2019

Comparing in-person and online modes of expert elicitation

Erin Baker
University of Massachusetts - Amherst, edbaker@ecs.umass.edu

Claire Cruickshank
University of Massachusetts Amherst, cruickshank_claire@hotmail.com

Karen Jenni
USGS, kjenni@usgs.gov

Steven Davis
University of California, Irvine, sjdavis@uci.edu

Comparing in-person and online modes of expert elicitation

Claire Cruickshank[a], Erin Baker[a], Karen Jenni[b], and Steven J. Davis[c,d]

[a] *Department of Mechanical and Industrial Engineering, University of Massachusetts, Amherst, MA, 01003 USA*
[b] *Science and Decisions Center, U.S. Geological Survey, Denver, CO, 80225 USA*
[c] *Department of Earth System Science, University of California at Irvine, Irvine, CA, 92697 USA*
[d] *Department of Civil and Environmental Engineering, University of California at Irvine, Irvine, CA, 92697 USA*

Corresponding author: edbaker@ecs.umass.edu

Keywords: probability elicitation, expert judgement, decision analysis, experiment

Abstract

Expert elicitation, a method of developing probability distributions over unknown parameters, traditionally involves in-person interviews by a trained analyst. There is growing interest in using the internet to enable participation of larger, more distributed groups of experts. However, analysts have questioned the quality of judgements elicited online rather than in person. We systematically compare online and in-person elicitation modes, finding no significant difference between the two modes across multiple measures: the two modes are similar in accuracy, uncertainty ranges, number of surprises, fatigue, and the substance of qualitative comments. These findings have an important caveat: many elicitation questions were subject to problems in online administration that made it impossible to compare to in-person results. We conclude that, although online elicitations represent a less resource-intensive option for large expert elicitations, they may require a higher level of testing and quality control since there is no analyst to catch errors or clarify small misunderstandings.

I. Introduction

Expert elicitations have been used for decades to provide decision support for important problems that elicitation protocols administered via in-person interviews have been subjected to rigorous testing (5-9). All of these protocols involve dynamic, interactive communications between the analyst and the expert or experts, with a goal of minimizing the biases and heuristics that seem to be a natural part of human reasoning (10, 11).

In recent years, a number of teams have begun using online expert elicitations instead of traditional in-person interviews (12-21). The appeal of at-a-distance elicitation is clear, as it

avoids the need to travel and enables access to a larger number of geographically diverse experts. In contrast to other distance methods (e.g., by physical mail), online elicitations are particularly attractive as they can provide participating experts with real-time graphical assessment tools and some level of immediate feedback. The automatic storage of the elicitation results further reduces the demands on the experts, who do not have to complete and return a questionnaire or a separate electronic file, for example. However, online elicitation methods have not been rigorously tested and are viewed with skepticism by some prominent analysts (22-24).

Despite this tension and skepticism, few previous studies have investigated the differences between in-person and online elicitations. One small study compared in-person elicitation with an emailed Excel-based tool, eliciting judgements on a single question (25). They found only minor differences in the reported judgements. Three studies used meta-analysis to pool results from multiple expert elicitations on energy technologies (26-28) and found mixed results in terms of central estimates and uncertainty range from different elicitation modes. However, each elicitation considered in these studies used a single elicitation mode, and so any detected differences may reflect other differences in the elicitations, such as question wording and form, or background information provided, rather than just elicitation mode.

In contrast to the limited literature comparing modes of expert elicitation, a large literature exists exploring the effects of different survey modes, including interviews, self-administered questionnaires, and computer assisted methods. Bowling (29) reviews numerous studies related to social, health and epidemiological research, and finds that survey mode is likely to have an impact on the quality of data collected. In particular, in-person interviews require less cognitive effort; are more likely to be completed; lead to longer, more detailed open responses; enhance the participant's motivation; and improve the accuracy of responses (29).

Here, we present the results of a controlled experiment aimed at gauging the effectiveness of elicitations conducted online compared to in-person interviews, including what we believe is the first analysis of the effect of elicitation mode on the accuracy of expert judgements. The details of our experimental design and analytic approach are described in Section II. In brief, during the 2017 spring semester at the University of Massachusetts, Amherst (UMass), we performed a controlled experiment comparing online and in-person elicitations designed according to best practice protocols. Undergraduate and graduate students served as proxies for experts and were asked a total of 20 questions: 39 individuals completed the in-person elicitation interview and 34 individuals completed the online elicitation survey. Questions were designed such that the students would have the necessary expertise to provide meaningful responses, and quantities were selected for which uncertainty would be resolved shortly after the elicitation so that we could evaluate the accuracy of the responses. For example, students were asked about the time that they would need to wait for an elevator in the main library at a specific date and time in the future. For each question, the student experts provided three values: 5^{th} and 95^{th} percentile estimates and a median. A complete list of the elicitation questions may be found in the SI, Table S1. Our research questions, metrics and hypotheses are shown in Table 1.

Table 1. Summary of the research questions and hypotheses

Research Question	Values or metrics used	Hypothesis
Do different modes lead to different central values?	Means of median estimates.	No difference.
Which mode results in a larger expressed uncertainty range?	Normalized uncertainty range.	In-person will have a larger uncertainty range and fewer surprises.
Which mode leads to fewer surprises?	Number of surprises.	
Which mode gives more accurate results?	Multiple quantile scoring rule (37).	In-person will have more accurate results.
Do modes vary in the level of fatigue and engagement for a lengthy survey?	Compare accuracy for questions at the beginning and end of survey, with better accuracy early indicating fatigue.	Fatigue will affect online results more than in-person.

II. Methods

Participants. We recruited students from the University of Massachusetts, Amherst (UMass). Students were recruited using posters, newsletters, and emails. Participants were randomly placed into either the in-person group or online survey group. Participants assigned to the online survey were emailed a web link and instructions on how to access the online elicitation; they were asked to complete the survey within two weeks of receiving the link. Participants assigned to the in-person group were also contacted by email to schedule an appointment time. On average five in-person elicitations were conducted each week; invitations were batched and

interviews were scheduled to occur within a week of initial contact. On completion, all participants received a thirty dollar gift voucher.

