Delphi with feedback of rationales: How large can a Delphi group be such that participants are not overloaded, de-motivated, or disengaged?

Ian Belton[a1], George Wright[1], Aileen Sissons[1], Fergus Bolger[1], Megan M. Crawford[2], Iain Hamlin[1], Courtney Taylor Browne Lūka[3], Alexandrina Vasilichi[4]

1. Department of Management Science, Strathclyde Business School, University of Strathclyde, 199 Cathedral Street, Glasgow G4 0QU
2. The Business School, Edinburgh Napier University, Unit 4, 10 Bankhead Terrace, Edinburgh EH11 4DY
3. Department of Psychology, University of Glasgow, Glasgow G12 8QQ
4. Department of Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB

[a] Corresponding author: ian.belton@strath.ac.uk

Abstract

In this paper, we investigate the effect of Delphi group size and opinion diversity on group members' information load as well as on their overall experience of the Delphi process - in terms of task involvement (enjoyment and interest) and in terms of group sway (the influence and helpfulness of others' rationales). For Delphi applications involving the exchange of rationales between participants, we found no evidence that group sizes of up to 19 participants cause information overload or de-motivation and disengagement of participants.

Keywords: Delphi, group processes, information load, information overload

Acknowledgement: This research is based on work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), under Contract [2017-16122000003]. The views and conclusions contained herein are

Introduction

Relative to individuals, groups potentially bring a variety of perspectives, experience and knowledge that should help to improve the accuracy of judgmental forecasting but this advantage may be reduced by several cognitive and social biases – such as anchoring (Tversky & Kahneman, 1974) and groupthink (Janis, 1972), respectively – that are manifest in (and amplified by) freely interacting groups. The Delphi technique was developed in the 1950s by the RAND organization as a structured group-based judgmental forecasting method for the defence sector (Dalkey & Helmer, 1963) but since then has been applied in a wide variety of contexts with the aim of improving the outcomes of group judgment (e.g. Linstone & Turoff, 1975; Rowe & Wright, 2011).

Delphi achieves its advantages over individual judgment and interacting groups by first surveying the *anonymous* opinions of several individual experts, who are thus able to make judgments free of any anchors provided by other group members or fear of group censure. Group members then receive feedback – usually edited and summarised by a facilitator – regarding the opinions of other experts in their nominal group. Next, each individual expert is invited to revise his or her opinion and then these revised opinions are again collated by the facilitator to be fed back in subsequent rounds of revision - or released as the final outcome if a sufficient degree of consensus (or a stable dissensus) has been reached. Between-round feedback typically includes summary statistics describing the group's responses and participants can also be asked to provide written rationales in support of their judgments (Bolger & Wright, 2011; Meijering & Tobi, 2016). Use of written rationales as feedback is now much more common that in earlier decades (Belton et al, 2019), thanks in part to the availability of online tools which allow the collation and feedback of multiple rationales to be managed more easily (see Aengenheyster et al., 2017 for a review).

Participation in a Delphi study requires a significant time commitment since panel members need to provide input on two or more occasions - sometimes separated by several weeks of elapsed time (Boulkedid et al., 2011; Rowe & Wright, 2011). As a result, keeping participants engaged can be difficult and minimising drop-out rates is an important consideration (Toma & Picioreanu, 2016; Turnbull et al., 2018). The quality of a Delphi study's findings depends on as high a proportion of the initial panel as possible completing the second and any subsequent rounds and contributing fully throughout (Hasson et al., 2000; Goluchowicz & Blind, 2011). Where the attrition rate is high, those with dissenting views are more likely to drop-out (Humphrey-Murto & de Wit, 2019) and the subsequent Delphi yield – now produced by a smaller and perhaps insufficiently diverse panel may be viewed as less credible by external stakeholders in the process or those in the wider community (Cairns, Wright, Fairbrother & Phillips, 2017; Landeta, 2006). In addition, if an individual has a negative experience as part of a Delphi process, he/she may be reluctant to take part in subsequent studies. Retaining a pool of experts for future projects can be an important consideration, for example where surveys are repeated over several years (e.g. Airaksinen, Halinen & Linturi, 2017; KISTEP, 2005, 2017; NISTEP, 2009, 2015, 2019) – especially since identifying and recruiting high-quality experts in the first place can be a time-consuming and challenging process (Bolger, 2018). Consequently, it is clearly important to understand the factors that may help determine whether a particular Delphi survey is likely to be well-received by its targeted participants, or not.

In terms of the number of expert panellists that are required, many ranges have been suggested: 5-20 (Rowe & Wright, 2001), 15-60 (Hasson et al., 2000), no more than 50 (Toma & Picioreanu, 2016; Witkin & Altschuld, 1995), or 15-30 for homogenous Delphi panels (Clayton, 1997) and 5-10 for heterogeneous panels (Delbecq, Van de Ven, & Gustafson, 1975). In a recent review of 63 Delphi studies, de Loë et al. (2016) reported a huge range of panel sizes, from fewer than 10 up to more than 1000. In the field of technology foresight, national-level surveys can involve many thousands of expert panellists, for example in Japan (2900 participants – NISTEP, 2009), The Republic of Korea (5450 participants – Choi & Choi, 2017) and China (nearly 3000 – Li et al., 2017)

In this study, we focus on both the size of a Delphi panel of participants and on the diversity of viewpoints within it. Current advice on these issues (Belton et al, 2019) is varied – perhaps because panel membership size is not crucial when qualitative feedback between rounds is not utilised and the feedback of panellists' viewpoints is simply numerical summaries of opinions. Nevertheless, a commonality in all advice is the imperative to utilise heterogeneous membership of Delphi panels, so that varied viewpoints are received and evaluated by Delphi panellists. Clearly, between-round feedback that involves the exchange and reading of rationales for numerical responses will lengthen panellists' time commitment for each Delphi item considered, whereas exchange of feedback between rounds in terms of medians and ranges of numerical responses will not.

Research on both Cognitive Load Theory (CLT; see e.g. Paas, Tuovinen, Tabbers & Van Gerven, 2003; Sweller, van Merrienboer & Paas, 1998) and information overload (for reviews, see Eppler & Mengis, 2004; Hwang & Lin, 1999) has found that an individual's performance in terms of decision accuracy and/or reasoning quality correlates positively with both the volume and complexity of information he/she receives before making a judgment or decision – up to a certain point. If further information is received beyond that point, the individual's performance will decline rapidly, due to cognitive overload. Studies of online consumer choice tasks suggest that having more information of greater complexity to review also makes individuals more likely to report negative perceptions of the decision process (Griefeneder, Scheibehenne & Kleber, 2010; Park & Jang, 2012), which may in turn lead to future task avoidance (Sthapit, Del Chiappa, Coudounaris & Bjork, 2019). Additionally, research suggest that individuals may change their judgment strategies when faced with information overload, for example by searching for less information (Cook, 1993; Swain & Haka, 2000) or using a simple decision heuristic to relieve the excess load (Agnew & Szykman, 2005; Biggs, Bedard, Gaber & Linsmeier, 1985).

