BMJ Global Health

A Delphi study to assess the effect of changes in language between the first and second editions of the WHO's Joint External Evaluation

Danique R Gigger ,[1] Jonna Messina Mosoff,[1] Meredith Pinto,[2] Dawn Mapatano,[1] Michael Mahar,[1] Anja Minnick[1]

To cite: Gigger DR, Mosoff JM, Pinto M, *et al*. A Delphi study to assess the effect of changes in language between the first and second editions of the WHO's Joint External Evaluation. *BMJ Glob Health* 2024;**9**:e013954. doi:10.1136/bmjgh-2023-013954

Handling editor Vijay Kumar Chattu

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjgh-2023-013954).

Received 13 September 2023
Accepted 29 March 2024

[1]Centers for Disease Control and Prevention, Atlanta, Georgia, USA
[2]RTI International, Research Triangle Park, North Carolina, USA

Correspondence to
Danique R Gigger;
ofn4@cdc.gov

ABSTRACT

Introduction Countries use the WHO Joint External Evaluation (JEE) tool—part of the WHO International Health Regulations (2005) Monitoring and Evaluation Framework—for voluntary evaluation of global health security (GHS) capacities. After releasing the JEE first edition (E1) in 2016, WHO released the JEE second edition (E2) in 2018 with language changes to multiple indicators and associated capacity levels. To understand the effect of language changes on countries' ability to meet requirements in each edition, we conducted a Delphi study—a method where a panel of experts reach consensus on a topic through iterative, anonymous surveys—to solicit feedback from 40+ GHS experts with expertise in one or more of the 19 JEE technical areas.

Methods We asked experts first to compare the language changes for each capacity level within each indicator and identify how these changes affected the indicator overall; then to assess the ability of a country to achieve the same capacity level using E2 as compared with E1 using a Likert-style score (1–5), where '1' was 'significantly easier' and '5' was 'significantly harder'; and last to provide a qualitative justification for score selections. We analysed the medians and IQR of responses to determine where experts reached consensus.

Results Results demonstrate that 14 indicators and 49 capacity levels would be harder to achieve in E2.

Conclusion Findings underscore the importance of considering how language alterations impact how the JEE measures GHS capacity and the feasibility of using the JEE to monitor changes in capacity over time.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ The WHO has published three editions of the Joint External Evaluation (JEE) (the first edition (E1) in 2016, the second (E2) in 2018 and the third (E3) in 2022), which countries voluntarily use to measure global health security (GHS) capacities to prevent, detect and respond to infectious disease threats.

WHAT THIS STUDY ADDS

⇒ We conducted a Delphi study—an established technique used across disciplines to solicit anonymous input from a group of experts in an iterative and structured environment—with 40+ GHS experts in 2021 to determine how language changes in the JEE indicators' scoring rubric (ie, requirements denoted by capacity levels) between E1 and E2 could affect a country's ability to meet requirements across 19 technical areas.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Results demonstrate that language changes make requirements for at least 14 indicators and 49 capacity levels more difficult to achieve in E2 compared with E1, which could result in lower scores due to language changes rather than true changes in capacity.

⇒ Interested parties may refer to countries' JEE scores from E1 and later editions to assess progress toward compliance with International Health Regulations (2005); however, parties should be aware of language changes in the scoring rubric between editions and how these changes may affect how monitoring capacities are perceived and measured over time.

INTRODUCTION

The WHO's International Health Regulations (IHR 2005) outline key capacities that countries need to prevent, detect and provide a public health response to the international spread of disease without unduly interfering with international traffic and trade.[1] IHR (2005) intended for States Parties to be compliant by 2012; in 2012, however, many States Parties requested a 2-year extension for compliance[2] and, by 2014, most States Parties had not indicated meeting core capacity standards.[3] To address this issue, the Review Committee on Second Extensions recommended that WHO develop options to 'move from exclusive self-evaluation to approaches