Question Design. We prepared twenty questions covering topics of general knowledge and interest to the UMass student population (See SI Appendix, Table S1). Topics were selected so the participants, UMass students, could be considered "experts" and would be able to make well-informed judgements. One set of questions, for example, related to aspects of campus life students encounter regularly: the UMass library, recreation center and catering services. A second set of questions covered popular culture, such as the opening weekend earnings for an upcoming high-profile movie. Although our experts were college students and not professionals, we believe that the findings from our controlled study will be indicative of results that would be found with professionals (34).

In addition, our questions were formulated to meet several requirements. First, we took care to construct our questions to avoid ambiguity, confusion and vagueness regarding the unknown parameter (1). Second, we designed questions where the answer was a single observable value that would be measurable in the months after the completion of the elicitations.

Different question orders were defined to enable us to investigate issues of fatigue and satisficing on the quality of responses. To determine the question order we first grouped questions into six themes, for example questions relating to the UMass library were grouped together. Questions within the same theme were placed in a random order, and the themed groups themselves were randomized. The SI Appendix, Table S2 lists the question orders used.

Elicitation protocols. A challenge was to design elicitation protocols that are as similar as possible while simultaneously allowing each elicitation approach to perform to its best potential as normally practiced. One of the strengths of in-person elicitations is the ability for both the

expert and the analyst to discuss the questions and responses, to ask questions of each other, to tailor the response mode to terms, units, and forms most comfortable for the expert, and to have real-time feedback about the implications of the experts' judgements. These opportunities are not available in a strictly online elicitation, but we chose not to limit interactions in the face-to-face interviews. Among the strengths of the online elicitation is the ability for experts to participate at times that are convenient to them, to come and go from the elicitation as their schedule demands, to access any resources they choose to inform their assessments. We chose not to limit this flexibility in the online elicitations. Another strength of the online format is that experts see their inputs displayed graphically in real-time and can manually adjust those inputs immediately. Although it would be possible to incorporate this kind of real-time graphing of the elicitation results into an in-person format, in our experience that is not common, so we did not add it to the face-to-face protocol.

Face-to-face protocol. All face-to-face elicitations were conducted by a single elicitor, a graduate student and a former high school teacher. From her background, she was comfortable with questioning techniques, including asking probing questions, allowing time for the participant to think about their answer and not being quick to accept a "don't know" response. In addition, she received training, guidance, and support in elicitation interviewing method from members of the research team with significant experience conducting in-person elicitations. This included emphasis on how to conduct an effective expert elicitation interview in a conversational style where the goal is to correctly capture, reflect and verify the expert's opinions.

The elicitation protocol followed the five phases of the Stanford interview process: motivating, structuring, conditioning, encoding and verifying (1). It was structured both to avoid ambiguity and to minimize the response burden on the participants. For example, questions were

clearly stated (e.g., How long will a person wait to use a computer in the public workstation of the library on Monday April 3rd? The wait time will be measured from when the first person arrives after 12:45pm, to when they gain access to a computer), but the expert was allowed to provide a response in whatever units or form she chose (seconds, minutes, hours).

As is typical, the interview script was structured to limit the effect of cognitive bias during the encoding phase. For example, we used follow-up questions to encourage participants to consider the reasons behind their initial judgements as well as to give participants the opportunity to examine all possible outcomes before assessing their judgement. In some instances, on reflection participants altered their subjective probability distributions. To avoid anchoring and adjustment, we asked experts to consider the upper and lower limits of the unknown parameter first. We used probing questions during the interview to reduce overconfidence. For example, experts were asked to explain various scenarios that might cause the observed value to fall below their low estimate (1). After encouraging the expert to consider all possible events, some experts decided to alter their judgements. We also prepared pie charts, in place of the standard probability wheel (1), as a visual aid to assist with encoding the probability judgements.

To provide context and allow for a consistent base of factual knowledge across all participants (22), we compiled a small set of background information relating to each question. We shared a brief summary of the background information and past data with participants at the beginning of each question. By carrying out background research, the analyst gained a better understanding of the topic and so was better equipped to challenge and engage the expert during the conversation.

The final phase, the verifying phase, asked the expert to consider and confirm or reevaluate their judgement. The analyst restated the initial elicitation results back to the expert in slightly different terms to prompt reflection and the expert verified or adjusted their responses made

some statements based on the elicited distribution to verify the judgements before moving on the next question in the elicitation interview.

The elicitation interviews were conducted on campus and participants had full access to the internet. Each participant approached the interview in a different way. Some participants looked up information using the internet and others used the pen and paper provided to carry out some calculations.

Online protocol and implementation. The online elicitation survey was administered by Near Zero, a non-profit organization. Near Zero developed software to elicit expert judgements specifically to inform climate and energy policy (16, 17). The software was customized for the purposes of our research study.

There are several differences between the face-to-face elicitation protocol and the online elicitation. First, the question wording and approach was adapted slightly to take advantage of the software's interactive graphical features. Both elicitation modes contained the same background information and definitions to allow for each question, but in the online elicitation, that information was provided by "rollovers" assigned to specific areas on the screen (i.e. when the participant rolled the mouse curser over the highlighted text, additional information was displayed; see Figure 3). Rollovers were used to avoid overwhelming participants with large sections of written instructions, but their use makes it difficult to ensure that all participants have seen the same information.

Similar to the face-to-face elicitation, the online software also gathered qualitative information via open-ended questions, giving participants an opportunity to type a written response. Participants' written comments provided valuable insights into the participants' thinking and allowed for transparency.

Questions were presented following one of the two pre-defined orders, with one question displayed per webpage. Three different types of response modes and tools (widgets) were used: box-and-whisker, time trend graph and direct entry (Figure 3 shows the box-and-whisker widgets; other widgets are shown in the SI Appendix, Section S2). The box-and whisker widget and time trend graph were used for quantile assessment; whereas the direct entry widget was used for probability assessment.