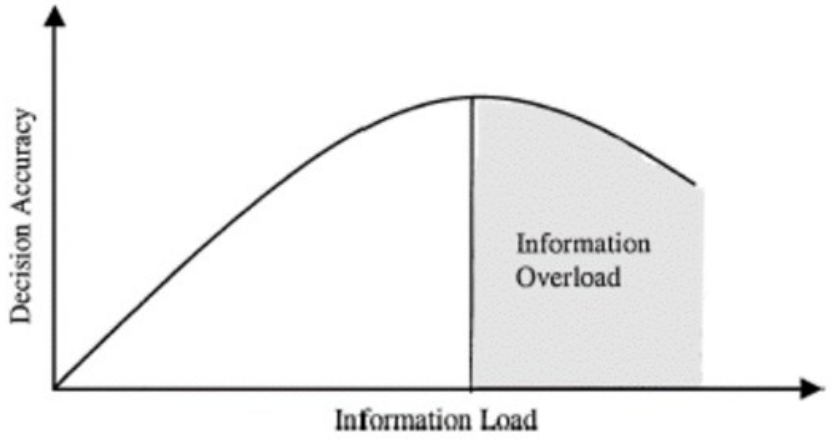


Figure 1. Effect of information overload on decision accuracy. Taken from Eppler and Mengis (2004).

A key question is: how much information is too much? CLT emphasises working memory constraints as determinants of cognitive capacity (Sweller, van Merrienboer & Paas, 1998). Miller (1956) famously estimated that working memory capacity was 7 ± 2 items (c.f. Cowan's 2001 review). But how does this figure translate to real-world information processing tasks? Park and Jang's (2012) tourism choice study compared participants' behaviour across choice sets with 1, 3, 10, 20 and 30 items. Consistent with CLT, participants were increasingly more likely to make a definitive choice between up to 20 alternatives, after which choice likelihood decreased – perhaps because of choice overload. Griefeneder et al. (2010) compared participant satisfaction across sets of consumer choices that differed in number (6, 15 or 30 items) and complexity (1 or 6 attributes of value attached to each choice item – such as size, colour, price, etc). When choice options were simple, i.e., contained few value attributes, there was no difference in satisfaction across choice sets with 6, 15 or 30 alternative items. When options were more complex, satisfaction with the decision reached reduced in a linear fashion as the number of items increased. In practice, the threshold for information overload is likely to be very task-specific.

In a Delphi setting with feedback of qualitative rationales, the issue of whether information provided prior to a judgment or choice is likely to cause a cognitive challenge will depend on a complex interaction between many factors including information volume (number and length of rationales and overall word count), complexity (diversity of views, type of

arguments used within the proffered rationales), and the expertise of the participant in relation to the given task.

Delphi panellists are often required to read through other panellists' written answers and rationales before revising their own judgments and this can be a cognitively demanding task, especially when completing long, multi-part Delphi surveys. In studies where rationales are provided, an increase in Delphi group size will typically mean more information for individual panellists to review and so there is likely to be a point at which the participants' task of reading and evaluating other group members' answers and rationales becomes unmanageable through information overload. In addition, groups that generate a wide range of diverse rationales – while obviously desirable for producing quality Delphi results – will also add to the cognitive burden on their members, since contrasting arguments and perspectives should, ideally, be carefully analysed and evaluated by participating panellists. But, note that a Delphi panellist with a strong interest in a question topic may (i) feel less overloaded than a panellist with a lesser interest, when both are presented with detailed rationales of other panellists, and also (ii) be prepared to spend more time evaluating such rationales. Recent automated methods for filtering or organising rationales such as Dynamic Argumentative Delphi (DAD - Gheorghiu et al., 2017) are subject to similar, potential, cognitive overload issues, since although these methods substantially reduce the number of rationales for panellists to review, there could still be too many for participants to carefully read and understand.

An 'overloaded' Delphi panellist could come away with a relatively negative perception of the experience, be less likely to participate in any further Delphi rounds and be less motivated to engage in any future Delphi studies. Also, the overloaded panellist might change his/her cognitive strategy (consciously or otherwise) to make the task more manageable, for example by only reading some of the answers or rationales, skim-reading the text, or engaging with the material on a more superficial level. Such a change in strategy could diminish the quality of the overall Delphi output – particularly if the effect was widespread across the group. Neither problematic issue would be immediately obvious, if at all, to the Delphi administrator.

There is very little empirical evidence for the effect of group size on members' experiences. Hackman and Vidmar (1970) explored the effect of group size (from two to seven) on

members' experience of various academic tasks. The study found that group size had a strong effect on member's reactions, with dissatisfaction increasing along with size. However, Hackman and Vidmar's study related to interacting groups, which are different in many respects from a Delphi survey. Boje and Murnighan (1982), the only researchers to date to have explored the relationship between Delphi group size and member experience, tested groups of three, seven and eleven on two almanac and two statistical questions. For each question, participants gave an answer and one supporting fact or reason. A follow-up questionnaire revealed that those in groups of three felt a larger group would have been more accurate, while those in groups of eleven felt they were less free to communicate their ideas. Group members' self-reported enjoyment, time sufficiency and rating of the quality of the method used (in general and applied to the given task) did not differ across group sizes. No published study has explored the relationship between the diversity of Delphi panellists' opinions and perceptions of the Delphi experience.

The present study explored the following research question: How do Delphi group size and opinion diversity influence group members' actual and perceived information load, as well as their overall experience of the Delphi process? Our aim was to learn more about how the variables group size and opinion diversity may operate, independently or in combination, to affect Delphi group members' experience of taking part in a survey.

Method

Participants

In total, 282 participants were recruited online using the 'Prolific' recruitment platform.

Thirty participants were recruited for the pre-study. Two hundred and fifty-two participants were recruited for the main study.