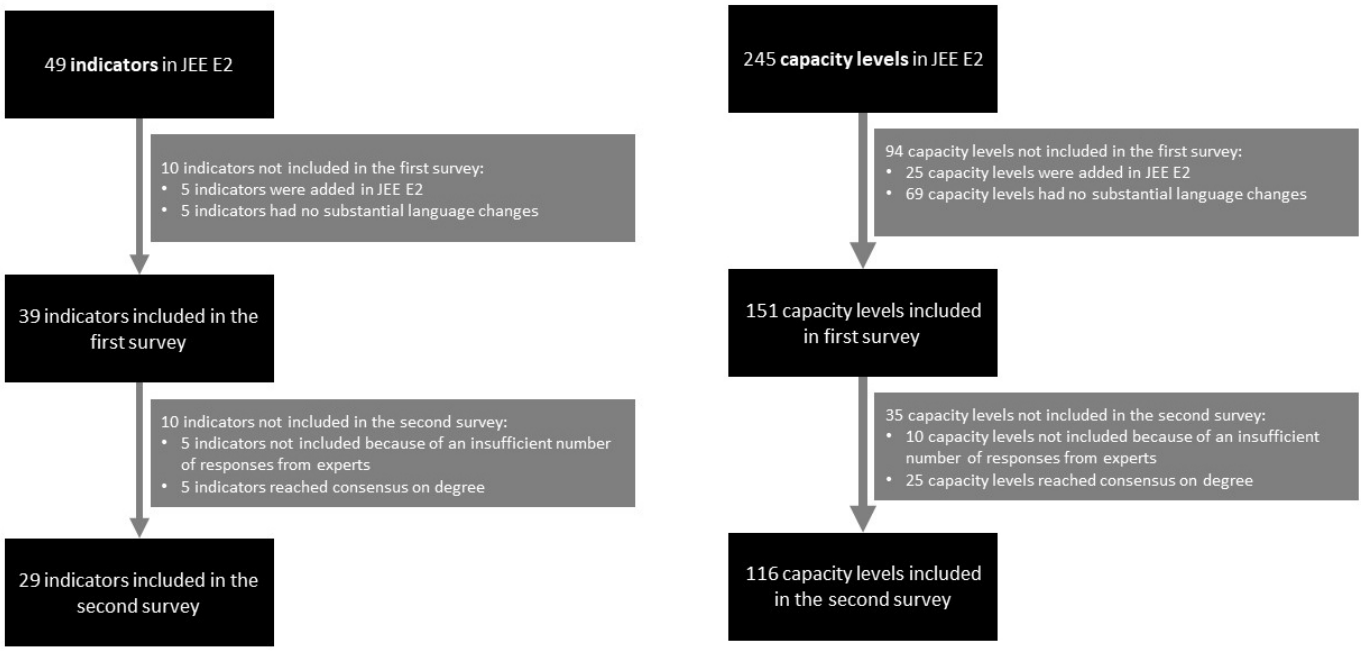

Figure 1 Joint External Evaluation (JEE) indicators and capacity levels asked in each iterative survey.

that combine self-evaluation, peer review, and voluntary external evaluations'.[3]

In response, WHO published the first edition (E1) of the Joint External Evaluation (JEE) in English as part of the IHR Monitoring and Evaluation Framework (MEF) in 2016.[4] As of December 2023, 122 countries have completed a JEE.[5] The JEE is a voluntary and multisectoral assessment, where external evaluators collaborate with a country to assess its capacity to manage infectious disease (ID) outbreaks and other health hazards. The JEE has 19 sections called 'technical areas'. Each technical area has subsections called indicators, and each indicator contains a scoring rubric denoted by five capacity levels that contain qualitative descriptions of specific requirements for each level. Countries receive a Likert-style score of 1 to 5, which reflects one of the five capacity levels for indicators, with '1' defined as 'No Capacity' and '5' as 'Sustainable Capacity'.

The JEE forms the basis for many countries' efforts to build global health security (GHS) capacity for managing ID outbreaks and other public health emergencies. Many countries track their JEE scores and the related IHR MEF State Party Self-Assessment Annual Reporting Tool (SPAR)[6] to gauge their progress towards IHR compliance. Additionally, 70+ countries are part of the Global Health Security Agenda (GHSA), which aims to prevent ID threats globally.[7] The GHSA has a target that by 2024 'more than 100 countries…will strengthen their capacities and demonstrate improvements in at least five technical areas to a level of 'Demonstrated Capacity' or comparable level, as measured by relevant health security assessments, such as those conducted within the WHO IHR MEF'.[7]

WHO released the JEE second edition (E2) in 2018, revising many indicators to incorporate lessons learnt from E1 implementation. The overall number of indicators increased from 48 to 49. In June 2022, WHO revised E2 and released the third edition (E3) of the JEE based on responses to public health emergencies such as Ebola and the COVID-19 pandemic. Since WHO recommends that countries complete the JEE every 5 years, many countries that conducted a JEE using E1 are due to repeat the assessment soon, and more than 20 countries have already completed a second JEE using other editions.[5]

As the JEE continues to change, there is a lack of shared understanding of how changes in the tool may affect how it measures GHS capacities. Without this knowledge, countries face the challenge of interpreting their GHS capacity at different points in time based on their scores across editions and ascertaining whether score changes reflect changes in capacity or changes to the tool's measurement criteria or both. From July to November 2021, we conducted a study using the Delphi method[8] to gain consensus from GHS experts on how language changes from E1 to E2 affect a country's ability to meet indicator and capacity-level requirements in both editions.

