a precise definition of an expert and expertise, and the critical necessity for panel members to show content experience through publications, presentations, and clinical practice. Additionally, predefined criteria for panel selection, a diverse and specialized panel, the requirement that steering committee members be content experts, and a minimum of 20 to 30 panel members for broader topics were highlighted. It was also noted that regional experts should provide consensus only for specific topics within their geographic area. Very strong consensus was reached regarding several procedural aspects: the responsibility for question development lies with the steering

Table 3. Likert Items, Results, and Consensus for Round 3

	Strongly Agree	Agree	Neither Agree Nor Disagree	Disagree	Strongly Disagree	Consensus
[1] Essential and nonnegotiable features of any Delphi study include the anonymity and blinding of panel members.	**12**	**2**	**1**	**0**	**2**	**83**
[2] Other essential features of a Delphi study include:						
[2.1] an independent steering panel not involved in voting	9	3	3	3	0	66
[2.2] an iterative process	**17**	**1**	**0**	**0**	**0**	**100**
[2.3] a heterogeneous panel of experts with both clinical and academic expertise	11	1	5	1	0	66
[2.4] controlled feedback between rounds	**13**	**4**	**1**	**0**	**0**	**94**
[2.5] a clear definition of "consensus"	**15**	**2**	**0**	**1**	**0**	**94**
[3] Delphi studies should be employed in situations where clear evidence is lacking or conflicting evidence exists on a particular topic.	**11**	**6**	**1**	**0**	**0**	**94**
[4] The Delphi method is not suitable for opinion polling.	4	3	9	2	0	38
[5] The Delphi method can be useful for establishing treatment guidelines.	**12**	**6**	**0**	**0**	**0**	**100**
[6] The CREDES criteria provide a useful guideline and should be considered when developing a Delphi study.	3	10	4	1	0	72
[7] Experts and expertise must be clearly defined	**14**	**3**	**1**	**0**	**0**	**94**
[8] Content expertise is critical and should be demonstrated through publications, presentations, and clinical practice.	**12**	**5**	**0**	**1**	**0**	**94**
[9] Panel members must be unbiased.	8	5	3	2	0	72
[10] Panel heterogeneity is crucial.	**11**	**4**	**3**	**0**	**0**	**83**
[11] Individuals with known strong and/or controversial opinions should be excluded from participation.	0	1	5	10	2	67
[12] Geographic variability is essential for enhancing the generalizability of the consensus.	9	4	4	1	0	72
[13] Regional experts can be included, but it is important to note that the results may primarily apply to the specific geographic area, and should be discussed in the manuscript.	**10**	**7**	**1**	**0**	**0**	**94**
[14] Panel members should have no financial interests or other conflicts of interest. This should be established upon acceptance of invitations. If a proposed panel member declares a conflict of interest, they should be excluded from participation.	1	4	9	4	1	Undecided

(continued)

Table 3. Continued

	Strongly Agree	Agree	Neither Agree Nor Disagree	Disagree	Strongly Disagree	Consensus
Panel members should meet the following minimum qualifications and experience to be eligible to serve on a Delphi panel:						
[15.1] A proven track record on the subject under study is essential.	**13**	**5**	**0**	**0**	**0**	**100**
[15.2] Experience with the Delphi method is beneficial but not mandatory.	**7**	**8**	**0**	**3**	**0**	**83**
[15.3] Predefined criteria for panel selection are essential and should be clearly documented in the manuscript or publication.	**13**	**4**	**1**	**0**	**0**	**94**
[16.1] Panel members can recommend friends or colleagues, provided that the core principles of the Delphi method are not compromised.	6	8	0	2	1	78
[16.2] Referrals from friends or colleagues should be avoided to minimize potential bias.	4	1	11	2	0	Undecided
[17.1] At least 1 member of the steering committee must be an expert in Delphi methodology. If this is not possible, the committee should invite a Delphi expert as an advisor.	**13**	**5**	**0**	**0**	**0**	**100**
[17.2] Steering committee members must be content experts.	**13**	**3**	**1**	**1**	**0**	**89**
[18] Both a diverse and specialized panels are suitable, with selection criteria contingent upon the subject under study.	**8**	**8**	**2**	**0**	**0**	**89**
[19] In the first round, open-ended questions using clear and unambiguous language should be used to facilitate genuine and diverse opinions.	**11**	**6**	**1**	**0**	**0**	**94**
[20] Panel members should be provided the opportunity to suggest additional questions while responding to round 1 queries if any pertinent topics are omitted. These proposed questions should then undergo assessment by the steering committee for potential inclusion in round 2.	**9**	**6**	**0**	**1**	**0**	**82**
[21] Likert-style or binary yes/no questions should not be included in round 1.	8	5	2	1	2	72
[22] Question development is the responsibility of the steering committee and should be informed by a comprehensive review of the existing literature.	**12**	**4**	**2**	**0**	**0**	**89**

(continued)

Table 3. Continued

	Strongly Agree	Agree	Neither Agree Nor Disagree	Disagree	Strongly Disagree	Consensus
[23] The steering committee should consider conducting a scoping review of the literature, which should be either published or shared with the panel members.	**7**	**8**	**2**	**1**	**0**	**83**
[24] There is no universally accepted optimal number of rounds, but in general, limiting the process to 3 rounds should be sufficient.	**9**	**7**	**2**	**0**	**0**	**89**
[25] The second round can include a combination of both open-ended and closed-ended questions.	**11**	**5**	**2**	**0**	**0**	**89**
[26] Closed-ended questions are the preferred format for round 2.	4	8	4	2	0	67
[27] If topics were not addressed or were omitted in round 1, they can be addressed in round 2.	**7**	**9**	**2**	**0**	**0**	**89**
[28] Controlled feedback and summarizing the results of each round are essential features of the Delphi method.	**15**	**2**	**0**	**1**	**0**	**94**
[29] Feedback can be organized either as a written summary or a graphical representation.	**11**	**6**	**1**	**0**	**0**	**94**
[30] Likert-style questions in early rounds are limiting and should be avoided; instead, the broader topic should be explored comprehensively first.	5	7	3	1	1	67
[31] The timeline for completing a Delphi project should not exceed 6 months in most instances.	**10**	**5**	**0**	**1**	**1**	**83**
[32] For most Delphi projects, the time between rounds should be 2 weeks, and the expert panel's response time should not exceed 4 weeks.	8	6	4	0	0	78
[33] The attrition rate from round 1 to the final round should not exceed 10%.	4	3	9	2	0	Undecided
[34] During the recruitment phase for panel members, a decline rate of up to 20% is considered acceptable. Higher rates could introduce bias and should be carefully considered, disclosed, and discussed.	5	9	3	1	0	78
[35] The final round is designated for voting on consensus, where statements should be rated using a 5-point Likert scale.	**11**	**4**	**3**	**0**	**0**	**83**

(continued)

Table 3. Continued

	Strongly Agree	Agree	Neither Agree Nor Disagree	Disagree	Strongly Disagree	Consensus
[36] For more concise measures, the Delphi manager can also be utilized, employing a 9-point Likert scale ranging from "not important" to "important but not critical" to "critical."	4	9	5	0	0	72
[37] In the final round, panel members who are outside of consensus can be contacted to either reconsider their stance or provide justification for their response.	2	6	4	5	2	Undecided
[38] The final round is intended for voting, and additional input should generally be avoided.	8	5	3	0	2	72
[39] The primary objective of the final round is to reach consensus through voting.	**11**	**5**	**1**	**0**	**1**	**89**
[40] Disagreement is as crucial as agreement and should be considered when assessing consensus.	**13**	**4**	**1**	**0**	**0**	**94**
[41] When reaching consensus on a specific item, disagreement should be limited to less than 10%.	3	3	9	3	0	Undecided
[42] If disagreement exceeds 10% despite apparent consensus, it should be concluded that consensus has not been reached unless the consensus rate is 90% or higher.	3	3	8	4	0	Undecided
[43] Neutral votes should not be considered as disagreement.	2	8	5	2	1	56
[44] Achieving 100% indicates "unanimous consensus," 90% to 99% indicates "very strong consensus," 80% to 89% indicates "consensus," and 70% to 79% indicates "weak consensus" should be used as a consistent naming convention across Delphi studies.	**7**	**8**	**1**	**2**	**0**	**83**
[45] 100% agreement is deemed necessary only in scenarios involving safety or life-saving measures.	10	4	3	1	0	78
[46] If panelists propose a consensus that deviates significantly from the current standard of care, achieving 100% consensus should be mandatory.	5	5	5	3	0	56
[47] Requiring 100% consensus in most scenarios is unrealistic given the diversity of opinions; therefore, a minimum of 80% consensus should be sought.	**9**	**7**	**1**	**1**	**0**	**89**
[48] In practice, consensus is the key measure of importance.	7	4	5	1	1	61