Materials

Pre-study

In the pre-study, we elicited rationales linked to binary ("yes/no") judgments that we then utilised for the main study, following a novel method based on the "Simulated Group

Response Paradigm" (SGRP). SGRP is described in detail in Bolger et al., (2020). Pre-study participants were given 10 short-term forecasting questions (see Appendix 1). Questions were written to be topical for the time at which the study was run (September 2018). For each question, participants were asked to answer "yes" or "no" and give three rationales to support their answer. For example, in response to the question "China and the US will resume trade talks designed to de-escalate the trade war", typical rationales included:

- (Answering "yes"): They have to talk. Money is falling off the stock exchanges left right and centre, both countries are huge players in trade, it benefits them both to find common ground.
- (Answering "no"): Internal economics will exert pressure on Trump to keep jobs and services in America, meaning no end to the trade stand-off with China.

Stimuli was presented online using Qualtrics survey software.

Main study

The main study was a 5 x 2 between-subjects design. Two independent variables were manipulated in the study:

1. Delphi group size (5 levels: 7, 10, 13, 16 and 19).
2. Opinion diversity (2 levels: low and high, defined below).

Participants were randomly allocated to one of 10 conditions. Stimuli were presented online using Qualtrics. The order in which participants received the 10 forecasting questions was also randomised to prevent order effects. In each condition participants were given 10 short-term forecasting questions (see Appendix 1). A two-round Delphi process was repeated for each question, as follows:

1. Participants were asked to answer "yes" or "no", give a rationale for their answer, and provide a confidence estimate on a scale from 50 to 100 per cent (an estimate of less than 50 per cent would mean that the participant should have chosen the other answer).
2. Participants were then shown a set of 6, 9, 12, 15 or 18 answers taken from those collected during the pre-study, each with an accompanying rationale. In the low diversity

conditions, each “yes” and each “no” answer was supported by a version of the same rationale (the same argument but worded differently). In the high diversity conditions, each “yes and each “no” answer was supported by three different rationales (repeated multiple times in the 12, 15 and 18 answer conditions). In every case, participants were shown a majority of answers and rationales contrary to their original answer (“no” if they answered “yes” and vice versa), in the ratio 2:1 (4 vs 2, 6 vs 3, 8 vs 4, 10 vs 5 and 12 vs 6 respectively).

3. Next, a participant was shown their original answer and rationale and given an opportunity to revise this, if she/he wished, after reading the other responses. Participants were also able to revise their confidence estimate at this point.

4. Lastly, participants completed the NASA-TLX satisfaction questions (described next) and the self-report group process questions (also described next), and provided basic demographic information.

Two separate attention checks were included in the stimuli to prevent completion of the online survey by “bots”. IP addresses and GPS locations of all completed responses were also checked to confirm that there were no duplicates.

Measures

There were three measures aimed at exploring participants’ experience of the online Delphi process:

1. The NASA-TLX task load index (Hart & Staveland, 1988). The NASA-TLX was chosen as a measure of perceived information load because it is the most relevant well-validated response scale that could be identified in the literature. The NASA-TLX is a self-report rating scale used to measure individuals’ subjective experience of “workload”, a hypothetical construct that “represents the cost incurred by a human operator to achieve a particular level of performance” (Hart & Staveland, 1988, p. 140; see also de Winter, 2014). The NASA-TLX comprises six sub-scales: mental demand, physical demand, temporal demand, performance, effort, and frustration. Each sub-scale is measured on a 21-point scale that is taken to represent 0-100 in increments of 5. A global score of 0-100 is obtained by first weighting the sub-scales, using a series of

pairwise comparisons where participants choose which item in each pair was a more important contributor to the level of workload they experienced, then calculating a weighted mean. The NASA-TLX sub-scales can also be analysed separately. The NASA-TLX has been used in over 700 studies across a wide range of domains, particularly those involving human-machine interaction, and is the most cited survey-based workload measure (Grier, 2015). It has been repeatedly evaluated for reliability, sensitivity, and utility during its lifespan. (See Appendix 2).

2. A set of nine self-report questions relating to satisfaction with the Delphi process, based on those used in Hackman and Vidmar (1970) and Boje and Murnighan (1982) but adapted to suit online Delphi groups. (See Appendix 3).
3. Time taken to complete the study. This was used as an indirect measure of participants' actual information load. There is empirical evidence for a positive correlation between information load/task difficulty and decision time (Iselin, 1988; Swain & Haka, 2000; Wright & Ayton, 1988).

Results

The present study was primarily exploratory in nature. However, based on the previous research on information load, we hypothesized that:

1. Increasing group size will increase actual and perceived information load
2. Increasing opinion diversity will increase actual and perceived information load

NASA-TLX

Participants' median score on the NASA-TLX index across all conditions was 53.00 (out of a possible 100, with a higher score indicating greater perceived task load). The descriptive statistics for each condition are set out in Table 1. Medians and interquartile ranges are reported for comparison with Grier's (2015) meta-analysis of 237 NASA-TLX studies. The present scores are comparable to the median scores obtained in studies relating to computer activities in general (54.00), video games (56.50) and medical activities (50.60). This suggests that participants found the task moderately challenging: more so than daily activities (18.30) or navigation (37.70), for example. Notably, even with the maximum group

size of 19 and coupled with high diversity in feedback rationales, the degree of the Delphi task's cognitive "challenge" was still viewed as moderate by participants.

Table 1.

NASA-TLX median scores (out of a possible 100) and interquartile ranges by condition

Group size	Low Diversity		High Diversity	
	Median	IQR	Median	IQR
7	56.00	15.67	53.50	13.00
10	53.67	13.50	54.33	20.91
13	52.33	17.34	54.00	15.33
16	47.67	17.83	56.67	18.50
19	50.67	19.34	54.33	13.66

ANOVAs were used to test for the effects of group size and opinion diversity on scores in the NASA-TLX scale and subscales – hence the reference to mean differences below and the mean scores shown in Figure 1. While the NASA-TLX items each produce ordinal data, the NASA-TLX scale as a whole generates interval data, which can properly be analysed using parametric tests (Carifio & Perla, 2008; Norman, 2010). It is not possible to carry out a multivariate non-parametric ANOVA and so a non-parametric approach would have required many more tests (increasing the risk of type I errors) and could not have tested for interactions between the two independent variables.