METHODS

We conducted two rounds of surveys using the Delphi method, which establishes reliable consensus through an iterative sequence of surveys combined with controlled feedback that allows for anonymity and interaction with other experts' opinions.[8] The Delphi method equally weighs participants' views by collecting opinions anonymously. This anonymity reduces biases that could arise from individuals with recognised status or power influencing majority opinion. The Delphi method allowed us to streamline participation from GHS experts because

Table 1 Characteristics of GHS experts participating on the Delphi panel (N=46)

Participants	n (%)
Experts with over 10 years of experience working in public health	32 (70)
Experts with over 20 years of experience working in public health	14 (30)
Experts who participated in writing E1	11 (24)
Experts involved with JEE revisions	39 (85)
Experts who participated as evaluators or observers for a JEE	21 (46)

GHS, global health security; JEE, Joint External Evaluation.

the process gives participants the flexibility to complete surveys asynchronously, which was critical with competing priorities during the COVID-19 pandemic.

Participant selection

We used purposive and snowball sampling to identify a GHS expert panel from across the US Centers for Disease Control and Prevention, other US government agencies, international ministries of health and public health agencies. We defined experts as individuals working professionally in English who contributed technical expertise in writing, revising and/or conducting a JEE. There is no standard number of participants in a Delphi panel nor a standard number of panels so we considered how questions on 49 JEE indicators and the COVID-19 pandemic would influence receiving responses to select our panel size.[9] To limit the data collection burden during the COVID-19 pandemic—which sought ongoing technical input from a finite pool of experts—we chose to convene a small panel consisting of at least five experts for each of the JEE technical areas during each survey round. Some experts served on multiple panels because of their expertise across several JEE technical areas.

Survey design

We reviewed the English versions of E1 and E2 to identify indicators and capacity levels with substantial language changes between editions. We defined a substantial change as any change in language that affected the meaning of the indicator or its capacity levels. For example, we did not consider the addition of 'the' in E2 as substantial, but we did consider a change from 'or' to 'and' as substantial. We did not include indicators added in E2 or removed from E1. Of the 49 indicators and 245 associated capacity levels in E2, we identified 39 indicators and 151 capacity levels with substantial language changes (figure 1).

We designed and sent two rounds of surveys to our expert panel using SurveyMonkey between July and November 2021. In both rounds, we included the JEE scoring rubric with indicators and associated capacity levels from E1 and E2 side-by-side (online supplemental appendix). For each language change in an indicator or capacity level, we asked experts to consider a scenario where they were conducting a JEE in a country using both editions at the same time and, using a Likert scale, indicate how requirements in the E2 rubric would affect a country's ability to achieve the same E1 capacity level. The Likert scale provided five options: (1) significantly easier, (2) slightly easier, (3) no change, (4) slightly harder and (5) significantly harder to meet the requirements in E2. We then asked experts to provide a narrative justification for their selection.

In the second round, we asked participants to review and consider justifications provided by experts in the first round and, using the same Likert scale, indicate how the requirements in the E2 rubric affect a country's ability to achieve a capacity level. We only included the Likert responses selected in the first round for the second round (eg, if no participants selected '3' in the first round, we only included options 1, 2, 4 and 5 in the second round). We did not include indicators and capacity levels in the second round if experts reached a consensus or if we did not have at least five experts respond to the survey in the first round.

Data analysis

We analysed responses to the Likert scale in Microsoft Excel using the inclusive median and IQR.[10] We created four mutually exclusive categorisations for the possible distributions of response data. In increasing order of consensus on how the language changes between E1 and E2 affect a country's ability to meet requirements: *no consensus* represents the distribution where experts did not agree at all (ie, experts selected a mix of scores 1–5); *change* describes where experts agreed that language updates in E2 change the requirements but did not

Table 2 Number and per cent of evaluated indicators (N=34) and capacity levels (N=141) where experts reached consensus on language changes from E1 to E2

	n (%)	
Consensus category	Indicators N=34	Capacity levels N=141
Consensus on degree (ie, IQR=0; score of 1, 2, 3, 4, OR 5)	14 (41)	54 (38)
Consensus on direction (ie, scores of '1 and 2' OR '4 and 5')	5 (15)	26 (19)
Consensus on change (ie, scores of '1 and/or 2' AND '4 and/or 5')	6 (18)	14 (10)
No consensus (ie, selection of scores 1–5)	9 (26)	47 (33)