(continued)

Table 3. Continued

	Strongly Agree	Agree	Neither Agree Nor Disagree	Disagree	Strongly Disagree	Consensus
[49] If more than 2 rounds using a Likert scale are conducted, the stability of responses should be checked and reported.	**7**	**10**	**1**	**0**	**0**	**94**
[50] For most Delphi projects, descriptive statistics are sufficient.	9	5	3	1	0	78
[51] For a more comprehensive and detailed analysis, interquartile ranges can be used to assess spread, and median absolute deviation can be included for interpreting the final round.	**7**	**8**	**3**	**0**	**0**	**83**
[52] In general, the Delphi panel should consist of at least 10 members, with a preferred range of 15 to 20 participants.	5	8	3	2	0	72
[53] For broader topics, the panel should include at least 20-30 members.	**9**	**7**	**1**	**2**	**0**	**89**
[54] An attrition rate of <10% is ideal for a study to be considered valid, provided the limitations from attrition are discussed in detail.	**10**	**5**	**2**	**1**	**0**	**83**
[55] An attrition rate of up to 20% is also acceptable, but the authors should thoroughly discuss and address the potential limitations and bias in detail.	4	10	1	3	0	78

Consensus was reached on these items were indicated in bold.

committee, with open-ended questions in round 1 and the option for both open and closed questions in round 2. The committee should also address topics omitted in round 1 in round 2. Controlled feedback between rounds should be provided in either written summaries or graphical formats. Additionally, if more than 2 rounds of Likert-style questions are used, it is essential to check for stability of responses. While there is no universally accepted number of rounds, limiting the process to 3 rounds is generally advisable. The panel also agreed that the Delphi method should be employed in situations where clear evidence is lacking or conflicting. Finally, during the final round of a Delphi study, the primary objective should be to reach consensus through voting, recognizing that while 100% consensus is unlikely, a minimum threshold of 80% is sought, with disagreement being considered a crucial component of the consensus process.

The previously published research pearls and editorial have established fundamental steps for conducting a Delphi consensus project.[11,12] However, these previous publications[11,12] primarily represent the informal opinions of members of the journal's editorial board. This study expands on those insights by employing a structured, scientific approach to consensus, incorporating new recommendations, and refining critical methodological steps. In addition, it identifies less critical elements that were previously thought to be important when designing a Delphi project.

Based on the final voting results in round 3, the panel recommends strongly considering the following steps when designing and executing a Delphi Expert Consensus study (Table 4). These recommendations were derived based on the level of consensus achieved. Critical steps were defined as those items that received unanimous or very strong consensus, while important steps were those that attained general consensus. Items that did not reach consensus were categorized as less important. Based on the consensus reached in round 3 and the classification of elements as critical, important, or less important, the steering committee refined and updated the methodological steps outlined in the research pearls (Table 5). These revised recommendations not only serve as a guideline for authors designing a Delphi study but also provide a framework for reviewing such manuscripts and assessing the quality of their methodology and design.

The critical and important key steps include an iterative process and a clear definition of consensus. Anonymity and blinding are also important, while the role

Table 4. Recommendations for the Design of a Delphi Study

Philosophy of Delphi		
Critical	Important	Less Important
Iterative process Clear definition of consensus Suitable where evidence lacking or conflicting Useful to establish guidelines	Anonymity and blinding	Independent steering panel
Steering Panel		
Critical	Important	Less Important
3-5 members One member must be Delphi expert Must be content experts		
Panel Creation		
Critical	Important	Less Important
Experts/expertise clearly defined Content experience critical Regional experts for regional consensus At least 10 members; preferably 15-20 Broader topics at least 20-30 members	Panel heterogeneity Experience with Delphi beneficial Diverse and specialized panel suitable Ideal attrition rate $\leq$10% Attrition rate up to 20% acceptable	Unbiased panel members Panel geographic variability COI or financial interest of panel Referral of friends or colleagues During recruitment attrition of up to 20% acceptable
Question Creation		
Critical	Important	Less Important
First round open questions	Panel members' input if topics omitted in round 1 Question development by steering committee Consider scoping review by the steering committee to be published or handed to panel	No Likert-style questions in round 1
Structure and Flow of Delphi		
Critical	Important	Less Important
Controlled feedback between rounds For more than 2 rounds with Likert scales, check stability of responses	No universal number of rounds Limit to 3 rounds is sufficient Second round open and closed questions Topics not addressed in round 1 can be addressed in round 2 Timeline for completion not to exceed 6 months Time between rounds not to exceed 2 weeks Time to respond for panel not to exceed 4 weeks Limit reminders to 2-3; if no response, exclude from future rounds	Closed questions for round 2
Final Round		
Critical	Important	Less Important
	Designated for voting 5-point Likert scale: strongly agree, agree, neither agree nor disagree, disagree, strongly disagree recommended Goal: reach consensus through voting	More concise 9-point Likert scale: not important to critical Panel members outside consensus should be contacted

(continued)

Table 4. Continued

Philosophy of Delphi		
Critical	Important	Less Important
Achieving Consensus		
Critical	Important	Less Important
Disagreement as crucial	100% unanimous, 90%-99% very strong, 80%-89% consensus, 70%-79% weak consensus	Disagreement <10%
	100% consensus unrealistic, minimum 80% recommended	100% consensus critical in life-saving or safety measures
Consensus Statement		
Critical	Important	Less Important
	For a comprehensive analysis: include median and interquartile ranges	Descriptive statistics

of an independent steering committee appears to be less critical. For the steering committee, critical features include comprising 3 to 5 members who are all content experts, with the requirement that at least 1 member must be a Delphi expert. For panel creation, critical steps include clearly defining the experts and their areas of expertise, as well as emphasizing the importance of content experience among panel members. The panel should consist of at least 10 members, with a preferred size of 15 to 20. For broader topics, a minimum panel size of 15 to 20 is recommended. Regional experts should provide consensus only for topics relevant to their specific geographic area. Important features for panel creation include panel heterogeneity and an ideal attrition rate of less than 10%, although rates of up to 20% are acceptable. The suitability of both diverse and specialized panels is essential, and prior experience with the Delphi method would be advantageous. Less critical features include geographic variability among panel members, the inclusion of unbiased participants, and the management of potential conflicts of interest, such as financial interests and the referral of friends or colleagues for panel inclusion.

For question creation, critical features include that the first round should consist solely of open-ended questions. Important considerations involve allowing panel input on any topics omitted in round 1, developing questions by the steering committee, and conducting a scoping review that should be shared with the panel or published. Less critical is the recommendation that Likert-style questions should not be included in round 1.

For the structure and flow of the Delphi process, critical steps include providing controlled feedback between rounds and checking the stability of responses if more than 2 rounds are conducted with Likert-style items. Important features include ensuring that the overall timeline does not exceed 6 months, with intervals between rounds not exceeding 2 weeks and response time for panel members limited to 4 weeks. Reminders should be limited to 2 to 3, after which panel members may be excluded from future rounds. An important feature of the final round is the use of a 5-point Likert scale for voting, with the goal of reaching consensus. Less critical features include the option to use a 9-point scale and the absence of a requirement to reach out to panel members outside the consensus.

To achieve consensus, it is critical to regard disagreements as equally important as agreements. When defining the strength of consensus, the following scale can be employed: 100% for unanimous consensus, 90% to 99% for very strong consensus, 80% to 89% for consensus, and 70% to 79% for weak consensus. Given that 100% consensus is unrealistic, a minimum threshold of 80% is recommended. In analyzing the consensus statements, a comprehensive analysis—including median and interquartile ranges—was deemed important, while the use of descriptive statistics for preliminary workup was considered less significant.