A one-way ANOVA found no main effect of group size on NASA-TLX score ($F(4, 242) = 0.81$, $p = .518$), and no linear or other trend was identified. There was no main effect of diversity on NASA-TLX score ($F(4, 242) = 2.94$, mean difference = 2.67, 95% CIs [-0.40, 5.73]. There was also no significant interaction found between group size and diversity. However, post hoc tests identified a significant difference in NASA-TLX score for participants in groups of 19 between high- and low-diversity conditions ($t(50) = 2.26$, $p = .028$, mean difference = 7.33, 95% CIs [0.81, 13.85], Cohen's $d = 0.56$ (a medium effect)), with task load perceived as higher in the high-diversity condition. See Figure 1 below.

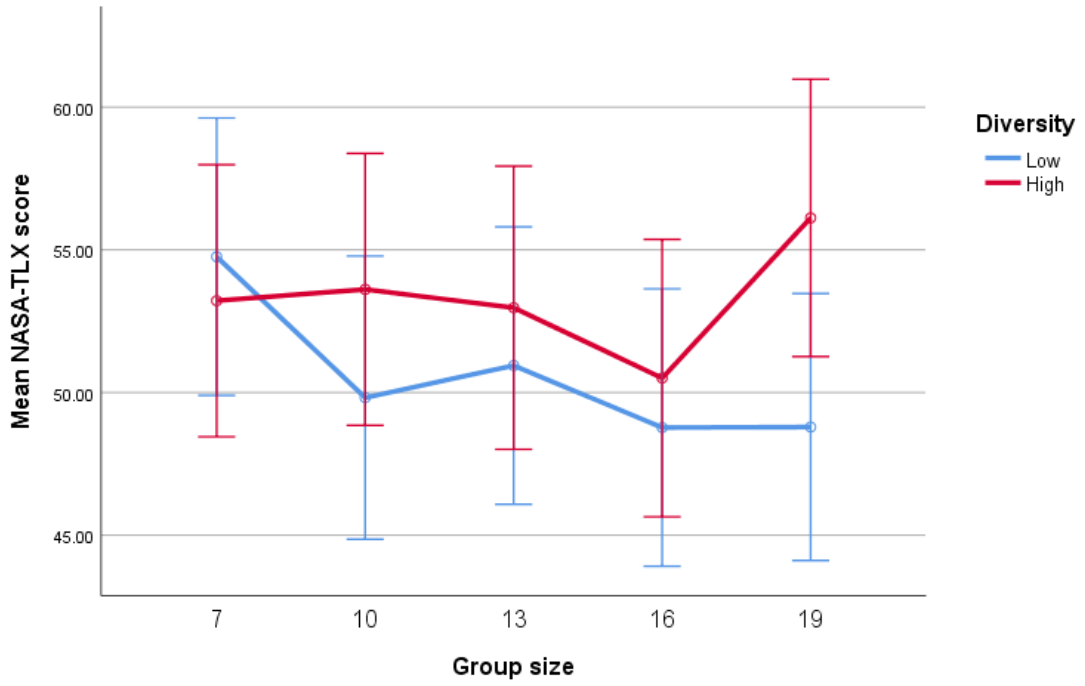


Figure 1. Mean NASA-TLX score by group size and diversity. Error bars are 95% confidence intervals.

NASA-TLX sub-scales

As described in the Method section, the NASA-TLX is composed of 6 sub-scales: mental demand, physical demand, temporal demand, performance, effort, and frustration. There is evidence that measurement of cognitive load using the NASA-TLX should take account of each sub-scale separately (Galy, Paxion, & Berthelon, 2018). ANOVAs were therefore carried out with group size and diversity as independent variables and each of the six sub-scales as dependent variables. No effect of group size was found for any sub-scale. Note, however, that the NASA-TLX subscales were originally devised to measure perceived information load within a variety of tasks and one of the subscales, that of "physical demand", had little prior face validity in our current experimental study. An effect of diversity was found only on the performance subscale ("how successful were you in accomplishing what you were asked to do?"), with participants in high-diversity groups perceiving that they were less successful than those in low-diversity groups[1] ($F(1, 242) = 4.64$, difference = .99, 95% CIs [.09, 1.89], p =

[1] This item is reverse-coded in the main NASA-TLX (a higher score indicates a lower perception of successful performance and so the scores were reversed for the purposes of this analysis].

.032, partial eta squared = .02 (a small effect). See Figure 2 below. Our tentative inference from this final result is that the salience of alternative viewpoints revealed in the high-diversity group setting may have focused participants' attention on the difficulty of the task that they faced[2].

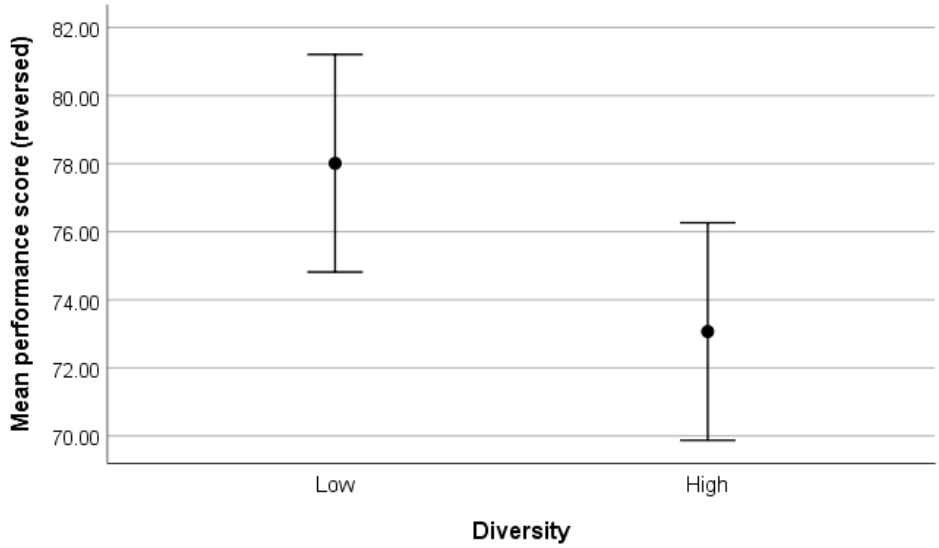


Figure 2. Mean NASA-TLX performance sub-scale score by diversity. Error bars are 95% confidence intervals.

Time taken to complete the study

Descriptive statistics for the time participants took to complete the study (in minutes) are presented in Table 3 below.

Table 3.

Means and standard deviations of total time taken (minutes) by group size and diversity

Group size	Low Diversity		High Diversity	
	Mean	SD	Mean	SD
7	41.43	15.162	48.90	31.139
10	50.02	24.489	34.79	12.884
13	44.88	25.137	57.72	39.745
16	48.86	18.026	49.86	24.372

[2] Care must be taken when analysing single items such as the NASA-TLX subscales, since these are ordinal data. However, Norman (2010) argues that ANOVA can be used, on the grounds that it is robust to non-normality.