In this software, the widgets are designed to reinforce best elicitation practice. For example, in the box-and-whisker widget, which was used for 18 of our 20 questions, extreme values (5^{th} and 95^{th} percentiles) are elicited before central values. The instructions shown on the top right corner of the webpage automatically update as the expert enters values, pointing to the next value to be elicited. As values are selected, the box-and-whisker plot is displayed (Figure 3b). If the participant needs to make changes to their values they can adjust the final plot directly. In our study, we set the initial number line that provides the starting point based on what we believed at the time was an appropriate and sufficiently large range. Although the participants can easily extend the range, selecting values not initially displayed, the inclusion of the initial high and low values in the question provided strong anchors (35). Nevertheless, we used this method as it is the common practice in online elicitations.

Data Analysis. Our quantitative data analysis focused primarily on the questions from our elicitation that were consistent between the in-person and online modes.

*Central Values***.** Considering each question independently, we compared the estimates of the median values across the two elicitation modes. For each question, we used a two-sided Welch's t-test to test the hypothesis that there is no difference in the average median estimates. The average median estimate, $\bar{x}_{jk}$ is calculated as:

$$\bar{x}_{jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} q_{a_2\ ijk}$$

where:

n_{jk}: The number of forecasts elicited from elicitation mode j, for question k.

a_2: The 50th percentile ($a_2 = 50$).

$q_{a_2\ ijk}$: The median (50th percentile) estimate for expert i, elicitation mode j, question k.

Uncertainty Range. We hypothesized that participants in the in-person elicitation would express a broader uncertainty range than online participants. The median-based normalized uncertainty range was defined as the difference between the 95th and the 5th percentile of the unknown parameter, normalized by the median (26, 27):

$$NUR_{ijk} = \frac{q_{a_3\ ijk} - q_{a_1\ ijk}}{q_{a_2\ ijk}}$$

where:

a_1 refers to the 5th percentile, a_2 to the 50th percentile, and a_3 to the 95th percentile.

$q_{a_y\ ijk}$: The value of the elicited a_yth percentile for expert i, elicitation mode j, question k.

NUR_{ijk}: The normalized uncertainty range for expert i, survey mode j, question k.

We also considered a second normalized uncertainty range following (36):

$$lNUR_{ijk} = \ln\frac{q_{a_3 ijk}}{q_{a_2 ijk}} - \ln\frac{q_{a_1 ijk}}{q_{a_2 ijk}}$$

We used a two-sided Welch's t-test to test the hypothesis that there is no difference in the average normalized uncertainty ranges.

Rate of Surprises. We hypothesized that over the full set of elicitation results the face-to-face experts would be "surprised" less often than the online experts. We defined a surprise (c_{ijk}) as the event that the observed value fell outside the experts assessed 5th to 95th percentile range:

$$c_{ijk} = \begin{cases} 0 & if\ q_{a_1\,ijk} < T_k < q_{a_3\,ijk} \\ 1 & otherwise \end{cases}$$

where T_k is the observed value for question k. We defined $\hat{p}_j$ as the proportion of surprises in our sample for a given mode (j) as follows:

$$\widehat{p_j} = \frac{1}{F_j} \sum_{k=1}^{7} \left(\sum_{i=1}^{n_{jk}} c_{ijk} \right)$$

We used a two-sided Welch's t-test to test the hypothesis that there is no difference in the proportion of surprises.

Accuracy of forecasts. We used the multiple quantile scoring rule (37, 38) to assess the accuracy of the forecasts. The multiple quantile scoring rule combines the three assessed quantiles and the observed value to give an accuracy score. We defined the accuracy score for expert *i*, survey mode *j*, and question *k*, A_{ijk},
as follows:

$$A_{ijk} = \sum_{y=1}^{3} \begin{Bmatrix} \left| T_k - q_{a_y\,ijk} \right| (a_y) & for\ T_k \geq q_{a_y ijk} \\ \left| T_k - q_{a_y\,ijk} \right| (100 - a_y)\, for\ T_k < q_{a_y ijk} \end{Bmatrix}$$

using notation from above.

Question by Question. We hypothesized the face-to-face elicitation would result in more accurate forecasts. As shown in Table 2 the direction of the effect was mixed, so we tested for significance using a two-sided Welch's t-test to test the hypothesis that there is no difference in the mean accuracy scores. The average score, m_{jk}, was defined as follows:

$$m_{jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} A_{ijk}$$

Aggregated. In order to aggregate across the questions, we adjusted the accuracy scores to a notionally common scale, using the min-max normalization. The normalized score, B_{ijk}, : is

$$B_{ijk} = \frac{A_{ijk} - \min_{ij} A_{ijk}}{\max_{ij} A_{ijk} - \min_{ij} A_{ijk}}, \qquad B_{ijk} \in [0, 1]$$

using notation from above.

This method, used for example in (39), is scale-independent and bounded, and the average normalized score, M_j, for survey mode j, was calculated as follows:

$$M_j = \frac{1}{F_j} \sum_{k=1}^{7} \sum_{i}^{n_{jk}} B_{ijk}$$

We used Welch's t-test to test the hypothesis that there is no difference between the mean of the online normalized scores and the mean of the in-person normalized scores.

Participant Fatigue. One potential benefit of in-person elicitation is that the interaction between the assessor and the expert is assumed to increase the amount of attention paid to the task, and to allow the assessor to help the expert in understanding the questions and express their judgements accurately, even late in the interview process. We hypothesized that participants in

the online survey were more likely to fatigue and to satisfice late in the survey process, answering questions more quickly and with less care. If so, we would expect to see less accurate responses for questions late in the elicitation than early in the elicitation, especially in the online version. We have two questions that appeared early (2nd and 3rd) in one version of the survey and late in the second version of the survey (18th and 16th/17th), allowing us to test this hypothesis directly. For each question and each mode, we carried out Welch's t-test to find out if the question appearing early in the elicitation obtained a better accuracy score.