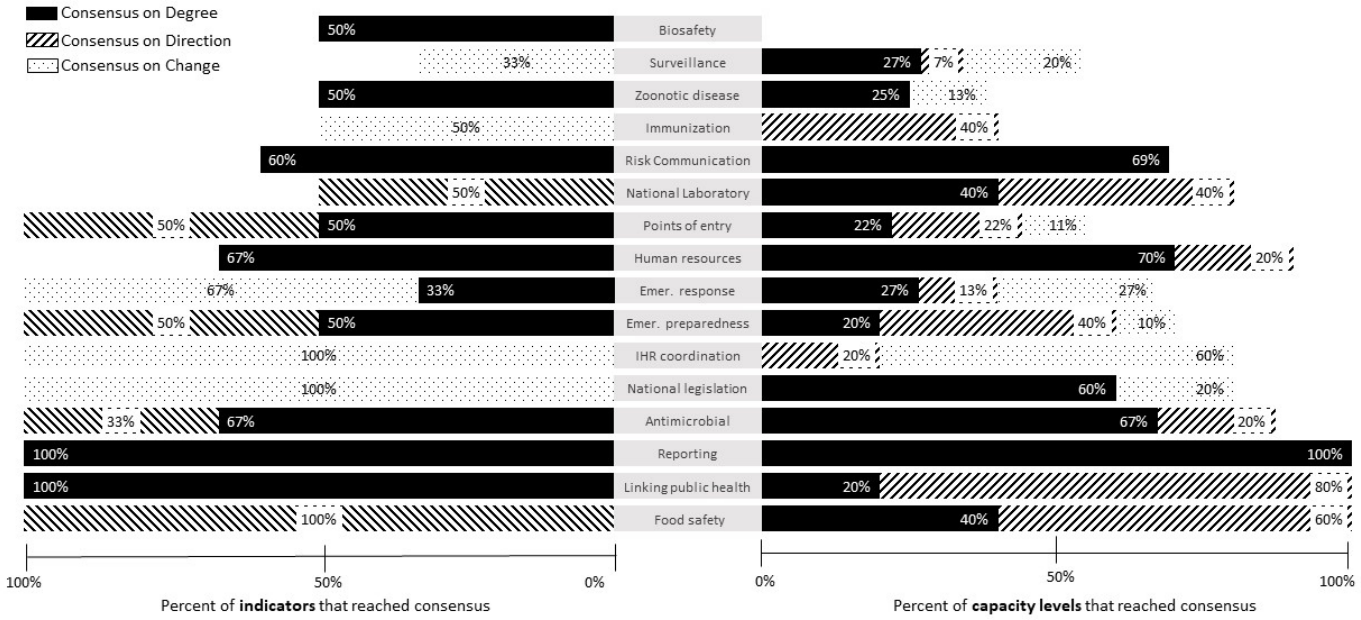

Figure 2 Per cent of evaluated indicators (N=34) and capacity levels (N=141) where experts reached consensus on degree, direction and change by technical area.

agree if updates make requirements easier or harder (ie, experts selected a mix of scores of '1 and/or 2' AND '4 and/or 5'); *direction* describes where experts agreed that language changes in E2 make the requirements easier or harder but did not agree if updates make requirements significantly or slightly easier or harder (ie, experts selected scores of '1 and 2' OR '4 and 5'); and *degree* describes an IQR of 0 where all experts selected the same score of 1, 2, 3, 4 OR 5.

Patient and public involvement

Patients and the public were not directly involved in the design of this study or the formulation of research questions and outcome measures.

RESULTS

Participants

We contacted GHS experts via email, and 41 individuals agreed to participate in round one; we received responses from 66% of experts (27/41). We recruited 5 new experts for round two; 46 agreed to participate, and we received responses from 89% of experts (41/46). We describe characteristics of the experts who responded to at least one survey in table 1. Seventy per cent of respondents (32/46) indicated that they have worked in public health for over 10 years, with 44% (14/32) having more than 20 years of experience. Twenty-four per cent of respondents (11/46) contributed to the text for E1, and 85% (39/46) contributed to revisions. Forty-six per cent (21/46) reported serving as an evaluator or observer for a JEE.

Extent of language changes to JEE E2 indicators and capacity levels

Of the 49 indicators and 245 capacity levels in E2, 10% (5/49) of indicators and 10% (25/245) of capacity levels were added (figure 1); 10% (5/49) of indicators and 28% (69/245) of capacity levels did not contain substantial language changes. In round one, we included the remaining 80% (39/49) of E2 indicators, which contained substantial changes from E1 in at least one capacity level, and 62% (151/245) of the individual capacity levels, which contained substantial changes.

Categories of consensus

In round one, experts reached a consensus on *degree* for 13% of included indicators (5/39) and 17% of included capacity levels (25/151). We did not receive at least five expert responses in the technical areas of 'Medical countermeasures and personnel deployment', 'Chemical events' and 'Radiation emergencies' so we excluded 13% of indicators (5/39) and 7% of capacity levels (10/151) in round two and did not evaluate the level of consensus reached. Overall, experts reached some level of consensus (on *degree*, *direction* or *change*) for 74% of evaluated indicators (25/34) and 67% of evaluated capacity levels (94/141) in E2 (table 2).