19	63.24	42.690	46.76	19.705

A one-way ANOVA found no main effect of group size or diversity on time taken. However, a polynomial contrast for group size identified a significant linear trend (difference = 8.42 minutes, 95% CIs [0.95, 15.89], p = .027), with participants in larger groups taking longer to complete the study than those in smaller groups. In addition, when the low- and high-diversity groups were analysed separately, different patterns were evident (see Figure 5). Planned contrasts found a significant linear trend amongst low-diversity participants (main effect of group size: F(4, 121) = 2.45, p = .050, partial eta squared = .08, polynomial contrast (linear): difference = 13.43, 95% CIs [2.79, 24.07], p = .014) but a significant fourth degree polynomial trend amongst high-diversity participants, with time taken peaking for those in groups of 13 and then falling from there (main effect = non-significant, polynomial contrast (order 4): difference = 12.36, 95% CIs [1.60, 23.12], p = .025). Overall, our inference is that even with group sizes of 19, respondents did not reject the task that they were given – since there was no abrupt decrease in time-taken over group-size increase, which would indicate such a rejection. Instead, the modest reduction observed in the high-diversity groups of 16 and 19 suggests a change towards a simpler cognitive strategy in order to manage information load.

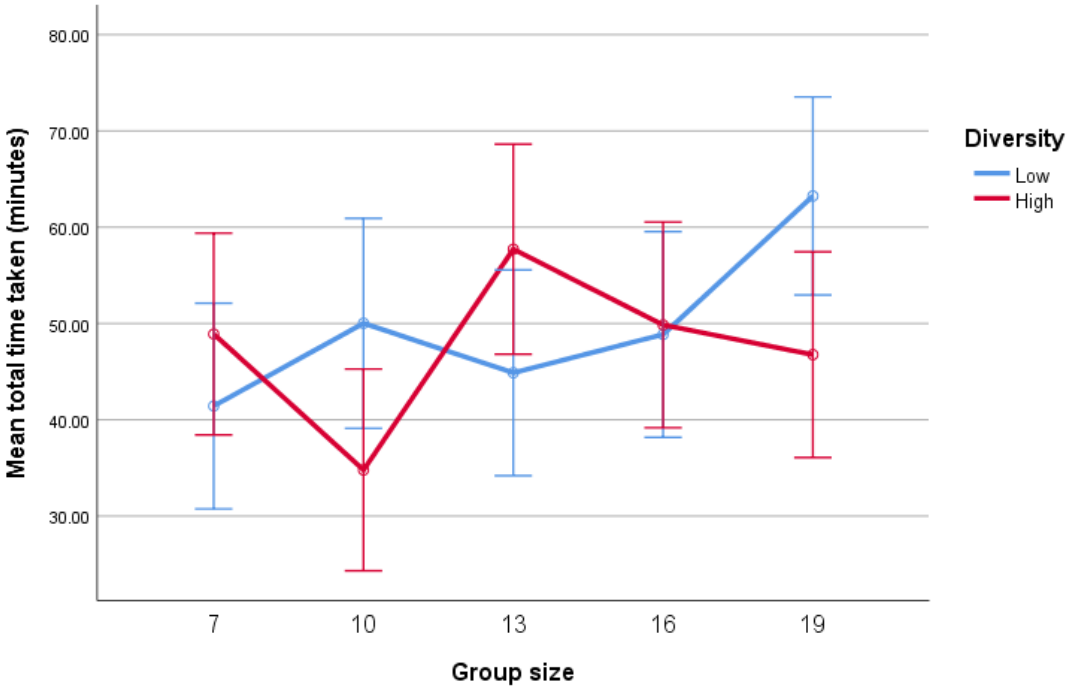


Figure 5. Mean total time taken (minutes) by group size and diversity. Error bars are 95% confidence intervals.

Overall, therefore, hypothesis 1 was supported in relation to actual information load (as measured by time taken) but not perceived information load (as measured by the NASA-TLX). Hypothesis 2 was not supported in relation to actual information load and was only supported for perceived information load to a very limited extent, as we have just discussed.

Self-report satisfaction questions

The nine satisfaction-focussed attitudinal questions given to participants (see Appendix 3) were tested for internal reliability using Cronbach's Alpha (Cronbach, 1951). The Alpha score achieved for all nine questions was .53, revealing that the set of 9 questions were not, together, a coherent single indicator of task and process satisfaction. Two pairs of items were selected that were highly inter-correlated:

- "I enjoyed working on the task" and "I was interested in the questions". Together, these two items produced a Cronbach's alpha of .80, revealing that they represented a reliable measure. We defined this two-item questionnaire as "**task involvement**".

- "The other group members' answers were helpful for completing my revised answers" and "How much influence did the group have on your personal, final ideas about what would be good answers to the questions?" Together, these two items produced a Cronbach's alpha of 0.74, again documenting a reliable measure. We defined this two-item questionnaire as "**group sway**".

Answers to each of the two Task Involvement questions were found to be uncorrelated ($p > 0.05$) with answers to each of the two Group Sway questions, and vice-versa, on analysis of our sample of 282 respondents, demonstrating the separability of our two conceptualizations.

Task involvement

An ANOVA was carried out to explore the effects of group size and diversity on participants' task involvement score, with group size and diversity as the independent variables and interest/enjoyment score as the dependent variables. No effect of diversity was found. A quadratic effect of group size was found, with interest and enjoyment increasing up to a group size of 13 and decreasing thereafter: difference = -0.68, 95% CI of difference [-1.26, -0.10], $p = .021$. The effect is illustrated in Figure 3 below. The pattern found reflects Eppler and Mengis' (2004) inverted u-curve for information load. Figure 3 shows that the pattern was very similar for high- and low-diversity groups. Notably, in the first question of the Task Involvement questionnaire, over 80% of respondents indicated that their experience above the mid-point of the enjoyment scale and on the second task-interest question, over 88% of respondents rated interest as above the mid-point of the scale, no matter the group size or level of diversity of opinion within their group.