As a second test, we considered a mixed linear effects model, with the normalized accuracy score as the dependent variable, and the question itself, the order in which it was asked, and the elicitation mode as potential predictor variables:

$$B_{ijk} = \alpha + \beta_1[Question] + \beta_2[QuestionOrder] + \beta_3[ElicitationMode] + \varepsilon$$

If the regression coefficient, β_i is statistically significant, it would indicate that variable i has a strong effect on the normalized accuracy score.

Statistical power, effect size, and multiple comparisons. We tested four hypotheses using five metrics (Table 1). Two hypotheses were tested question by question, and had smaller power. Hypothesis 1 was that the central values elicited by the two elicitation modes would not differ significantly, and we compared the mean of the 50th percentile estimates from each elicitation mode for each question. Our realized sample size of n=34 (online) and n=39 (face-to-face) and a two-sided Welch's t-test provides a power of 0.68 to detect a "medium" difference (using Cohen's d measure of 0.5 as a medium effect (30)) at a significance level of 0.10. This means that the probability of a type II error (accepting the null hypothesis of no significant difference when a medium difference is present) is about 0.32. Hypothesis 3 was that face-to-face results would be more accurate than online results, and we compared the mean of accuracy scores from

each mode for each question. For these two question-by-question comparisons, we defined significance at the 0.05 level, but test at the $p \leq .0071$ level after applying a Bonferroni correction for multiple comparisons (40). For Hypotheses 2 and 4 related to surprises, uncertainty ranges, and the normalized accuracy scores, our metrics allowed aggregating across questions and thus increased power of the statistical test. Our actual sample size yields n=233 for the online condition and n=270 for the face-to-face condition. In these cases the one-sided Welch's t-test provides a power of 0.83 to detect a small difference (Cohen's d measure of 0.2) at a significance level of 0.1, or a power of 0.95 to detect a small-medium difference (0.3) at a significance level of 0.05.

III. Results

Qualitative differences between elicitation modes

The most important difference in the results of the two modes was that responses to numerous questions in the online protocol could not be compared directly with the in-person responses. For various reasons, online responses to 13 of 20 questions were disqualified from direct comparison.

For eight questions, the observed value (realized shortly after completion of the elicitations) fell outside the initial, pre-determined, range shown on the online survey tool. While the number line of the online tool was dynamic and allowed for participants to expand and provide answers outside of the originally displayed range, the initial values nevertheless provided a strong anchor. For four of these, the pre-determined units of measurement in the online survey were also not the most appropriate choice – questions asked for responses in minutes when the true answer (i.e. the later-observed value) turned out to be less than four minutes. This anchoring led to significantly lower accuracy for these questions than for unanchored responses in the in-person elicitation or for online responses where the initial range included the realized value.

For two questions, we were not able to obtain realized values to judge accuracy; one question was worded differently for the two modes; and for two questions the online software simply did not function correctly.

During in-person elicitations, experts naturally provide comments and explanations of their thinking through the interview process. The online elicitation provided optional, open-ended questions with each elicitation question, asking respondents to explain or justify their responses. Somewhat to our surprise, online participants made extensive use of this option, providing substantial and meaningful feedback on 94% of such open questions overall.

The online elicitation was typically completed in less time than the in-person version. In-person interviews (scheduled for 2 hours) ranged in duration from 64 minutes to 123 minutes, with median duration of 90 minutes. For the online elicitation, the software tracked the duration each question was displayed on the screen, giving us an approximation of time spent on each response. Online surveys ranged in duration from 11 minutes to over 20 hours (which we assume came from the participant leaving the page open while away from the task), with median duration of 36 minutes.

Among the in-person participants who scheduled an interview, 39 out of 45 recruited participants (87%) attended and 37 of the 39 completed all questions in the elicitation. Among the online participants, 47 participants were emailed a link to the elicitation and 34 elicitations were fully completed (72%).

Quantitative comparison of elicitation results.

Figure 1 summarizes the responses to each of the seven questions for which in-person and online responses were directly comparable: in each panel, the distributions of the elicited median values from online and in-person elicitations are shown on top (green) and bottom (purple), respectively. Each individuals' response is also represented by a thin line reflecting the 5^{th} to 95^{th} percentile of elicited values, with red lines indicating "surprised experts" (where the true answer to the question ended up being outside the boundaries of that expert's 90% confidence interval).