No consensus

For 26% of evaluated indicators (9/34) and 33% of evaluated capacity levels (47/141), experts did not reach consensus on how language changes affect a country's ability to meet requirements. Experts reached the least consensus in the 'Biosafety and biosecurity' technical area (figure 2), which contains substantial changes in two indicators, P.6.1 and P.6.2, and nine corresponding capacity levels. Although experts stated that requirements would be the same for indicator P.6.1 (figure 3), they did not reach a consensus on indicator P.6.2 or any capacity level within it. Qualitative responses contextualise this lack of consensus; for example, P.6.2 capacity

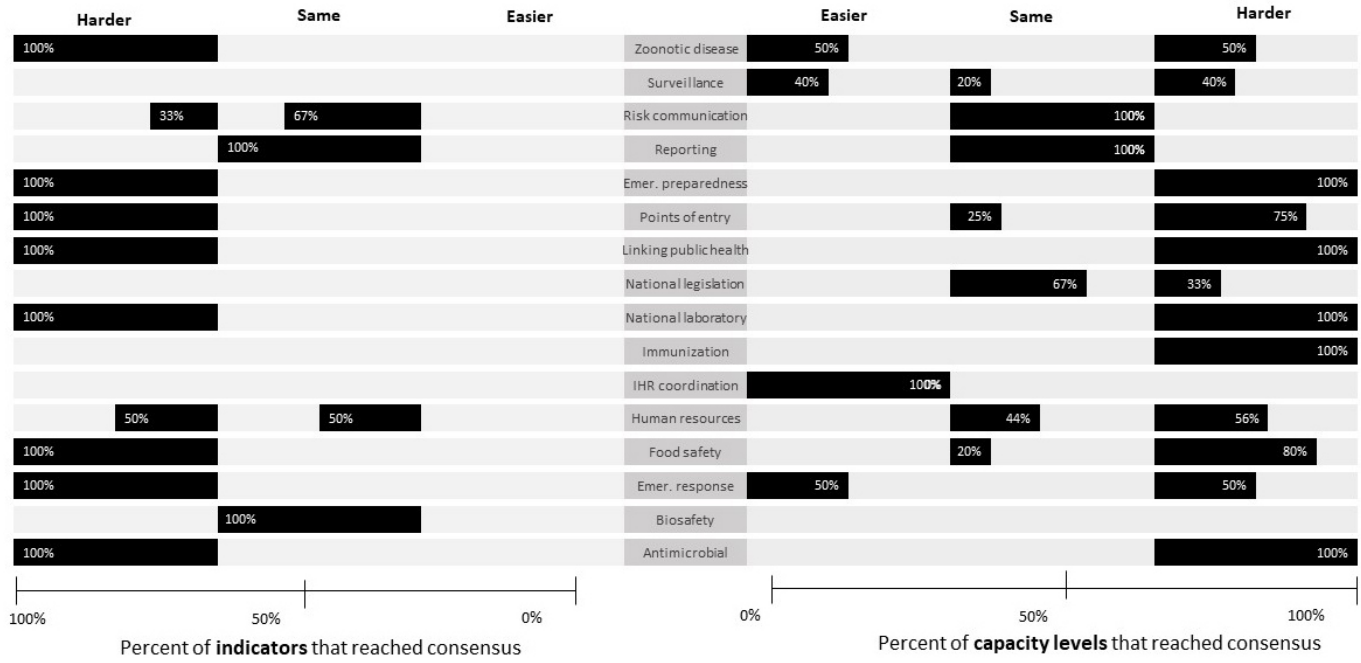

Figure 3 Per cent of indicators (n=19) and capacity levels (n=80) that reached consensus on direction or degree of change with easier, harder, or the same requirements in E2 by technical area.

level 3 responses ranged from 'the revised language uses 'specific' and 'training proportionate to the risks' which slightly reduces the burden of evidence needed for a country to achieve this score' (selection of 2: slightly easier) to the 'changes average out' (selection of 3: no change) to '…the specificity here would make it slightly harder to achieve if only because it limits what is acceptable…' (selection of 4: slightly harder).

Consensus on change

For 18% of evaluated indicators (6/34) and 10% of evaluated capacity levels (14/141), experts reached consensus on change (table 2). Change was the most common distribution of responses for the 'IHR coordination, communication and advocacy' technical area, which contains indicator P.2.1. Experts reached consensus on change for indicator P.2.1 and 60% (3/5) of its capacity levels (figure 2). Providing justifications for their selections, experts indicated that meeting requirements for P.2.1 capacity level 2, for example, is easier in E2 because the language 'is more clear and will allow for better responses' (selection of 1: significantly easier), and it 'broadens the option to meet the requirements' (selection of 2: slightly easier). Other experts indicated that meeting requirements would 'now be harder to achieve because coordination would have to occur within a ministry in addition to between different ministries' (selection of 4: slightly harder), and this addition 'will require more resources [and] while it may serve to clarify coordination activities, the requirement likely complicates the scoring' (selection of 4: slightly harder).