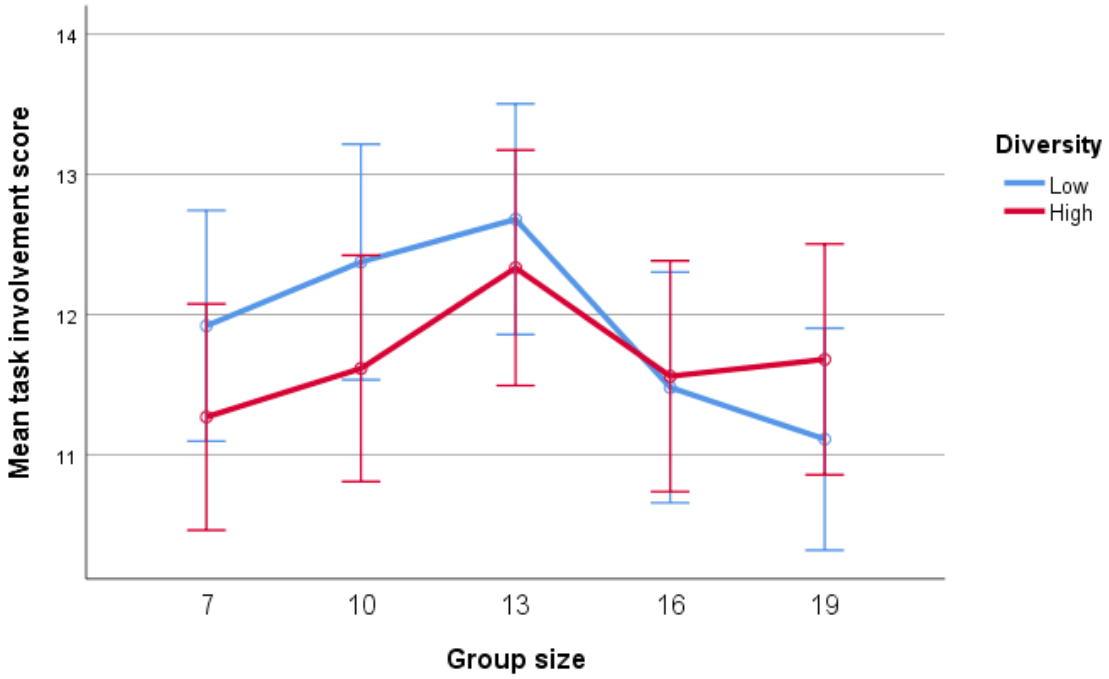


Figure 3. Mean task involvement score by group size and diversity. Error bars are 95% confidence intervals.

Group sway

An ANOVA was used to explore the effects of group size and diversity on group sway score. There was no main effect found for either group size or diversity. However, the picture is different for the low- and high-diversity groups. Amongst low-diversity participants, there was no effect of group size on group sway. For high-diversity participants, however, a linear effect was found (difference = -1.03, 95% CIs [-1.98, -.08], p = .034), with group sway perceived to be highest for groups of 7 and lowest for groups of 19. See Figure 4 below. Our inference from this result is that when the other group members all made the same argument in support of their answer, the size of that group made no difference. Conversely, where the group presented a range of different arguments, the larger groups were perceived as being more useful and influential.

In the second question of the Group Sway questionnaire, respondents' rating of group-based influence was roughly evenly split above and below the mid-point of the scale, whilst on the first question, over 68% of respondents rated perceived-helpfulness of the group as

above the mid-point of the scale. Overall, therefore, others' opinions were seen to be of both moderate influence and helpfulness in forming participants' second round judgments.

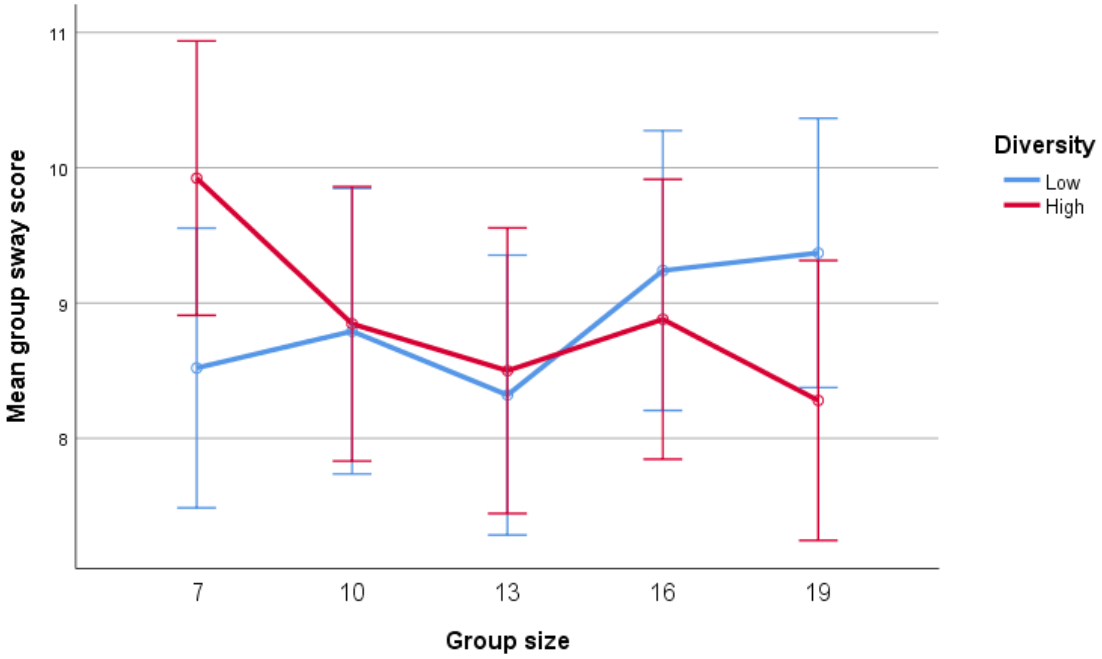


Figure 4. Mean group sway score by group size and diversity. Error bars are 95% confidence intervals.

Discussion

This study set out to explore whether the size of a Delphi group or the diversity of the opinions represented affect participants' experience of the process. Overall, even with a maximum group size of 19 and coupled with high diversity of qualitative feedback from the other panellists, the degree of the challenge presented by the Delphi task to individual participants was moderate and roughly equivalent to that of participating in computer gaming. Notably, participants' task involvement (enjoyment and interest) was quite high - at least 80% of respondents rated their experience as above the mid-point of the scale, no matter the group size or diversity of opinions utilised as between-round feedback. Additionally, the qualitative between-round feedback from their fellow group members was seen to have been of moderate influence and helpfulness.