The results of this Delphi consensus project on conducting a Delphi consensus study reinforce previous recommendations.[11,12,15,18-21] However, in contrast, this Delphi study now establishes a set of clearly defined steps that should be followed when designing and conducting a Delphi study, providing a structured framework for its implementation. The use of Delphi techniques has significantly increased over the past few decades, addressing a crucial gap in achieving consensus among experts. Khodyakov et al.[25] showed an average growth rate of 300% per decade since the 1970s, with the number of published articles rising from 219 in the 1970s to 6,587 in the 2010s, as indexed in PubMed. In the field of medicine, the number of published articles grew from 36 in the 1970s to 4,302 in the 2020s.[25]

In principle, there are 2 approaches to the Delphi method.[15] The classic Delphi relies on anonymity, iteration, feedback, and statistical summary. It begins with a round of open-ended questions and concludes

Table 5. Updated Principal Steps of the Delphi Technique

	Critical and Important	Recommended	Less Important	Comments
Principle Philosophy	**Ensure**: • clear definition of consensus • iterative process • anonymity and blinding			
[1] Define the problem		• Consider a narrative or scoping review by steering committee		
[2] Steering Panel	• 3-5 members • Must be content experts • At least 1 Delphi expert		• Independent steering panel, can also vote	
[3] Panel Selection	• Expertise must be clearly defined • Content experience critical, demonstrated through publications, presentations, and clinical practice • Panel heterogeneity important • Regional experts for regional topics • Minimum panel size of 10	• Panel size preferably 15-20 • Size for broader topics at least 20-30 • Previous Delphi experience • Attrition rate <10% but 20% acceptable	• Unbiased panel • Geographic variability unless regional consensus is the aim • Referral of friends or colleagues • Panel can have conflict of interests as long as they declare	
[4] Question Development	• Responsibility of steering committee • Input of panel members for round 2 if topics omitted in round 1			
[5] Round 1	• Open-ended questions: clear and unambiguous language • Fast distribution via e-mail or other mediums	No Likert-style questions in round 1		
[6] Summary and Feedback	• Qualitative analysis providing the basis for rounds 2 + 3 • Controlled feedback between rounds to panel: written summary or graphical representation			
[7] Round 2	• Combination of open- and closed-ended questions • Ranking techniques can be included			Consensus begins forming

(continued)

Table 5. Continued

	Critical and Important	Recommended	Less Important	Comments
[8] Summary and Feedback	• Qualitative analysis providing the basis for rounds 2 + 3			
[9] Round 3/Final Round	• Designated for voting on consensus • Voting via 5-point Likert scale on presented statement	• A 9-point Likert scale is a feasible alternative	• No strong requirement to contact panel members outside consensus	
[10] Consensus	• Minimum agreement for consensus 80% • Disagreement as important	• 100% unanimous, 90%-99% very strong, 80%-89% consensus, 70%-79% weak consensus		Consensus has been formed
[11] Conclusion	• Summarize consensus based on the established criteria	• Feedback to panel and submission/ publication		Final consensus statement
[12] Important Considerations	• Timeline for completion not to exceed 6 months • Time between rounds not to exceed 2 weeks • Time for panel not to exceed 4 weeks • Limit reminders to panel to 2-3; if no response, exclude from future round • If more than 2 rounds with Likert-style questions, check stability of responses • No universal number of rounds but limited to 3 rounds in general sufficient			

with a round of closed questions, ultimately resulting in a group summary and the aggregation of consensus.[15] The modified Delphi technique employs a similar methodology, with the primary distinction being the inclusion of an in-person meeting facilitated by a skilled moderator.[15] This meeting allows for discussion of the results and re-evaluation of findings through dialogue among panelists.[15] In the final rounds, panelists use a 9-point Likert scale to vote on the items deemed appropriate.[15] Naturally, the latter approach could introduce bias, as younger panel members might feel intimidated and alter their final votes. Additionally, the discussions during the in-person meeting could influence panelists to change their votes based on group dynamics. To mitigate these potential issues, the classic Delphi method may be the preferable choice. This project has based its expert consensus on the classic Delphi approach.

The reporting quality of consensus studies is frequently inadequate, with common issues including panel composition, definitions of consensus, and the thresholds for achieving it.[22] Other concerns involve the roles of the steering committee, chair, and co-chair, as well as potential conflicts of interest among panelists and funding sources.[22] Data extracted from included studies often revealed additional aspects not captured in the extraction form, such as justifications for deviations from the protocol or incentives to encourage panelist participation.[22] Spranger et al.[26] have proposed reporting guidelines for health sciences studies, which encompass 15 items. These guidelines cover essential aspects, including title and abstract requirements, as well as sections on discussion, limitations of findings, and dissemination strategies.[26]

Later, Gattrell et al.[27] developed the ACCORD checklist as a reporting guideline applicable to all consensus-based studies. This checklist includes 1 item for title wording, 3 items for crafting the introduction, 21 items for reporting methodology, 5 items for presenting results, 2 items for the discussion, and 3 items for addressing additional information.[27] The authors believe that it will assist authors in crafting accurate and detailed manuscripts, thereby enhancing the completeness and transparency of reporting while providing readers with clarity on the methods used to reach consensus. Additionally, the checklist will clarify the rigor of the consensus methods employed to guide the recommendations. This checklist seems highly comprehensive and should be considered when drafting a consensus manuscript.

Limitations

Delphi consensus is often criticized for its small participant sample size, which may not accurately represent a larger group of experts.[28] Additionally, panel selection is frequently nonrandom, as panelists are typically chosen based on their expertise and clinical experience.[29] This can introduce bias, as these experts may hold strong opinions and potentially have conflicts of interest, such as being paid consultants for implant or drug companies. Moreover, the Delphi method has faced criticism for its low sensitivity and replicability, as outcomes can depend heavily on both the opinions expressed and the size of the panel itself.[30,31] Increasing the number of panel members may lead to different consensus outcomes for certain items. Additionally, panel members may not always be impartial. However, these concerns may be more closely related to study execution than the methodology itself. Boulkedid et al.[32] investigated the quality of reporting in Delphi studies over a 30-year period and found that the methodological quality was insufficient for several items. Specifically, only 39% of studies reported response rates for all rounds, 60% indicated whether feedback between rounds was provided, and only 77% detailed the methods used to achieve consensus.[32] One could argue that these shortcomings reflect more on the authors and the review process than the Delphi technique itself. Other limitations include the potential biases introduced by the steering committee members and their preconceived notions about the Delphi method. Additionally, the selection process for panelists may have influenced outcomes, compounded by the small number of participants. The study's reliance on English may exclude valuable insights from thought leaders in other languages. Furthermore, the emphasis on application to orthopaedic surgeons may limit the generalizability of our findings to other medical and health care domains, as well as to broader contexts outside of health care.

Conclusions

Unanimous agreement was reached for iteration, the ability to establish treatment guidelines, a proven track record of panel members, and the requirement for at least 1 steering committee member to be a Delphi expert. Very strong consensus was reached on several key requirements: a clear definition of consensus, controlled feedback between rounds, precise definitions of expert and expertise, and the need for panel members to show experience through publications and clinical practice. Criteria for panel selection should ensure diversity and specialization, with steering committee members being content experts and a minimum of 20 to 30 panel members for broader topics. Regional experts should provide consensus on specific topics only. The steering committee should develop questions, with open-ended questions in round 1 and both types in round 2. Limiting the process to 3 rounds is advisable, aiming for at least 80% consensus in the final round.