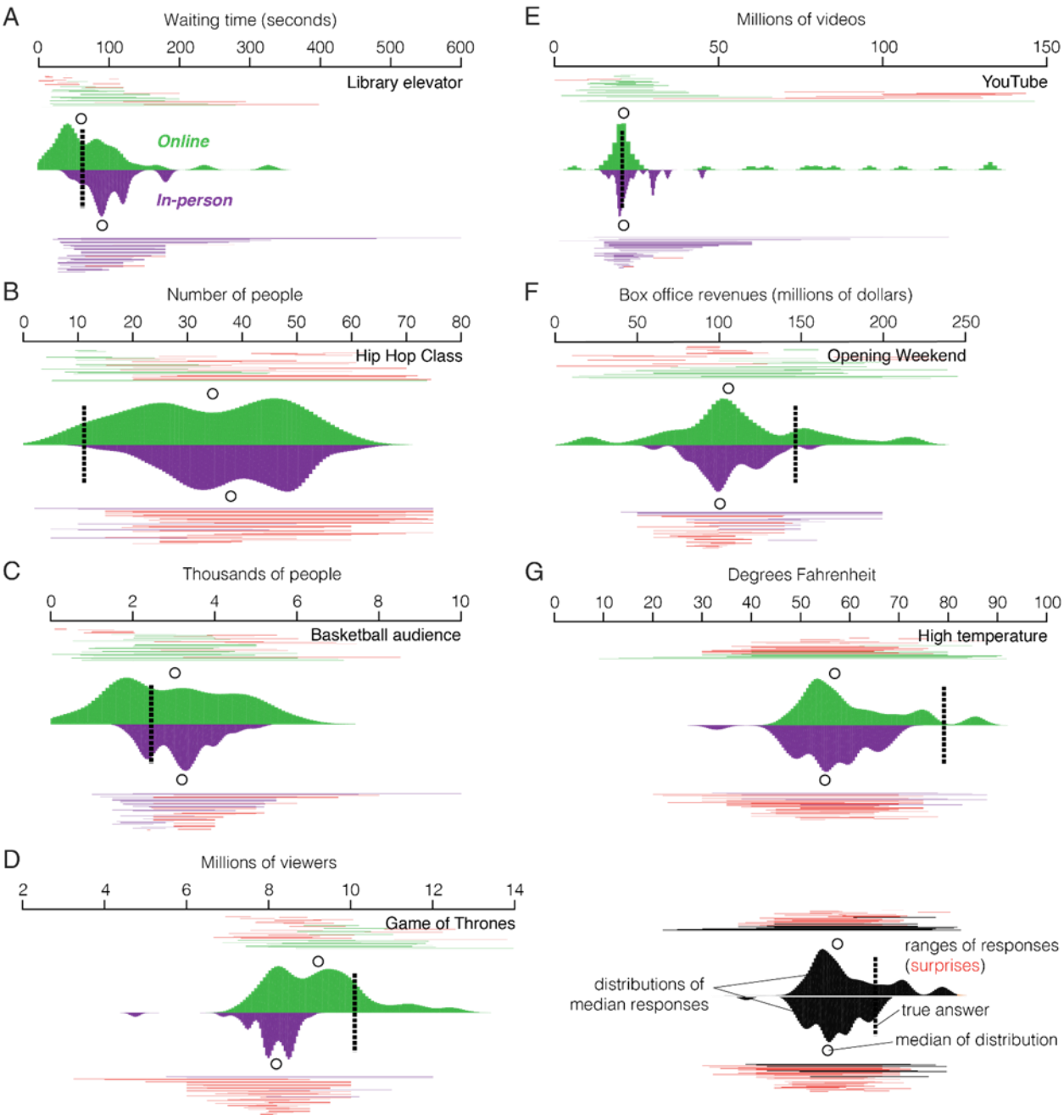

Figure 1. Summary of results for online (upper, green) and in-person (lower, purple) elicitations.

Similar estimates of central values. In a question-by-question comparison, we find that the median of the elicited median values (indicated by open circles in Fig. 1) was significantly different between modes in two of the seven questions (Table 2). However, the differences were

in opposite directions: for the "Game of Thrones" question (number of viewers of season premiere; Fig. 1D), the online results were closer to the true answer; for the "YouTube" question (number of videos downloaded; Fig. 1E), the in-person results were closer. These contrasting results, and the lack of significant difference between the two elicitation modes for the other 5 comparable questions, support our hypothesis that there is little reason to expect a difference between the central values for the two elicitation modes.

Similar uncertainty ranges and levels of overconfidence. Both the online and in-person elicitations produced a higher rate of surprises than they would if experts were perfectly calibrated, with surprise rates of 51% (n = 233) and 55% (n = 270), respectively. These proportions are not significantly different (t=-1.02 df= 489, p = .031, Cohen's d = 0.09). Similarly, Figure 2A shows that the normalized uncertainty ranges from the online elicitations (n=233, m=0.86, sd=0.86) and the in-person elicitations (n=270, m=0.82, sd =0.65) were not significantly different (t=0.56, df=428, p= 0.57, Cohen's d = 0.05).

Similar accuracy of judgements. In a question-by-question comparison, we found a statistically significant difference in accuracy (measured by a multiple quantile scoring rule) in three of the seven questions, but the effect is not consistent. Forecasts from the in-person elicitation were more accurate in two of the cases and the online forecasts more accurate in one (Table 2). Furthermore, when the accuracy scores are normalized to allow comparison across all questions (Figure 2B), no significant difference is found between overall accuracy in the two elicitation modes (t= 0.51, df=453, p=0.61, Cohen's d = 0.05).

Lack of evidence for fatigue effects. We found no significant effect of question order on accuracy for either elicitation mode; neither using a mixed linear effects model, nor in direct comparisons using questions that appeared early in one survey version and late in the second version. Thus,

we found no evidence to support the idea of participant fatigue or satisficing in either elicitation mode.

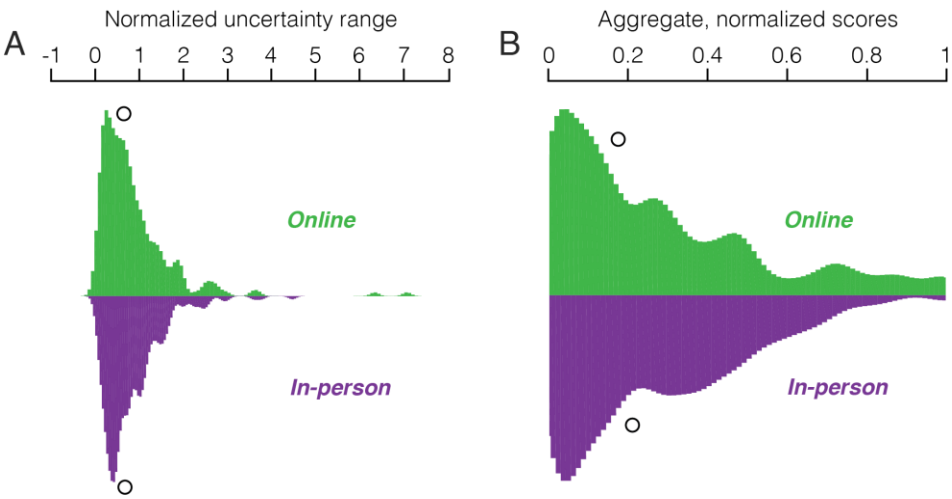

Figure 2. Comparison of the aggregated online and in-person results for the (A) normalized uncertainty range (numerically lower values indicate less uncertainty was expressed), and (B) normalized accuracy scores (numerically lower scores indicate more accurate results).