Consensus on direction

For 15% of evaluated indicators (5/34) and 19% of evaluated capacity levels (26/141), experts reached consensus on direction (table 2). The 'Food safety' technical area—which contains indicator P.5.2 and five capacity levels with substantial changes—had the highest percentage of consensus on direction; experts reached consensus on direction for indicator P.5.2 and 60% (3/5) of its capacity levels (figure 2). Providing justifications for P.5.2 capacity level 2, one expert reported that 'the inclusion of a full plan [versus] just focal points would make the updated category slightly harder' (selection of 4: slightly harder); another expert stated that 'having a national food safety plan including triggers and identified [human resources] is more difficult than having identified focal persons, as this would require x-ministry concurrence' (selection of 5: significantly harder).

Consensus on degree

For 41% of evaluated indicators (14/34) and 38% of evaluated capacity levels (54/141) experts reached consensus on degree (table 2). The 'Reporting' technical area had the highest percentage of consensus on degree. Experts reached consensus on degree for the one indicator (D.3.1) and four associated capacity levels with substantial changes, indicating that the requirements are the same as E1 (figure 3). One expert noted for D.3.1 capacity level 2, for example, that the E2 language 'is nearly identical', while another indicated that they did not 'interpret the change as adding any difficulty'; a third stated it 'would not impact the ability to achieve the level' (selections of 3: same).

Overall ability to meet indicator and capacity level requirements in JEE E2

Experts reached consensus on the direction or degree of language changes for 56% of evaluated indicators (19/34) and 57% of evaluated capacity levels (80/141) (table 2). Of those that reached consensus on direction or degree, experts indicated that 74% of indicators (14/19) and 61% of capacity levels (49/80) would be more difficult to achieve in E2 compared with E1; requirements for 26% of indicators (5/19) and 30% of capacity levels (24/80) would remain the same; 0% of indicators (0/39) and 9% of capacity levels (7/80) would be easier to achieve. The 'Emergency response operations'; 'IHR coordination, communication, and advocacy'; 'Surveillance'; and 'Zoonotic disease' technical areas were the only technical areas where experts concluded that at least one capacity level would be easier to achieve in E2 (figure 3).

DISCUSSION

In this Delphi study, we examined the language changes between indicators and capacity levels in E1 and E2 to understand how changes in language may influence how country capacities are perceived, measured and reported. For evaluated indicators and capacity levels where experts reached consensus on the direction or degree of the language change, experts indicated that it would be harder for countries to meet requirements using E2 compared with E1 for the majority—74% of indicators (14/19) and 61% of capacity levels (49/80). This means that experts agreed it is harder to meet requirements for at least 29% of all indicators (14/49) and 20% of all capacity levels (49/245) in E2.

To facilitate accurate interpretation of results using qualitative measurement tools, a tool must be valid and reliable. Changing language in a measurement tool could influence the ways in which people interpret how the tool measures its outcome of interest (validity); an observed score by different raters, at different times, or measurement errors in the tool itself could indicate that changes in scores may not reflect true change (reliability).[10–13] For JEE technical areas where experts did not reach consensus on the direction or degree of change, experts' qualitative justifications indicated a range of interpretations on how the change would affect a country's ability to meet requirements. Differences in interpretations have broad implications for the JEE. Authors of the JEE may intend only to clarify language and have no intention to make requirements harder or easier in capacity levels; however, when different pools of experts conduct JEEs in different countries using 'clarified' language in a newer edition without clear definitions for key terms, they may score countries differently based on their interpretation of requirements within a specific capacity level, indicator or technical area. Given that changing a measurement tool could impact its validity and reliability, an increase or decrease in a country's capacity level using the scoring rubric in E2 versus E1 may not accurately reflect a change in the country's capacity and may indicate a change in how the tool itself measures capacity.

When there are substantial changes between editions, it may be beneficial to include guidance that identifies when language changes are meant to provide context or clarity on what was written in a previous edition and are not intended to add or eliminate requirements. Overall, it may be helpful to consider including definitions, explicitly describing requirements and standardising key terms used throughout indicators and capacity levels in the JEE.

This study has several limitations, most notably in the participant retention rate. We conducted this study during an active response to the COVID-19 pandemic globally, and individuals who met the criteria for experts were also in a pool of individuals mobilised to respond to COVID-19. There may be differences in the views of individuals who were invited and agreed to participate compared with individuals who were invited but did not participate; however, we have little information about the views of identified experts who chose not to participate. Because the Delphi technique relies on a panel of experts completing multiple rounds of surveys, both identification of participants who met the inclusion criteria and the retainment of those individuals between survey rounds was a challenge.