Disclosures

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: E.H. is on the Editorial Board for *Arthroscopy* and *The Hive—Musculoskeletal Journal*. J.C. is a board or committee member of the American Orthopaedic Society for Sports Medicine, Arthroscopy Association of North America, and International Society of Arthroscopy, Knee Surgery, and Orthopaedic Sports Medicine; is a paid consultant for Arthrex, CONMED Linvatec, Ossur, and Smith & Nephew; and is a paid presenter or speaker for Smith & Nephew. R.F.L. is a consultant or advisor for Smith & Nephew and Ossur; receives funding grants from Smith & Nephew and Ossur; has received travel reimbursement from Smith & Nephew; has a patent with royalties paid to Robert F. LaPrade; is on the Editorial Boards for *AJSM*, *JEO*, *KSSTA*, *JKS*, *OTSM*, and *JISPT*; and has received research grants from AOSSM and AANA. F.M. is a consultant or advisor for Exactech and Arthrex. B.A.M. is a consultant for Arthrex and Bioventus. G.M. has received funding grants from Smith & Nephew, has received speaking and lecture fees from Smith & Nephew, is a consultant for Smith & Nephew, and is on the Editorial Board for the *Journal of Arthroscopy* and *JBJS*. J.C.M. is on the Editorial Committee for International Society for Arthroscopy, Knee Surgery and Orthopaedic Sports Medicine (ISAKOS). I.M. is a consultant for Stryker and has served as faculty in educational events organized by Smith & Nephew and Arthrex. I.W. is a consultant or advisor DePuy Synthes Mitek Sports Medicine, Smith & Nephew, CONMED Corporation, and Bioventus LLC; is on the Editorial Board for the *American Journal of Sports Medicine* and *Journal of Arthroscopy*; is a board member of AANA, ISAKOS, and AAC; and is associate editor of the *Orthopaedic Journal of Sports Medicine*. N.P.J.P. has received travel reimbursement from the Arthroscopy Association of North America and The Society of Military Orthopaedic Surgeons, is a board member of The Society of Military Orthopaedic Surgeons, and is an associate editor for *Arthroscopy*. All other authors (P.B., D.B., A.G., S.H., S.H-M., E.H., J.C.M., M.N., U.R., Z.S., S.W.) declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: What it is and what it isn't. *BMJ* 1996;312:71-72.
2. Blanco MA, Capello FC, Dorsch JL, Perry G, Zanetti ML. A survey of evidence-based medicine training in US and Canadian medical schools. *J Med Libr Assoc* 2014;102:160-168.
3. Maggio LA, Cate OT Chen HC, Irby DM, O'Brien BC. Challenges to learning evidence-based medicine and educational approaches to meet these challenges: A qualitative study of selected EBM curricula in U.S. and Canadian medical schools. *Acad Med* 2016;91:101-106.
4. Simon F. Evidence-based medicine: Fact or proof? *Eur Ann Otorhinolaryngol Head Neck Dis* 2020;137:361-362.
5. Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome. *Cochrane Database Syst Rev* 2017;2:MR000033.
6. Sheridan DJ, Julian DG. Achievements and limitations of evidence-based medicine. *J Am Coll Cardiol* 2016;68:204-213.
7. Furberg CD. To whom research findings apply? *Heart* 2002;87:570-574.
8. Woodbury MG, Kuhnke JL. Evidence-based practice vs. evidence informed practice: What's the difference? *Wound Care Canada* 2014;12:18-21.
9. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN. Revolutionizing healthcare: The role of artificial intelligence in clinical practice. *BMC Med Educ* 2023;23:689.
10. Subbiah V. The next generation of evidence-based medicine. *Nat Med* 2023;29:49-58.
11. Hohmann E, Brand JC, Rossi MJ, Lubowitz JH. Expert opinion is necessary: Delphi panel methodology facilitates a scientific approach to consensus. *Arthroscopy* 2018;34:349-351.
12. Hohmann E, Cote MP, Brand JC. Research pearls: Expert consensus based evidence using the Delphi method. *Arthroscopy* 2018;34:3278-3282.
13. Hohmann E. Editorial commentary: Wider acceptance of medical expert consensus requires strict adherence to Delphi panel methodology. *Arthroscopy* 2024;38:250-252.
14. Minas H, Jorm AF. Where there is no evidence: Use of expert consensus methods to fill the evidence gap in low-income countries and cultural minorities. *Int J Ment Health Syst* 2010;4:33.
15. Khodyakov D, Grant S, Kroger J, Bauman M. RAND methodological guidance for conducting and critically appraising Delphi panels. https://www.rand.org/pubs/tools/TLA3082-1.html. Accessed September 18, 2024
16. Green B, Jones M, Hughes D, Williams A. Applying the Delphi technique in a study of GP's information requirements. *Health Social Care Comm* 1999;7:198-205.
17. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey. *J Adv Nurs* 2000;32:1008-1015.
18. Shang Z. Use of Delphi in health sciences research: A narrative review. *Medicine (Baltimore)* 2023;102:e32829.
19. Beiderbeck D, Frevel N, von der Gracht HA, Schmidt SL, Schweitzer VM. Preparing, conducting, and analyzing Delphi surveys: Cross-disciplinary practices, new directions, and advancements. *MethodsX* 2021;28:101401.
20. Nasa P, Jain R, Juneja D. Delphi methodology in healthcare research: How to decide its appropriateness. *World J Methodol* 2021;11:116-129.
21. Humphrey-Murto S, Wood TJ, Gonsalves C, Mascioli K, Varpio L. The Delphi method. *Acad Med* 2020;95:168.
22. van Zuuren EJ, Logullo P, Price A, Fedorowicz, Hughes EL, Gattrell WT. Existing guidance on reporting of consensus methodology: A systematic review to inform

ACCORD guideline development. *BMJ Open* 2022;12:e065154.
23. Hohmann E, Angelo R, Arciero R, et al. Degenerative meniscus lesions: An expert consensus statement using the modified Delphi technique. *Arthroscopy* 2020;36:501-512.
24. Hurley ET, Matache BA, Wong I, et al. Anterior shoulder instability part I—diagnosis, nonoperative management, and Bankart repair—an International Consensus Statement. *Arthroscopy* 2022;38:214-223.
25. Khodyakov D, Grant S, Kroger J, Gadwah-Meaden C, Motala A, Larkin J. Disciplinary trends in the use of the Delphi method: A bibliometric analysis. *PLoS One* 2023;18:e0289009.
26. Spranger J, Homberg A, Sonnberger M, Niederberger M. Reporting guidelines for Delphi techniques in health sciences: A methodological review. *Z Evid Fortbild Qual Gesundhwes* 2022;172:1-11.
27. Gattrell WT, Logullo P, van Zuuren EJ, et al. ACCORD (ACcurate COnsensus Reporting Document): A reporting guideline for consensus methods in biomedicine developed via a modified Delphi. *PLoS Med* 2024;21:e1004326.
28. Akins RB, Tolson H, Cole BR. Stability of response characteristics of a Delphi panel. *BMC Med Res Methodol* 2005;5:37.
29. Khodyakov D, Savitsky TD, Dalal S. Collaborative learning framework for online stakeholders engagement. *Health Expect* 2016;19:868-882.
30. Powell C. The Delphi techniques: Myth and reality. *J Adv Nursing* 2003;41:376-382.
31. Rennie D. Consensus statements. *New Eng J Med* 1981;304:665-666.
32. Boulkedid R, Abdoul H, Loustau M, Sibony O, Alberti C. Using and reporting the Delphi method for selecting healthcare quality indicators: A systematic review. *PLoS One* 2011;6:e20476.

Appendix 1. Summary Round 1 Questions

Philosophy of Delphi

[1] What are the essential features of the Delphi method? What are the nonnegotiable features of the Delphi method?

Summary:

The majority of the panel mentioned blinding and anonymity, an independent steering panel not involved in voting, a heterogenic panel of leading experts with a track record both academically and clinically, iteration, controlled feedback between rounds, and a clear definition of consensus as the nonnegotiable features of Delphi. Please find below a few of the responses.

- The Delphi method is essentially a survey performed following stringent, well-defined standards, resulting in expert synthesis of the best available evidence on a subject.
- It is mandatory to determine the relevant problem to solve, identify and engage a panel of anonymous experts, and create a summary report, and repeat until a consensus is reached (iteration, at least 3 rounds), with the help of a facilitator to manage the process.
- The essential features of the Delphi method are that it provides an expert consensus on topics when level 1 evidence may not exist or on important emerging topics whereby the literature has not caught up to the current findings of an important topic.

Nonnegotiable

Inclusion of leading experts in the field of the subject being studied

Heterogenic panel of anonymous experts on the matter

Independent study facilitator

Blinding of group participants to limit groupthink bias

Balanced inclusion of participants both in terms of geography and practice setting

Evidence-based question generation based on a thorough literature review

Open-ended initial questioning to limit bias

Use of a Likert scale to assess the level of agreement with each statement

Anonymous setup

Controlled feedback to panel

Iteration

Clear criteria for defining and measuring consensus must be established and include the level of agreement required for a consensus

Consensus was defined as achieving between 80% and 89% agreement, strong consensus was defined as between 90% and 99% agreement, and 100% agreement was unanimous

The same questionnaire presented several times (iteration) with controlled feedback

Evaluation (and determination of consensus) takes place via descriptive analyses of the quantitative data, which is supplemented by a systematic evaluation of the open responses

[2] When is the Delphi method the most appropriate method and, conversely, when is it not a suitable tool?

Summary: Below are the most common responses. The majority believed that Delphi should be employed if strong evidence does not exist or when there is a high level of controversy.