IV. Discussion and Recommendations

Our quantitative results suggest no difference in performance between online and in-person expert elicitations. Among comparable questions, we find no statistically significant difference between the normalized uncertainty ranges, the number of surprises, or the normalized accuracy of the forecasts when all questions are considered together. Although there are statistically significant differences between the central values and the accuracy for a few specific questions, these showed no pattern; each elicitation mode came out as "better" about half the time. We also find no evidence of fatigue or an impact from question order on either elicitation mode. The statistical power of our sample size and tests was sufficient to successfully detect "moderate"

differences (Cohen's d of 0.5 or greater (30)) on a question-by-question basis and to detect even small differences for measures where the responses to all questions could be aggregated.

There are reasonable explanations for the countervailing differences between the modes on specific questions. For example, in the Game of Thrones question (Fig. 1D), experts used an interactive chart in the online questionnaire that was static in the in-person elicitation. Previous evidence suggests that online surveys can have superior results in such cases (31, 32). In contrast, the question on YouTube (Fig. 1E) required calculations, in which the analyst can be a helpful guide.

Our findings of mode equivalence, however, are subject to an important caveat: as described above, only 7 of the 20 elicitation questions led to directly comparable responses. In-person elicitations permit natural and dynamic responses by the interviewer to the individual expert. Although artificial intelligence may someday provide similar interactivity to online elicitations, for now there remain many possibilities for unexpected errors to occur in online elicitations that are likely to be obviated or avoided by in-person interviews. Of greatest concern is whether the results in the online elicitation that we disqualified might have been regarded as legitimate in an elicitation in which the true answer is fully unknown, which is commonly the situation when expert elicitations are undertaken.

We have a few relatively straightforward practical suggestions for improving online elicitations. For example, for box-plot type questions, rather than showing an initial number line anchored on values the study designers select, respondents should be asked first for their own "low" and "high" estimates, and a number line generated based on those values. Such an approach was used by Wiser et al. (21) in a recent large-scale online elicitation of wind energy experts. Similarly, it should be possible to let participants choose from a range of units, and then generate the initial

number line based on their preferred units. More generally, software-based surveys should be subject to more rigorous testing than in-person protocols before being deployed. Moreover, results from online studies could be improved if the results are analyzed on an ongoing basis; allowing for questions where the software is mal-functioning, or in which participants seem to have misunderstood the question, to be revisited (with those who have already responded) and revised (for future respondents) before the end of the elicitation. Unfortunately, there may be unanticipated miscommunication and ambiguity even given the most careful testing of online surveys, and analysts must therefore treat the results of online elicitations carefully and especially conservatively.

All of these suggestions indicate that designing and implementing robust online elicitations will require a great deal of up-front work, including all the preparation that best practices for in-person elicitations require (22) as well as additional testing specific to at-a-distance elicitations where software issues can arise and communication between the assessment team and the participating experts must be more formalized and structured. The time and cost savings for on-line elicitations over in-person elicitations may therefore not be as large as many analysts have hoped. Nonetheless, on-line elicitation may enable studies where the costs and logistics of in-person elicitations are too formidable.

Our conclusions are of course limited by the context of our experiment (students standing in as experts, and a straightforward, well-defined set of questions with near-term realizations of true answers). In the future, structured field-tests of both modes with real experts and real questions of interest are important to substantiate our findings and explore relative merits of different elicitation modes. As the first large-scale controlled experiment into this question, however, this

study provides an initial and positive answer to the question of the effectiveness of conducting expert elicitation online.

In-person interactions are likely to remain the "gold standard" for investigators whose focus is obtaining numeric, probabilistic estimates of potential future outcomes, or who are equally interested in the conversations and joint learning between the analysis team and experts that occur during elicitations, and in studies where interaction between experts is considered a crucial part of the knowledge acquisition process (33). But increasingly studies have other goals. For example, analysts may be interested in eliciting judgments from a very large number of experts, or from experts who are widely dispersed geographically, or they may want to perform and repeat elicitations at regular intervals. Our results suggest that online elicitations, carefully developed and implemented, may be a viable method to achieving such broader goals.

Acknowledgements: Funding for this study was provided by the Alfred P. Sloan Foundation. Any use of trade, product, or firm names in this publication is for descriptive purposes only and does not imply endorsement by the U.S. Government.

.

References

1. Morgan M, Henrion M & Small M (1990) *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis,* (Cambridge Univ Press, New York), pp 346.

2. Cooke R (1991) *Experts in Uncertainty: Opinion and Subjective Probability in Science,* (Oxford Univ Press, New York), pp 336.

3. O'Hagan A, *et al* (2006) *Uncertain Judgements: Eliciting Experts' Probabilities,* (John Wiley & Sons, Hoboken, NJ), pp 321.

4. Oppenheimer M, Little C & Cooke R (2016) Expert judgement and uncertainty quantification for climate change. *Nat.Clim.Change* 6(5): 445-451.

5. Wallsten T & Whitfield R (1986) Assessing the risks to young children of three effects associated with elevated blood-lead levels. *Office of Scientific and Technical Information* ANL/AA-32 ON: DE87004219

6. Clemen R & Reilly T (2000) *Making hard decisions with decision tools,* (Duxbury/Thomson Learning, Pacific Grove, CA), pp 848.

7. Jenni K & van Luik A (2010) in *Geological Repository Systems for Safe Disposal of Spent Nuclear Fuels and Radioactive Waste,* eds Ahn J & Apted M (Woodhead Publishing, Boca Raton, FL), pp 580-609.