Another limitation is our inability to obtain consensus from experts on the direction or degree of change for 44% of indicators (15/34) and 43% of capacity levels (61/141). The Delphi method entails continued deployment of surveys until experts reach consensus, based on iterative qualitative justifications. Although consensus may have been reached through additional survey rounds, many participants did not change their quantitative selection—even after seeing other experts' justifications—and the drop off in participant retention indicated that subsequent surveys would likely not produce sufficient responses per technical area. Participants also frequently provided minimal information in their qualitative responses, likely because of survey fatigue with the number of items asked and lack of time in the pandemic response context. This drop off in response rate is consistent for Delphi studies with a large number of items included in each survey.[14] Finally, although all participants work professionally in English, we did not identify if they consider English as a first language, and it is unknown if that may have impacted their interpretation of requirements and ability to reach consensus.

Overall, this Delphi study has identified that language changes between JEE editions may substantially change the interpretation of requirements in the JEE and subsequent scoring during country assessments. This study has broad implications for any analyses using results from, or based on, more than one JEE edition. For example, several globally recognised indices based on the JEE, including the SPAR,[6] the Global Health Security Index from the Nuclear Threat Initiative and Johns Hopkins Center for Health Security,[15] and the ReadyScore from

Resolve to Save Lives[16] may be impacted by changes between JEE editions.

The JEE is not designed for intercountry comparisons; it is a tool created to support WHO member countries in establishing a quantitative baseline assessment of IHR core capacities to then measure their progress over time.[17] However, the COVID-19 pandemic highlighted gaps in the JEE's ability to predict GHS readiness as JEE scores were not predictive of countries' COVID-19 response performance.[18 19] Given the updates in E2 and now in E3 based on lessons learnt from the COVID-19 response, there may be a misconception of where to focus efforts if countries use E1 to identify priority areas for improvement (ie, indicators with lower scores). Additionally, a capacity level of '4' or 'Demonstrated Capacity' in E2 or E3 may have substantially different requirements than the previous edition, which may affect how country capacities for added or eliminated requirements are perceived and measured. Many countries that completed E1 are due to repeat the assessment soon, and more than 20 countries have already completed a second JEE using other editions. Interested parties may refer to countries' scores from E1 and later editions in their assessment of progress towards IHR compliance and current initiatives to improve GHS capacities such as the GHSA 2024 target. However, it is important to be aware of language changes in the scoring rubric between editions and how these changes may affect monitoring capacities over time.

At the time of this study, WHO had not yet published E3; however, now that countries are conducting JEEs using E3, further research is needed to understand the effect of language changes from E1 and E2 to E3.

Acknowledgements The authors gratefully acknowledge all participants who provided survey data for this Delphi study.

Contributors All authors planned, designed and conducted the Delphi study. DRG and JMM analysed the data and wrote the first draft. MP and AM edited the draft. All the authors reviewed and approved the final version. DRG and JMM accept full responsibility of the overall content as the guarantors.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD

Danique R Gigger http://orcid.org/0000-0003-1279-164X

REFERENCES

1 International health regulations 2005. World Health Organization; 2005.
2 Fischer JE, Katz R. Moving forward to 2014: global IHR (2005) implementation. *Biosecur Bioterror* 2013;11:153–6.
3 World Health Organization. Implementation of the International Health Regulations (2005): report of the Review Committee on Second Extensions for establishing national public health capacities and on IHR implementation: report by the Director-General. World Health Assembly, 2015.
4 World Health Organization. IHR Monitoring and Evaluation Framework, Available: https://extranet.who.int/sph/ihr-monitoring-evaluation [Accessed 25 Jun 2022].
5 World Health Organization. IHR (2005) Monitoring and Evaluation Framework. Joint External Evaluation (JEE) tool.. 2023. Available: https://extranet.who.int/sph/jee [Accessed 25 Jul 2023].
6 International Health Regulations 2005. World Health Organization; States Parties self-assessment annual reporting tool, 2018. Available: https://www.who.int/publications/i/item/WHO-WHE-CPI-2018-16
7 Global Health Security Agenda. ghsagenda.org. A partnership against global health threats 2023, Available: https://globalhealthsecurityagenda.org/
8 Linstone HA, Turoff M. The Delphi Method. MA: Addison-Wesley Reading, 1975.
9 R Avella J. Delphi panels: research design, procedures, advantages, and challenges. *IJDS* 2016;11:305–21.
10 Stevens SS. On the theory of scales of measurement. *Science* 1946;103:677–80.
11 Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. *Am J Health Syst Pharm* 2008;65:2276–84.
12 Scholtes VA, Terwee CB, Poolman RW. What makes a measurement instrument valid and reliable *Injury* 2011;42:236–40.
13 Sullivan GM. A primer on the validity of assessment instruments. *J Grad Med Educ* 2011;3:119–20.
14 Gargon E, Crew R, Burnside G, *et al*. Higher number of items associated with significantly lower response rates in COS Delphi surveys. *J Clin Epidemiol* 2019;108:110–20.
15 Cameron E, Nuzzo J, Bell J. Global health security index: building collective action and accountability, 2019., Available: https://www.jstor.org/stable/resrep44061.1 [Accessed 28 Jul 2023].
16 Resolve to save lives: prevent pandemics: Resolve to Save Lives, 2022. Available: https://resolvetosavelives.org/prevent-epidemics [Accessed 25 Jul 2023].
17 Traore T, Shanks S, Haider N, *et al*. How prepared is the world? identifying weaknesses in existing assessment frameworks for global health security through a one health approach. *The Lancet* 2023;401:673–87.
18 Haider N, Yavlinsky A, Chang Y-M, *et al*. The global health security index and joint external evaluation score for health preparedness are not correlated with countries' COVID-19 detection response time and mortality outcome epidemiology & infection. *Epidemiol Infect* 2020;148:e210.
19 Nguyen L, Brown MS, Couture A, *et al*. Global health security preparedness and response: an analysis of the relationship between joint external evaluation scores and COVID-19 response performance. *BMJ Open* 2021;11:e050052.