- Most useful for topics that are common and have a high level of controversy in the literature due to a lack of high-quality evidence, ambiguity, or conflicting high-quality evidence.
- It is less suitable when there is robust evidence supporting a certain treatment option/plan of action, for example.
- It can be great for decision-making algorithms, as well as discussion of technical pearls around a procedure.
- Appropriate to establish guidelines, solving complex issues, helpful for novel interventions, highly polarized topics, if evidence supports more than 1 answer.
- Not suitable for a representative opinion, not suitable for a very small panel, for analysis of literature, not suitable for rapid decision-making, not suitable if Level of Evidence (LOE) I exists or if there is overwhelming evidence in support of a single answer, not suitable to replace LOE I randomized controlled trials, not suitable if expert panel cannot be accessed

[3] What are the limitations of Delphi? How would you mitigate or overcome these limitations?

Summary: The panel is concerned with a plethora of potential limitations, and most of them are potential valid concerns.

- Level V evidence, which cannot be overcome
- Inherent bias in the selection of study participants (one method of overcoming this is having the study facilitator who is not involved in the voting process scan recent association meetings [AOSSM, AANA, ESSKA, etc.] and journal publications for potential participants based on academic productivity/expertise on the study subject. Also, it is important to ensure geographic and practice type variability to increase generalizability of findings)
- Lack of standardized process for generating the questions (thorough literature review to generate the questions, independent generator of questions not involved in voting process, use of a Likert scale when responding)
- Lowest common denominator of expert
- Uses controlled feedback, no open discussion
- Time (and money) consuming

- Results highly dependent on the expertise of the panel/responders
- Limited ability to (1) generalize the results and (2) contribute to theory-building with the Delphi study
- Lack of objectivity in selection of panel members
- Categorization of comments from experts
- Participant attrition: the overall consensus loses robustness
- "Experts" who get overlooked in a Delphi consensus that may have important input, which could sway the directions and recommendations of the method
- Attainment of consensus (or dissensus) does not imply the "correct" answer
- Dropout of panelists during process, response fatigue
- Expert selection bias, attrition, lack of face-to-face interaction, and an over-reliance on subjective judgments

[4] Can you comment on the CREDES criteria, and how would you perceive their relevance and applicability?

Summary: Most support the use of CREDES as a helpful guideline, but some feel that they are not appropriate for orthopaedic studies as they were developed for palliative care.

Helpful guidelines for conduction of a Delphi: important criteria to use going forward to help standardize how these studies are being performed, similar to the PRISMA guidelines for conducting systematic reviews.

- It was developed for palliative care specifically, so I have not used it as it seems less relevant given the differences in specialties.
- The CREDES criteria reflect a comprehensive set of guidelines, the vast majority of which is applicable to MSK research.
- Aspects of the CREDES criteria such as the panel selection, iterative rounds, and data analysis are instrumental in enhancing Delphi studies and are both relevant and applicable to Delphis across all research areas.
- A reporting guideline is very important. But it must apply to all areas of health research and to all variants of Delphi. That is why CREDES is difficult.
- The CREDES criteria represent a well-thought-out process that ensures the majority of Delphi studies follow a standardized process.
- I have reviewed the CREDES criteria. These are obviously well thought out. As it pertains to applications to medicine (and specifically to orthopaedics) I believe that consensus does not need to be defined in binary terms (yes/no). Stratifying consensus and reporting on the strength of consensus provides valuable information.
- Junger et al. (5) published the CREDES criteria to allow for rigorous conduct of studies using the Delphi technique. Much like the CONSORT-like guideline, as is used for randomized controlled trials, it allows for better guidelines for Delphi studies.
- I fully support the application of the CREDES criteria as a standard approach to Delphi studies. For me they seem to be understandable and feasibility in implementation is given.
- I am not very familiar with it, but I believe that the CREDES criteria are a set of guidelines designed to improve the reporting quality and transparency of Delphi studies. They ensure that the Delphi process is well-documented and that the findings are reliable and reproducible. They are a great start, as they offer good guidelines on study design and provide useful definitions.

Panel Creation

[5] Which criteria would you apply when selecting a Delphi panel?

Summary: Panel creation should be balanced and heterogenic unless a very specific narrow topic, which requires special expertise, is investigated. Some panel members mentioned panel size, but this variable needs to be examined in more detail.

- Content expertise regarding a particular topic based on academic productivity on that topic, also clinical practice experience
- Geographic variability to ensure balanced participation
- Inclusion of allied health with expertise when appropriate
- Absence of/limited financial bias regarding the subject
- Participants need to both actively treat and study the topic (i.e., they need to be performing research or give lectures at scientific meetings on this topic)
- Heterogeneity of experts—facilitators of the study have the responsibility to make sure that we don't have a bias or unbalance in the expert panel
- Size of the panel—for statistical matters, I would have at least 20 experts in the panel and if we aim to differentiate between groups, each subgroup should have at least 20 experts. It is not necessarily the more the merrier, but 20 would be my lower end, except for very, very specific topics, where experts are very rare.
- The size of the panel often depends on the complexity of the topic, but larger panels, such as 50 members, complicate logistical considerations and make it more difficult to reach a consensus.
- According to CREDES criteria: transparent information on recruitment of the panel, sociodemographic

details including information on expertise regarding the topic in question, and (non)response and response rates over the ongoing iterations should be reported.
- Clearly define "experts"
- Individuals with very strong opinions and limited willingness to reflect on feedback and other views should not be chosen

[6] What qualifications and experience would you deem essential for individuals to have in order to be eligible to serve on a Delphi panel? We are interested in understanding what you believe are the minimum necessary qualifications versus the desired or ideal qualifications.

Summary: Most of the panel believed that expertise and a track record within the topic to be investigated is essential.

- Minimum necessary qualifications would be a proven track record of either academic (ideally) and/or clinical productivity on the subject being studied, work and research experience for at least 2-3 years.
- Experience performing Delphi method studies would be an asset.
- A geographically diverse panel helps to achieve a broader perspective and generalization of consensus. The expert can be defined as someone with knowledge and experience on a particular subject. However, it is practically difficult to measure experience quantitatively. Panel selection should adhere to predefined criteria. In the current project, I would say someone with experience in academics, writing scientific papers, and prior knowledge on what a Delphi method is, is the target. A homogeneous group, on the other hand, may be more reliable in a particular study objective. This homogeneous panel is suitable when resolving unsettled issues of a focused problem like the current topic.
- The selection of a panel or voting members in a Delphi study should be based on objective and predefined criteria and related to the problem under study, but the definition of an expert can vary depending on the purpose of the study.
- Minimal qualification: Subject matter expertise and professional expertise in the investigated topic. Committed availability and willingness to actively participate.
- Desired/ideal qualification (in addition to the items above): Multiyear professional expertise with recognized track record of achievements, conducted and published research in the investigated subject, communication skills and the ability to articulate complex ideas clearly and effectively.

[7] When selecting a panel member, do you believe that choosing a friend or colleague, who is perceived as an expert, may introduce bias or interfere with the Delphi process? If so, how would you mitigate this potential influence?

Summary: Friends and colleagues could be recommended as panel members as long as selection criteria and anonymity/blinding are not compromised. On the contrary, some panel members believed that this practice would introduce bias and should be avoided if possible.

- Not necessarily as long as the person is not closely involved in the research process itself. However, I would recommend avoiding including too many peer groups to not risk a completely biased subset of experts.
- As long as the selection criteria are not compromised, choosing a colleague is reasonable.
- Yes. There is a potential for bias. However, if there is a research team/steering committee who is overseeing the development of the Delphi panel, then one mitigation strategy would be to have that friend or colleague presented to the other directors who would then decide on exclusion/inclusion as a group.
- If an alternate to the friend is appropriate, they should be selected in lieu. If the "friend" really fits a desirable demographic, then close attention to anonymity should reduce bias. Comments from participants may need to be slightly modified to maintain anonymity.
- No, as it should ideally be anonymous, and panel members should be blinded to the identities of other members.

[8] When creating a Delphi panel, should you prioritize a panel with a diversity of experiences and backgrounds or a more focused, specialist panel? What factors ensure the effectiveness of Delphi within this context?

Summary: There was no clear trend. Diversity of experiences and backgrounds were preferred, but others felt that a specialist panel was more appropriate to achieve consensus.

Diversity should be prioritized for reasons stated above.