Importantly, analysis of the time that participants took to complete the Delphi task of responding to ten forecasting questions indicated that participants did not disengage with the Delphi task, even when the Delphi group size was increased to 19. However, there was some indication that participants shortened their processing of highly diverse feedback when group size reached 16 and above. Further research is needed to explore the link between increased information load and possible changes in panellist behaviour to reduce cognitive strain, such as reading fewer of the rationales or engaging with the material at a more superficial level. Process-tracing methods such as mouse- and eye-tracking would be useful for this purpose (Maldonado, Dunbar & Chemla, 2019; Schulte-Mecklenbeck et al., 2017; Fiedler & Glockner, 2012). Panellists behavior will also, no doubt, be influenced by their degree of interest in both the Delphi topic and in the similar or alternative rationales of others – when presented as feedback between Delphi rounds.

In short, in our experimental, rather than real-world study, we found no evidence to suggest that running a Delphi application using qualitative rationales as between-round feedback in groups of 19 or fewer will overload, demotivate, or disengage participants. It is worth noting that this was primarily an exploratory study and further research would be required to ascertain whether these findings will generalise to other, less simple Delphi tasks. Experiments such as the present study are always a trade-off since they involve creating an artificially simple and controlled environment so that the effect of specific variables can be tested. This ability to isolate effects comes at the price of reduced external validity. Certain kinds of Delphi tasks, for example where there are many, detailed questions and/or where panellists provide particularly long, complex rationales, might overload participants in groups smaller than 19. The reverse is true for studies where only few and/or brief rationales are provided. The tipping point for information overload may also be extremely sensitive to the format of the online tool used to deliver the Delphi survey and the exact procedure used. In addition, in many real-world contexts the size of Delphi panel needed will be dictated to some extent by the goals of the survey. The scope of the topic investigated might require a sample much larger than 20 (e.g. 100 or more) in order to capture relevant expertise and generate legitimate, representative results. In such cases, it may make sense to apply a mechanism for reducing the volume of rationales presented to

panellists while retaining as much of the argumentation as possible (e.g. DAD – Gheorghiu et al., 2017).

References

Aengenheyster, S., Cuhls, K., Gerhold, L., Heiskanen-Schüttler, M., Huck, J., & Muszynska, M. (2017). Real-time Delphi in practice – a comparative analysis of existing software based tools. *Technological Forecasting and Social Change, 118,* 15–27.

Agnew, J. R., & Szykman, L. R. (2005). Asset allocation and information overload: The influence of information display, asset choice, and investor experience. *Journal of Behavioral Finance, 6*(2), 57-70.

Airaksinen, T., Halinen, I., & Linturi, H. (2017). Futuribles of learning 2030 – Delphi supports the reform of the core curricula in Finland. *European Journal of Futures Research, 5*, 2.

Belton, I., MacDonald, A., Wright, G., & Hamlin, I. (2019). Improving the practical application of the Delphi method in group-based judgment: A six-step prescription for a well-founded and defensible process. *Technological Forecasting & Social Change, 147*, 72-82.

Biggs, S. F., Bedard, J. C., Gaber, B. G., & Linsmeier, T. J. (1985). The effects of task size and similarity on the decision behaviour of bank loan officers. *Management Science, 31*(8), 919-1054.

Boje, D.M., & Murnighan, J.K. (1982). Group confidence pressures in iterative decisions. *Management Science, 28*(10), 1187-1196.

Bolger, F. (2018). The selection of experts for (probabilistic) expert knowledge elicitation. In Dias, L.C., Morton, A., & Quigley, J. (Eds.), *Elicitation: The science and art of structuring judgement* (pp. 393-444). Cham, Switzerland: Springer.

Bolger, F., Rowe, G., Belton, I., Crawford, M., Hamlin, I., Sissons, A., Taylor Browne Lūka, C., Vasilichi, A., & Wright, G. *The simulated group response paradigm: A new approach to the study of opinion change in Delphi and other structured-group techniques.* Manuscript submitted for publication. Pre-print available at www.doi.org/10.31219/osf.io/4ufzg

Boulkedid, R., Abdoul, H., Loustau, M., Sibony, O., & Alberti, C. (2011). Using and reporting the Delphi method for selecting healthcare quality indicators: A systematic review. *PLoS ONE*, *6*(6), 1-9.

Cairns, G., Wright, G., Fairbrother, P., & Phillips, R. (2017). 'Branching scenarios' seeking articulated action for regional regeneration – a case study of limited success. *Technological Forecasting and Social Change, 124,* 189-202.

Carifio, L., & Perla, R. (2008). Resolving the 50 year debate around using and misusing Likert scales. *Medical Education, 42*, 1150–1152.

Chewning, E. G., & Harrell, A. M. (1990). The effect of information load on decision makers' cue utilization levels and decision quality in a financial distress decision task. *Accounting, Organizations and Society, 15*(6), 527-542.

Cook, G. J. 1993. An empirical investigation of information search strategies with implications for decision support system design. *Decision Sciences*, 24, 683–699.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24 1*, 87-114; discussion 114-85.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science, 9*(3), 351-515.

De Winter, J.C.F. (2014). Controversy in human factors constructs and the explosive use of the NASA-TLX: A measurement perspective. *Cognition, Technology & Work, 16,* 299-302.

Eppler, M.J., & Mengis, J. (2004). The concept of information overload: A review of literature from organizational science, accounting, marketing, MIS, and related disciplines. *The Information Society, 20*(5), 325-344.

Fiedler, S., & Glöckner, A. (2012). The dynamics of decision making in risky choice: An eye-tracking analysis. *Frontiers in Psychology, 3,* article 335.

Galy, E., Paxion, J., & Berthelon, C. (2018). Measuring mental workload with the NASA-TLX needs to examine each dimension rather than relying on the global score: an example with driving. *Ergonomics, 61,* 517-527.

Gheorghiu, R., Dragomir, B., Andreescu, L., Cuhls, K., Rosa, A. Curaj, A., & Weber, M. (2017). *New horizons: Data from a Delphi survey in support of European Union futures policies in research and innovation.* Retrieved from https://ec.europa.eu/jrc/sites/jrcsh/files/fta2018-paper-c3-cuhls.pdf

Goluchowicz, K., & Blind, K. (2011). Identification of future fields of standardisation: An explorative application of the Delphi methodology. *Technological Forecasting and Social Change, 78*(9), 1526-1541.

Griefeneder, R., Scheibehenne, B., & Kleber, N. (2010). Less may be more when choosing is difficult: Choice complexity and too much choice. *Act Psychologica, 133,* 45-50.