8. Spetzler CS & Staël von Holstein C-AS (1975) Probability encoding in decision analysis. *Manage Sci* 22(3): 340-358.

9. Aspinall W (2010) A route to more tractable expert advice. *Nature* 463(7279): 294-295.

10. Tversky A & Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157): 1124-1131.

11. Garthwaite P, Kadane J & O'Hagan A (2005) Statistical methods for eliciting probability distributions. *J Am Stat Assoc* 100(470): 680-700.

12. Anadon L, Bosetti V, Bunn M, Catenacci M & Lee A (2012) Expert judgments about RD&D and the future of nuclear energy. *Environ Sci Technol* 46(21): 11497-11504.

13. Aspinall W, Cooke R, Havelaar A, Hoffmann S & Hald T (2016) Evaluation of a performance-based expert elicitation: WHO global attribution of foodborne diseases. *PLoS One* 11(3)

14. Bojke L, *et al* (2010) Eliciting distributions to populate decision analytic models. *Value Health* 13(5): 557-564.

15. Dodd P, Yuen C, Sismanidis C, Seddon J & Jenkins H (2017) The global burden of tuberculosis mortality in children: A mathematical modelling study. *Lancet Glob Health* 5(9): e898-e906.

16. Inman M & Davis S (2012) Energy high in the sky. expert perspectives on Airborne wind energy systems. *Near Zero* 2018(06/18)

17. Inman M & Davis S (2012) How low will solar photovoltaic prices go? summary of a near zero expert elicitation. *Near Zero* 2018(06/18)

18. Leal J, Wordsworth S, Legood R & Blair E (2007) Eliciting expert opinion for economic models: An applied example. *Value Health* 10(3): 195-203.

19. Ricci E, Bosetti V, Baker E & Jenni K (2014) From expert elicitations to integrated assessment: Future prospects of carbon capture technologies. *SSRN Electronic Journal*

20. Speirs-Bridge A, *et al* (2010) Reducing overconfidence in the interval judgments of experts. *Risk Analysis* 30(3): 512-523.

21. Wiser R, *et al* (2016) Expert elicitation survey on future wind energy costs. *Nat Energy* 1(10): 1.

22. Morgan M (2014) Use (and abuse) of expert elicitation in support of decision making for public policy. *Proc Natl Acad Sci USA* 111(20): 7176-7184.

23. Colson A & Cooke R (2018) Expert elicitation: Using the classical model to validate experts' judgments. *Rev Environ Econ Pol* 12(1): 113-132.

24. Rai V (2013) Expert elicitation methods for studying technological change under uncertainty. *Environ Res Lett* 8(4): 041003.

25. Grigore B, Peters J, Hyde C & Stein K (2017) EXPLICIT: A feasibility study of remote expert elicitation in health technology assessment. *BMC Med Inform Decis Mak* 17(1)

26. Anadon L, Nemet G & Verdolini E (2013) The future costs of nuclear power using multiple expert elicitations: Effects of RD&D and elicitation design. *Environ Res Lett* 8(3): 034020.

27. Verdolini E, Anadon L, Lu J & Nemet G (2015) The effects of expert selection, elicitation design, and R&D assumptions on experts' estimates of the future costs of photovoltaics. *Energy Policy* 80: 233-243.

28. Nemet G, Anadon L & Verdolini E (2017) Quantifying the effects of expert selection and elicitation design on experts' confidence in their judgments about future energy technologies. *Risk Anal* 37(2): 315-330.

29. Bowling A (2005) Mode of questionnaire administration can have serious effects on data quality. *J Public Health* 27(3): 281-291.

30. Cohen J (1988) *Statistical power analysis for the behavioral sciences,* (L. Erlbaum Associates, USA), pp 400.

31. Dillman DA, *et al* (2009) Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the internet. *Soc Sci Res* 38(1): 1-18.

32. Fricker S, Galesic M, Tourangeau R & Yan T (2005) An experimental comparison of web and telephone surveys. *Public Opin Q* 69(3): 370-392.

33. Hanks T, Abrahamson N, Boore D, Coppersmith K & Knepprath N (2009) Implementation of the SSHAC guidelines for level 3 and 4 PSHAs: Experience gained from actual applications. *US Geological Survey Open-File Report* 2009-1093: 1-66.

34. Visser P, Krosnick J, Lavrakas P & Kim N (2014) in *Handbook of research methods in social and personality psychology*, (Cambridge University Press, New York, NY), pp 402-440.

35. Marquard J & Robinson S (2008) in *Decision Modeling and Behavior in Complex and Uncertain Environments. Springer Optimization and Its Applications, vol 21.* (Springer, New York, NY), pp 33-55.

36. Douglas J, Ulrich T, Bertil D & Rey J (2014) Comparison of the ranges of uncertainty captured in different seismic-hazard studies. *Seismol Res Lett* 85(5): 977-985.

37. Jose V & Winkler R (2009) Evaluating quantile assessments. *Oper Res* 57(5): 1287-1297.

38. Grushka-Cockayne Y, Lichtendahl K, Jose V & Winkler R (2017) Quantile evaluation, sensitivity to bracketing, and sharing business payoffs. *Oper Res* 65(3): 712-728.