- Heterogeneity is more important in a more ambiguous, future-oriented setting.
- If you want an answer/consensus from specialists, you can just include specialists in the panel. This would mean you only have responders with a specific amount of experience that needs to be determined.

- The labeling of panel members as "experts" is tricky. Despite its controversy, as far as I know, experts are commonly used in the Delphi studies for panel members. Experts should possess relevant knowledge, expertise, and experience related to the topic under investigation. The common goal behind using experts is to increase the qualitative strength of recommendations or consensus.
- The decision of diversity versus high specialization should be made when the research topic is defined and scoped. For multifaceted topics that are complex, a diverse panel might be the best solution; if the topic is very focused and specialized/technical, a specialist panel might be the best choice.
- Diverse panels might have a more visionary, innovative approach to the topic while achieving consensus might be more challenging. Specialty panels might become quite narrow in their assessment of a topic while consensus might be reached easier.

Question Creation

[9] For first-round, open-ended questions, what are the characteristics or qualities of a well-written, effective question?

Summary: Most preferred open-ended questions for round 1, which should be neutral with room for individual responses.

- Questions should be clear and concise, focusing on a specific point, with ample room for individual interpretation and responses.
- Clear questions without risk of misinterpretation or ambiguity
- Well-crafted open-ended questions should be clear and concise, without implying a single correct answer. Additionally, an ideal open-ended question should be neutral, encouraging thorough and creative responses.
- It should be using simple language that avoids jargon or complex terms. It should be directly related to the main topic, tapping into the panelists' expertise and encouraging insightful responses. The question needs to be open-ended, prompting detailed and thoughtful answers rather than simple yes or no responses. Neutrality is crucial, so the question should be unbiased and objective, allowing for genuine and diverse opinions.

[10.1] Do you believe it is more effective for the research team/steering committee to formulate questions for round 1, or should the panel be given the opportunity to submit topics of interest?

Summary: In general, the panel preferred question development by the steering committee. However, some felt that there should be room for the panel to add questions for round 2 if important topics were not addressed by the steering committee.

- Generally, questions should be generated by the steering committee that are based on a thorough literature search/review, are more relevant and of higher quality than ones generated by the panel-at-large, which have the potential to be more "off-the-cuff" and less evidence-based. However, allowing for potential feedback is useful.
- Develop the questions and projections based on a suitable framework and discuss this in a pre-workshop together with a set of selected experts.
- A smaller panel can then give indications or vote for topics or specific projections—however, this also depends on the topic.
- The idea of being given the opportunity to submit topics or questions (or at least write in additional questions) is beneficial. If the research team feels that the panel omitted certain important questions, then these can be added on subsequently.

[10.2] Can you envision where utilizing both techniques would be beneficial?

Summary: In addition to [10.1], the panel felt that there should be an option to add questions.

- Yes, for studies focusing on a topic with less high-quality evidence, the latter option could prove to be more fruitful.
- Yes. The latter may be helpful if the topic is relatively narrow or focused. Otherwise, a large panel may provide less relevant questions that create distractions and add "noise."
- Yes, where the topic/field is relatively unexplored and new.

Structure and Flow of Delphi

[11] How would you determine the optimal number of rounds? What factors dictate the minimum number of rounds to achieve consensus and substantive valid outcomes?

Summary: In general, rounds should be limited, and the majority felt that 3 rounds should be sufficient. Some felt that reaching consensus is more important than the number of rounds.

- 3 voting rounds are sufficient. However, the utilization of the above-stated CREDES criteria may be helpful in deciding this.
- Minimum of 3 rounds, initially broad, then close ended, allow feedback on statements, and then voting.
- The process can continue until there is little change between rounds i.e. a steady state is reached.
- Three rounds are enough in most cases. However, depending on the level of consensus achieved in each round, additional rounds can be conducted to further

refine the opinions. At the end, the strength of this method is founded in iteration.
- Deadly, the number of rounds is not fixed in advance—the Delphi process is only ended when consensus is reached, or the judgments no longer converge (stability). However, this rarely happens, usually some judgments remain in dissent and the stability of the judgments is not checked.
- The optimal number of rounds depends upon when a consensus can be obtained. For the first round of the survey, statements were categorized as "essential" and retained for the second round if both $\geq$70% of respondents agreed *and* <20% disagreed. In particular, if there are no further questions or comments after the third round, having a minimum of 3 rounds is effective. The factors which dictate the minimum number of rounds to achieve consensus and to provide substantive valid outcomes statements are established when a final consensus of >75% agreed and <10% disagree with the questions.
- Obtaining consensus is more important than limiting the number of rounds to achieve the outcome. I'm on round 6 for another Delphi study currently.
- Very high response rate over 5 rounds of Delphi voting is an achievable goal (Turnbull 2018). This was done through e-mail reminders, receiving phone calls or text messages, with only 4% panelists reported being bothered by these reminders. The impressive response rate of 90% over 5 rounds of the Delphi may have been related to what appears to be a highly motivated panel and multiple reminders (Turnbull 2018).
- In order to avoid "false" consensus (participants with dissenting views start dropping off), studies have instructions to participants emphasizing the importance of completing all rounds. A useful example is provided by Sinha in their paper. It should also be clear to participants that they do not need to conform.
- If one of these rounds struggles in producing the expected outcome, additional rounds should/can be concluded with the goal to "sharpen" the conclusion and the quantitative consensus building.
- Typically, 3 rounds are considered sufficient for most studies. In the first round, open-ended questions gather a wide range of opinions. The second round refines these responses, presenting them back to the panel for further input and narrowing down the options. The third round aims to finalize consensus by allowing panelists to re-evaluate their positions based on the collective feedback.

[12] For the second round, should open- or closed-ended questions be employed? When should the panel transition from open-ended to closed-ended questions?

Summary: Most members felt that round 2 should mainly consist of closed-ended questions.

- The panel should transition to closed-ended questions once the responses to that specific question fail to produce additional unique responses after the previous voting round. As such, questions may be converted to closed-ended ones at different rates depending on each round's responses.
- If first round is open, second should be closed and then a third round is needed from my perspective.
- I personally wouldn't count the rounds with open-ended questions to be honest. For me, these are pre-phases of the actual Delphi study.
- For rounds 1 and 2 panelists should have the opportunity to suggest modifications or suggest new items.
- As a Delphi study progresses, questions should become more tailored and direct. Therefore, later rounds should avoid open-ended questions to facilitate a final consensus among expert panelists [11]. For the second round, closed-ended questions may be employed if the original stance of each expert is similar enough. Ideally, the second round of a Delphi study should provide participants the opportunity to provide feedback on all responses. The third round would then be conducted by administering a questionnaire with close-ended questionnaire to panelists or participants to reach consensus.
- In the second round the responses from round 1 are summarized and compiled by the facilitator. This summary is then used to create a more focused and structured questionnaire for the next round. Therefore, mainly closed-ended questions should be used.
- I'm not sure if the transition from open-ended to close-ended questions takes place at the second or third round, but once the questions have been refined then closed-ended questions are needed so that a consensus can be evaluated. In other words, you need "yes-no" questions to determine how many "yes" responses and how many "no" responses.
- The second round should be close-ended questions, but still more open than the third round. This is to ensure efficiency and to reduce the risk of straying outside of the scope of the study.

[13] How should feedback be organized between rounds? Does this vary between rounds with open-ended questions and closed-ended questions?

Summary: There is overall agreement that feedback between rounds is an essential part of Delphi.

- Feedback should be standardized between rounds.
- Depends on the structure. Sometimes no feedback in terms of aggregation or intervention of the facilitator is also valid.
- There needs to be a balance between data on consensus and stability and so on vs. no feedback at all. Sometimes I get the impression that too much

feedback leads to a quicker survey fatigue and hence probably poorer results.
- Feeding back percentages and providing individual round scores for every item would be nice. This provides a visual means for the panelists of assessing the diversity of responses. It also allows them to check that researchers have recorded correct responses.
- Important: quick and easy to understand (no complicated statistics, not too much text), graphical frequency tables are nice.
- Feedback should be organized by question/topic and panelists should have the opportunity to comment or dialog. Ideally the final round is performed in person as during the 2018 International Consensus Meeting for MSK infections.
- I would use a quantitative scale in rounds 2 and 3 to collect level of agreement/disagreement. In round 2 and subsequent rounds I would allow for comments and explanations for rating.