Grier, R.A. (2015). How high is high? A meta-analysis of NASA-TLX global workload scores. *Proceedings of the Human Factors and Ergonomics Society 59th Annual Meeting*, 1727-1731.

Hackman, R.J., & Vidmar, N. (1970). Effect of group size and task type on group performance and member reactions. *Sociometry, 33*(1), 37-54.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology, 52,* 139-183.

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing*, *32*(4), 1008-1015.

Helgeson, J.G., & Ursic, M.L. (1993). Information load, cost/benefit assessment and decision strategy variability. *Journal of the Academy of Marketing Science, 21*(1), 13-20.

Humphrey-Murto, S., & de Wit, M. (2019). The Delphi method – more research please. *Journal of Clinical Epidemiology, 106,* 136-139.

Hwang, M.I., & Lin, J.W. (1999). Information dimension, information overload and decision quality. *Journal of Information Science, 25*(3), 213-218.

Iselin, E.R. (1988). The effects of information load and information diversity on decision quality in a structured decision task. *Accounting, Organizations and Society, 13*(2), 147-164.

Janis, I. L. (1972). *Victims of Groupthink: A Psychological study of foreign-policy decisions and fiascoes.* Boston: Houghton Mifflin.

KISTEP (2005). *The Future Perspectives and Technology Foresight of Korea: Identifying challenges and opportunities for Korea's economy and society (Korean language only).*

KISTEP (2017). *The 5th science and technology foresight (2016-2040): Discovering future technologies to solve major issues of future society.*

Landeta, J., Barrutia, J., & Lertxundi, A. (2011). Hybrid Delphi: A methodology to facilitate contribution from experts in professional contexts. *Technological Forecasting & Social Change, 78,* 1629-1641.

Linstone, H.A., & Turoff, M. (1975). *The Delphi method: Techniques and applications*. Reading, Mass.: Addison-Wesley.

Maldonado, M., Dunbar, E., & Chemla, E. (2019). Mouse tracking as a window into decision making. *Behavior Research Methods, 51,* 1085-1101.

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

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

NISTEP Foresight Center (2009). *The 9th science and technology foresight – contribution of science and technology to future society – the 9th Delphi survey. NISTEP report no. 140*. Retrieved from http://hdl.handle.net/11035/696

NISTEP Foresight Center (2015). *The 10th Science and Technology Foresight – Scenario planning from the viewpoint of globalization (Summary Report). NISTEP report no. 164.* Retrieved from http://hdl.handle.net/11035/3079

NISTEP Foresight Center: Dai 11kai GijutsuyosokuChoosa, S&T Foresight 2019 Soogoo Hosokusho (Report of the 11 Foresight in Japanese), NISTEP Report No. 183. Retrieved from https://doi.org/10.15108/nr183.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15,* 625-632.

Paas, F., Tuovinen, J.E., Tabbers, H., & Van Gerven, W.M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63-71.

Park, J. & Jang, S. (2013). Confused by too many choices? Choice overload in tourism. *Tourism Management, 35,* 1-12.

Rowe, G., & Wright, G. (2011). The Delphi technique: Past, present, and future prospects – Introduction to the special issue. *Technological Forecasting and Social Change*, *78*, 1487-1490.

Schulte-Mecklenbeck, M., Johnson, J. G., Böckenholt, U., Goldstein, D. G., Russo, J. E., Sullivan, N. J., & Willemsen, M. C. (2017). Process-tracing methods in decision making: On growing up in the 70s. *Current Directions in Psychological Science, 26*(5), 442-450.

Sthapit, E., Del Chiappa, G., Coudounaris, D. N., & Bjork, P. (2019). Determinants of the continuance intention of Airbnb users: Consumption values, co-creation, information overload and satisfaction. *Tourism Review, 75*(3), 511-531.

Swain, M.R., & Haka, S.F. (2000). Effects of information load on capital budgeting decisions. *Behavioral Research in Accounting, 12,* 171-198.

Sweller, J., van Merrienboer, J.G., & Paas, F.G.W.C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-296.

Toma, C., & Picioreanu, I. (2016). The Delphi technique: Methodological considerations and the need for reporting guidelines in medical journals. *International Journal of Public Health Research*, *4*(6), 47-59.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131. Wright, G. & Ayton, P. (1988). Decision time, subjective probability, and task difficulty. *Memory & Cognition, 16*, 176-185.

Appendix 1 – Forecasting Questions

19 questions were pre-tested for spread of answers and quality of rationales given. The following 10 questions were chosen for use in the study.

1. Kim Jong Un will publicly announce that North Korea will give up its nuclear weapons.
2. The UK Prime Minister Theresa May will face a vote of 'no confidence' in the House of Commons.
3. The personal details of over 1,000 customers will be stolen from a UK bank or financial-service provider.
4. The UK will expel one or more Russian diplomats.
5. A category 5 hurricane (most dangerous) will make landfall on the US mainland.
6. China and the US will begin trade talks designed to de-escalate the trade war.
7. An African country will announce that it is prepared to set up a processing centre for migrants to the EU.
8. Russian President Vladimir Putin will send troops into territory belonging to another country (other than Syria).
9. The Israeli army will kill more than 10 Palestinians.
10. Twenty or more people will die from Ebola in Africa.

Appendix 2 – NASA-Task Load Index (TLX)

NASA Task Load Index

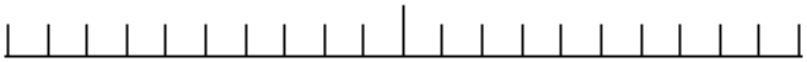
Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?



Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

Appendix 3: Satisfaction self-report questions

Questions

1. This group was too small (in number of members) for best results on the task it was trying to do.
2. This group was too large (in number of members) for best results on the task it was trying to do.
3. My group was creative on this task.
4. *The other group members' answers were helpful for completing my revised answers.
5. *I enjoyed working on the task.
6. I had enough time to work on the task (7-point Likert scale from Too little to Too Much).
7. I was interested in the questions.
8. *How would you rate this process, as a way of making forecasts? (7-point Likert scale from Very Poor to Excellent).
9. How much influence did the group have on your personal, final ideas about what would be good answers to the questions? (A great deal of influence, considerable influence, moderate influence, little influence, almost no influence).

Notes

- The above list is a selection of self-report questions adapted from Hackman and Vidmar (1970). Each item (apart from numbers 6, 8 and 9) is answered using a seven-point Likert-type scale anchored at the ends with "not at all true" and "very true".
- Three additional items were added (marked *). Numbers 5 and 8 are adapted from Boje and Murnaghan (1982).