39. Bickel J (2007) Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decis Anal* 4(2): 49-65.

40. Miller R (1981) *Simultaneous Statistical Inference*, (Springer New York, New York, NY), pp 299.

Additional Figures and Tables

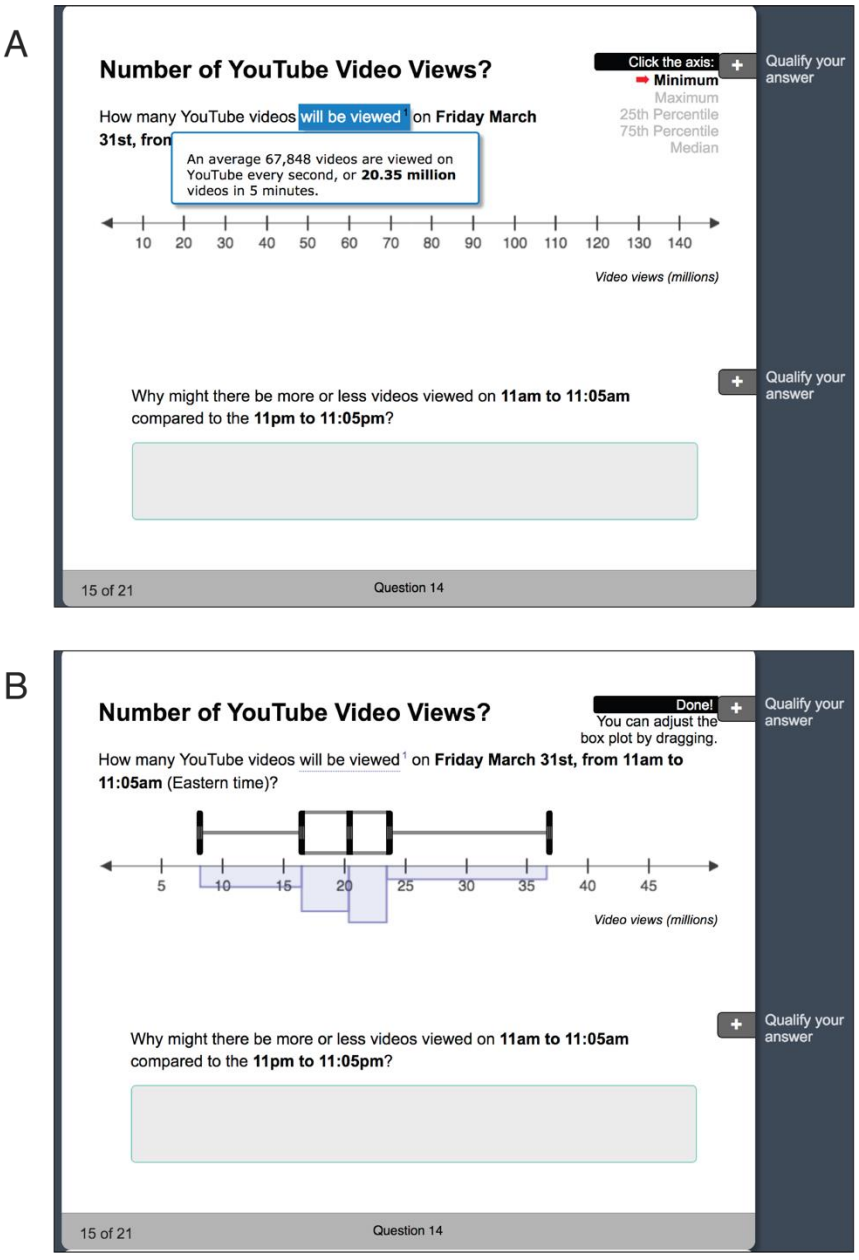

Figure 3. Illustration of box-and-whisker widget. Panel (A) shows a rollover box used to provide definitions and background information, and the instructions for numeric entries in the upper right of the screen. Panel (B) shows a completed assessment and the note informing the expert that they can still make adjustments to their inputs.

Table 2. Comparison of median estimates and normalized score.

		Median Estimate (H_0: no difference in average median estimates)						**Normalized Accuracy Score** (H_0: no difference in average normalized score)					
		Median estimate		**Effect size**	**Welch's two-sided t-test**			**Normalized accuracy score**		**Effect size**	**Welch's two-sided t-test**		
		Online	In-person	Cohen's d *				Online	In-person	Cohen's d *			
	Question (N_{online}, N_{F2F})	M (SD)	M (SD)	d (90%CI)	*t*	*df*	*p*	M (SD)	M (SD)	d (90%CI)	*t*	*df*	*p*
A	Library elevator (33, 37)	78.8 (65.5)	99.9 (32.8)	-0.41 (-0.82, -0.011)	-1.67	45	0.101	0.16 (0.21)	0.11 (0.09)	0.35 (-0.05, 0.76)	1.42	43	.1629
B	Hip hop class (34, 39)	33.8 (14.2)	38.6 (10.4)	-0.38 (-0.77, 0.015)	-1.59	59	0.118	0.36 (0.28)	0.39 (0.18)	-0.16 (-0.55, 0.23)	-0.67	54	.5084
C	Basketball attendance (33, 39)	3048 (1493)	3173 (739)	-0.11 (-0.50, 0.29)	-0.436	45	0.665	0.32 (0.28)	0.18 (0.13)	0.69 (0.29, 1.1)	2.77	44	*.0081* †
D	Game of Thrones (33, 39)	9.22 (1.17)	8.07 (0.70)	1.2 (0.80, 1.7)	4.98	50	*7.88E-06* †	0.24 (0.21)	0.42 (0.20)	-0.86 (-1.27,-0.45)	-3.61	65	*.00058* †
E	YouTube (34, 39)	42.8 (37.5)	24.2 (7.04)	0.71 (0.31, 1.1)	2.85	35	*0.007* †	0.18 (0.30)	0.025 (0.03)	0.75 (0.35,1,16)	2.99	33	*.0052* †
F	Opening weekend (34, 39)	113 (46.3)	105 (19.2)	0.23 (-0.17, 0.62)	0.92	42	0.364	0.29 (0.25)	0.27 (0.16)	0.10 (-0.29,0.49)	0.43	56	.6693
G	High temperature (32, 38)	60.1 (10.1)	56.2 (7.62)	0.44 (0.039, 0.85)	1.81	56	0.076	0.37 (0.23)	0.45 (0.22)	-0.36 (-0.77, 0.04)	-1.51	64	.1358

* Cohen's d is the difference between the means (here online – in-person) divided by the pooled standard deviation (30). Effect sizes are typically categorized as small (0.2), medium (0.5), or large (0.8).

† Using a Bonferroni correction for multiple measures comparison (40), $p \leq .0071$ would be considered equivalent to a standard $p \leq .05$ significance.