[14] Would you consider employing Likert-style questions in early rounds, followed by eliminating items that fail to achieve consensus as a viable strategy, or do you perceive it as potentially deviating from Delphi principles?

Summary: There is no agreement with regards to Likert yes/no style questions in early rounds.

- I think employing Likert-style questions is important from the get-go, but items that fail to achieve consensus should not be eliminated, as the reporting of a lack of consensus re: a specific question is a potential source for future research direction.
- For me this is fine, but in both ends—both if consensus is reached quickly or if a certain structure of dissent is visible. If there is just a general lack of consensus without any pattern, I would keep the questions and see if more iteration leads to consolidation or polarization, for example.
- No, I think that this can over complicate the process—broadly agreeing/disagreeing is fine initially.
- Incorporating Likert-style questions in early rounds could prove to be beneficial in establishing a baseline of the group's opinions, behaviors, and attitudes on a given topic. However, if employed, these questions should only be administered in early rounds. Eliminating items that fail to achieve consensus as a viable strategy could compromise the true results of a Delphi study, especially if the group is small.
- The Delphi method can be modified to suit the topic that is being addressed. Likert-style questions are reasonable, but one concern is that too many options (for example 7 or 9) are quite taxing on the panelist. I have experienced this first-hand. Long lists of Likert-style questions may also increase the potential/risk of fatigue.
- I would recommend using Likert-style questions/rating features already in round 2 as well as in any subsequent rounds. I would eliminate items that do not reach a predefined threshold very early.
- Overall, no. This is because it would be too limiting, and that it deviates from the traditional approach of Delphi studies. However, there is value to this approach, as it would allow us to get a wide range of opinions and measure the level of agreement on various items right from the start.

[15] What is the optimal maximum time frame between rounds and for the overall completion of a Delphi project?

Summary: Suggested timelines to complete a Delphi project varied widely.

- 4 weeks for round 1, then 2 weeks per round.
- Overall completion targets 8 weeks.
- I would say anything between 2-6 weeks between the rounds is ok.
- Optimal time is for me less than 4 weeks and 3-5 months in total.
- The optimal maximum time frame between rounds should be 2 to 3 weeks. An ordinary Delphi study is comprised of at least 4 rounds. Two to 3 weeks is enough time allotted to ensure most or all participants respond, while giving participants enough time to develop their thoughts and opinions without feeling rushed through the study. The maximum time for completion of a Delphi project should be 2 years. Studies that contain less participants with more similar ideologies could conclude in as little as 4 to 6 months.
- Four weeks between rounds, overall, 6 months
- Long enough that there is adequate time to process the information from the previous round to drive the specific questions for the next round, but not so much that the panelist forgets what was presented in the previous round. I would guess a few weeks to a couple of months between rounds is probably about right for an expansive topic (such as anterior or posterior shoulder instability).
- Two, maximum 3 weeks, with one reminder e-mail.
- Ideally 6 weeks, not more than 3 months.
- My recommendation would be 2 to maximally 4 weeks duration for every round. The overall duration depends on the number of rounds conducted. My recommendation would be not to exceed a total duration of 6 months.
- No exact time, as it depends on the time commitments of each individual member. For example,

professors during a busy grant writing season would require a larger time frame than most other participants. Perhaps 2-4 weeks would be a good balance, with a 1- to 2-week period of data analysis between each round, and for 3 rounds, would thus be 12-18 weeks (4-6 months).

[16] What constitutes an acceptable attrition rate of panel members?

Summary: The majority of the panel suggested that an attrition rate of 10% and less is acceptable. An important point was that the response rate for each round was often not reported.

- <10%
- <10% to 20%
- Although there is no set cutoff for attrition rate for Delphi panels, high attrition rates can be problematic. If individuals with opinions that disagree from the consensus end up dropping out of the panel, there can be significant response bias [13]. In general, an attrition rate of 20-30% can be expected and would be deemed acceptable [14].
- I really don't know, but applying a common rule in scientific studies, we can say that a dropout rate of 20% to 30% can be expected after some rounds. Of course, this should be taken into consideration when determining the panel sample size.
- The response rate is often not reported (Sinha 2011; Humphrey-Murto, Varpio 2017).
- Depends on the size of the panel. Larger panels can tolerate higher attrition rates while small panels with very high expertise levels should be more cautious on maintaining engagement. As a rule of thumb, I would suggest that in panels with more than 10 panelists a rate of 20% is acceptable.
- Typically, 30% or less, but can be variable, and it should not necessarily reflect the quality of the study.

Final Round

[17] What are your thoughts on the optimal structures and objectives for the final round?

Summary: The last round should be for final voting and consensus. Interestingly, most panel members did not address how this should be achieved.

- The main objective for the final round is to vote on agreement with the final statements, not necessarily to achieve consensus on all the statements. Disagreement can be as important as agreement.
- Prior to the beginning of the final round, there should be ample time for analysis of discussion from the previous rounds. Data from previous rounds should be presented in a way that is easy to evaluate, and controlled feedback should be provided while maintaining anonymity amongst included panelists. Ideally, there will have been a sufficient number of previous rounds to have drawn the group closer to reaching a consensus statement with a high percentage of agreement while maintaining stability of responses throughout the process.
- The repetitive and interactive survey rounds should be useful for gathering qualitative information, improving framing of the statements for panel members, and achieving consensus, which is the final goal.
- My thoughts on the optimal structures and objectives to the final round are to achieve consensus on the topics. Sometimes, this means that areas where one may feel are essential, are found not to be included in the final Delphi statement. It is important to recognize that a Delphi panel is a "give-and-take," whereby there are times that one's strong beliefs in a particular area may not be supported by the experts that are on the panel.
- Ideally, this should be a series of closed-ended concise questions (more complex questions should have been parsed out into 2 or more questions during the earlier rounds). The objective is to evaluate consensus (yes/no/abstain).
- Again, must be fit for purpose. If need to limit the number of items for whatever reason, have participants rank their top 3 or top 10 items. If you are more interested in all the items deemed important, you could have a "verification" round, where all participants confirm important items to include.
- The final round should focus on measuring the consensus/agreement and the elimination of findings and items that fail in a predefined threshold of agreement.

Achieving Consensus

[18] What percentage of agreement between panel members do you consider to determine and ensure robust and meaningful consensus?

Summary: Suggestions ranged from 70% to +90% and some authors suggested to include the frequency of disagreement as an additional variable limiting this to 10%.

- 90-99% agreement with a statement constitutes robust (strong consensus/unanimous) agreement.
- I go with interquartile ranges and they should be below 1/5 of the scale that you are using.
- 80% consensus, 90% strong consensus, 100% unanimous
- *80-100% strong*
- 75 agreement, less than 10% disagreement is widely used.
- 80%
- In general, 70% consensus agreement with a statement with disagreement from fewer than 10% of individuals would ensure robust and meaningful consensus [14].
- *In my humble opinion, as much as possible. However, around 75% should be considered adequate.*

- At least 70%
- The percentage of agreement between panel members to ensure a robust and meaningful consensus for the final round should be >75% agreed and <10% disagree.
- That differs from topic to topic. I have appreciated studies that have explored strength of consensus: 100% unanimous consensus; 90-99% very strong consensus, 80-89% consensus, 70-79% weak consensus. Or something similar.
- 90% would be ideal. Less than that probably does not qualify as consensus.
- Depends on purpose. A typical definition of agreement would be that if 70% of respondents agreed or strongly agreed that an item should be included. Items where 70% disagreed or strongly disagreed should be discarded, and items not meeting these criteria would be sent to the next round for re-ranking.
- Others may decide a median score above a defined threshold is adequate (Tammela 2013). There is great variability in the definition, with some studies recommending 51% for agreement whereas others suggest 70-80% (Hasson et al. 2000). Authors should clearly define how consensus is defined, including the threshold that will be used to determine if consensus was reached.
- Another potential issue is the change in consensus definition and selection criteria between rounds (Boulkedid et al. 2011). There is risk in forcing consensus as it may reduce the credibility of the results.
- A majority vote does not guarantee consensus. It simply reflects that a majority is in favor, but there could be stakeholder groups who consider the outcome unacceptable. For OMERACT, consensus indicates that although the result of the process may not be everyone's preferred choice, the aim is to reach an agreement that all participants can accept as a "working arrangement." It is also noteworthy that OMERACT does not seek to force consensus as this ultimately leads to poor acceptance in the long term (Humphrey-Murto, Crew 2019).