APPENDIX

Below is an illustrative example of the Delphi study round 1 and round 2 questions.

IHR Coordination, communication and advocacy: Indicator *P.2.1*

JEE Edition/ Indicator Number	E1: P.2.1	E2: P.2.1
Indicator	A functional mechanism is established for the coordination and integration of relevant sectors in the implementation of IHR.	
Capacity level 1	Coordination mechanism between relevant ministries is not in place	Coordination mechanism **within and** between relevant ministries, **including government agencies**, is not in place
Capacity level 2	Coordination mechanism between relevant ministries is in place. National Standard Operating Procedures (SOPs) or equivalent exists for the coordination between IHR NFP and relevant sectors	Coordination mechanism **within and** between relevant ministries is in place. National SOPs or equivalent exists for coordination between the National IHR Focal Point and relevant sectors
Capacity level 3	A multisectoral, multidisciplinary body, committee, or taskforce addressing IHR requirements **on surveillance and response** for public health emergencies of national and international concern is in place and participated in latest event	A multisectoral, multidisciplinary body, committee, or taskforce addressing IHR requirements for public health emergencies of national and international concern is in place and has participated in the latest event **or simulation exercise**
Capacity level 4	Multisectoral and multidisciplinary coordination and communication mechanisms are tested and updated regularly through exercises **or through the occurrence of an actual event.** Action plan developed to incorporate lessons learnt **of** multisectoral and multidisciplinary coordination and communication mechanisms	Multisectoral and multidisciplinary coordination and communication mechanisms **are in place**, tested, and updated regularly through exercises **or after-action reviews based on the occurrence of an actual event.** Action plan developed to incorporate lessons learnt **from** multisectoral and multidisciplinary coordination and communication mechanisms
Capacity level 5	Annual updates on the status of IHR implementation to stakeholders across all relevant sectors **conducted**	Annual updates on the status of IHR implementation to stakeholders **(including WHO and other IHR States Parties across all relevant sectors) are conducted and confirm the efficiency and effectiveness of the coordination, communication and advocacy arrangements** across all relevant sectors

P.2.1 Capacity Level 1 (repeat questions for each capacity level):

P.2.1	Capacity Level Definition JEE E1	Capacity Level Definition JEE E2
Capacity Level 1	Coordination mechanism between relevant ministries is not in place	Coordination mechanism within and between relevant ministries, including government agencies, is not in place

Question 1: If you were conducting a JEE in a country using these two editions at the same time, to what extent would this change in language in the second edition affect a country achieving this capacity level? (select one)

1. ***Significantly easier*** to meet the requirements of the capacity level in the second edition
2. ***Slightly easier*** to meet the requirements of the capacity level in the second edition
3. ***No change*** in requirements to meet capacity levels (the score would be the same) in the second edition
4. ***Slightly harder*** to meet the requirements of the capacity level in the second edition
5. ***Significantly harder*** to meet the requirements of the capacity level in the second edition

Question 2: Please provide a justification for your selection.

Overall Indicator P.2.1:

Question 1: Overall, how do the changes in all capacity levels 1-5 from the first edition to the second edition affect the scope of this whole indicator? (select one)

1. ***Significantly easier*** to meet the requirements of the capacity levels in the second edition
2. ***Slightly easier*** to meet the requirements of the capacity levels in the second edition
3. ***No change*** in requirements to meet capacity levels (the score would be the same) in the second edition
4. ***Slightly harder*** to meet the requirements of the capacity levels in the second edition
5. ***Significantly harder*** to meet the requirements of the capacity levels in the second edition

Question 2: Please provide a justification for your selection.