[19] Is there a scenario where achieving 100% agreement among panel members is deemed necessary?

Summary: Some felt that 100% agreement must be reached in cases where safety or life-death scenarios require unanimous agreement. The majority felt that 100% is not a realistic goal in medicine.

- No, because in this case, Delphi is probably the wrong method to be used.
- No, the reality is some people will always disagree if a panel is large enough, very few things are unanimous.
- Requiring 100% can lead to response to bias due to increased attrition. There may be some ethical considerations where 100% agreement is required. However, in most scenarios, an agreement percentage of 70% or higher is adequate.
- Not in orthopaedics, because I think that insisting on 100% agreement might have medicolegal implications. If a group of experts achieve 100% agreement, then does this become a "standard of care"? In my opinion, it is okay to accept that there are always exceptions to the rule . . . or that there always a few contrarians in our midst. And that's okay.
- If the panelists propose a consensus that deviates dramatically from the current standard of care, then 100% consensus would be best.
- If 100% consensus is required the topic normally deals with safety-critical decisions, given legal requirements, or ethical standards. The first decision to make is to decide whether the Delphi method is the appropriate instrument to reach a 100% consensus result—if so, only small high-expertise panels should be used.
- As above, perhaps 90% or higher if it is a life-or-death issue, such as protocols for life-saving measures in the ER or ICU.

[20] Which aspect is more important: Consensus or stability of responses?

Summary: Many members argued that consensus is the goal of Delphi and easier to measure.

- Stability of responses
- Overall consensus and final voting
- For a good study the stability of responses is more important.
- Ideal Delphi outcomes maximize both consensus and stability of responses. Stability is the consistency of individual responses as they progress through subsequent rounds [15]. Although it remains important to obtain a sufficiently high level of consensus at the termination of the Delphi session, a criterion that must be satisfied throughout this process is stability of responses. In other words, consensus is the primary endpoint of interest, but it cannot be adequately attained without appropriate levels of response stability throughout the Delphi process.
- Both! but in practice it's always about consensus.
- The aspect that is more important is to achieve consensus among the Delphi panel members. There are times that some topics cannot be included because there are differences of opinion, but that is the important part about these type of surveys. Achieving consensus is important.
- This might slightly differ based on the topic and goals of the study. I assume that normally consensus is the key driver for the results. Although if stability of responses was seen as challenging issue during the

execution of the study, I would suggest rather adding additional rounds to stabilize and create reliability and confidence in the outcomes.
- Consensus is more important, as it is easier to measure. The definition of stability of responses is controversial, as some argue that the commonly used method for measure stability.

Consensus Statement

[21] What considerations should guide the selection of statistical methods? Are there any standard approaches that should be used?

Summary: Most agreed that descriptive statistics are sufficient and no detailed analysis is required.

- I think this depends entirely on the topic and what is being looked for, many consensus studies only require descriptive statistics: as increases in agreement percentage, convergence of standard deviations of importance ratings, and decreases in number of comments as rounds progress are all indicative of the group progressing toward a consensus. Weight Kappa values can be calculated for within-subject agreement level to account for the magnitude of disagreement between individual answers following a given round.
- Percentage Agreement: proportion of panelists who agree on a particular item. Interquartile Range: to determine the spread of responses. A smaller IQR indicates higher consensus. Median Absolute Deviation (MAD): to measure consensus, especially useful when data are not normally distributed.

Appendix 2. Summary Round 2 Questions

[1] Please rate the importance of nonnegotiable features, with 1 being the most important and 6 being the least important. Additionally, feel free to include any other items you consider mandatory and nonnegotiable.

Ranking:

[1] An iterative process
[2] Anonymity and blinding
[2] A heterogeneous panel of experts with both clinical and academic expertise
[3] A clear definition of consensus
[4] Controlled feedback between rounds
[5] An independent steering panel not involved in voting

[2] Please rate the limitations of the Delphi method according to their importance in introducing bias and error, thereby reducing validity, with 1 being the most important and 9 being the least important. Additionally, feel free to include any other items you consider mandatory and nonnegotiable.

Ranking:

[1] Dependence on the panel and potential selection bias
[2] Inherent bias
[3] Lack of standardization for generating questions
[4] Level of Evidence V (LOE V)
[4] Potential omission of relevant experts, which may influence the direction of the results
[5] Lack of panel objectivity
[6] Agreement based on the lowest common denominator
[7] Geographic bias
[7] Limited generalizability

[3] Could you please comment on the composition of the steering panel?

[a] Is there a minimum number of panel members required, and should they all be experts in the field and the Delphi method?

This question appears to have caused some confusion, with a few responses likely referring to the overall size of the Delphi panel rather than the steering panel. However, the majority agreed that the steering panel should consist of experts in the field, with at least 1 member being proficient in the Delphi technique. The suggested size of the steering panel ranged from 3 to 5 members. Additionally, several participants proposed that the Delphi expert should be an independent panel member, serving as a consensus advisor.

[b] Is it acceptable to have a steering panel member who is not an expert in the field but is familiar with the Delphi method?

Sixty percent of respondents believed that one member of the steering committee does not need to be a content expert but should instead be a Delphi expert. The remaining respondents felt that all members of the steering committee should be content experts but that a Delphi expert should be involved separately to guide the consensus process and prevent errors.

[c] Or would you consider it mandatory for at least 1 member of the steering committee to be a Delphi expert?

While admittedly a repetitive question, we felt it was important to clarify the composition of the steering panel. Similar to question 3b, the majority agreed that the steering panel should include at least 1 member with expertise in Delphi methodology.

[4] The minimum and ideal size of the panel remain unclear. Could you please specify the minimum and ideal number of panel members?

In this question, we focused on the ideal size of the Delphi panel. Most respondents felt that the panel size

should depend on the specific topic. For specialized topics, a minimum of 10 to 15 members was suggested, while broader, more diverse topics might require 10 to 30 members. One panel member referenced a bootstrap study indicating that 23 participants can lead to response stability across rounds. A minority recommended having at least 30 members, emphasizing the importance of ensuring geographical diversity.

[5] The use of friend or colleague referrals requires further exploration. Could you please specify the conditions under which you consider such referrals appropriate and when they are not? For this item, please use bullet points and a yes/no structure.

This question addressed whether friends or colleagues should be part of the Delphi panel, an important issue on which the panel could not reach consensus. Some members felt it was appropriate, provided that the friend or colleague is an expert in the field and the process is sufficiently blinded to ensure independence. However, they noted that if the individual is not a content expert, or if there's a risk of confirmation bias (e.g., "I know what they think anyway"), it would be inappropriate. Others strongly believed that under no circumstances should friends or colleagues be included. However, it was emphasized that if this approach is used, the criteria for inclusion must be clearly defined. One member made a valid point: if the panel and responses are kept anonymous, experts who know each other could still participate. Nevertheless, the risk of sharing or discussing responses remains if participants are aware they are part of the same project.

[6] What are the optimal methods for handling nonresponders after sending out a survey? How should researchers remind panelists to participate, and how many reminders should be sent? If a panelist is dropped from 1 round, should they be allowed to participate in the next round or dismissed from the study?

There appears to be general consensus on this point: sending reminders is reasonable but should be limited to 2 to 3 e-mails. If there is no response after that, the individual should be excluded from future rounds. However, a few panel members suggested allowing nonresponders to participate in subsequent rounds, particularly when the topic is highly specialized and the number of experts is limited.

[7] Does exceeding the acceptable attrition rate of panelists in a Delphi study invalidate its findings? Can the study still be considered publishable despite this? How should researchers proceed in such a situation?

Panel members suggested that an attrition rate of around 20% might be acceptable, provided the panel is large enough. However, others argued that any dropout rate can undermine the validity of the study and may lead to a false consensus. It was agreed that dropout rates should be reported in detail, with potential limitations thoroughly discussed. If these requirements are met, the study can still be considered for publication.

[8] How can researchers designing a Delphi project best prevent burnout or responder fatigue among participants? What methods can be used to incentivize or encourage panelists before or during the Delphi project?

Responder fatigue is a known issue, and several suggestions were made to mitigate it: limit the number of questions, minimize open-ended questions, avoid redundancy, focus on multiple-choice questions early, provide regular progress updates, send thank-you messages, inform panelists of the expected workload before the study begins, allow for feedback, share results, and remind participants of potential coauthorship opportunities.