Appendix S4. Calibration and information scores for all expert assessments (including expert #10) and robustness analysis.

While the main text presents results with Expert 10 excluded, Table 1 (below) shows the results with all experts included. Again, the performance-weighted decision maker (PW) outperforms the equal-weights decision maker (EW) in terms of both calibration and information. The PW combination gives expert 8 100% of the weight and the other nine experts no weight. The EW decision maker is less informative than most of the individual experts; only expert 10 has a lower information score. The combined score is the product of each expert's calibration score and the information score on the calibration variables. The last two columns in Table 1 show each expert's information scores relative to the EW decision maker. For all variables, these scores range from 0.9556 (expert 10, the least informative expert) to 3.268 (expert 2, the most informative expert). This range is an informative benchmark for the robustness analysis presented below (Tables 2-4).

Table 1. Calibration and information scores for expert assessments, including expert #10. EW represents the combined estimate of all expert assessments weighting each equally. PW indicates the combined estimate of experts' assessments using the performance-weighted criterion. Larger calibration scores indicate a greater probability that an expert's estimates encompassed the 'true' values of the calibration variables. Larger information scores indicate a narrower interval between the expert's 5$^{th}$ and 95$^{th}$ percentiles (i.e., greater certainty), regardless of whether that interval actually encompassed the 'true' values of the calibration variables. See the text and Appendix S1 for further explanation of how the weights, calibration scores, and information scores were calculated.

| | | Information score | | | Information relative to EW | |
|---|---|---|---|---|---|---|
| Expert ID | Calibration score | All variables | Calibration variables | Combined score | All variables | Calibration variables |
| 1 | 0.4920 | 2.635 | 3.747 | 1.844 | 1.743 | 2.943 |
| 2 | $7.543\times10^{-12}$ | 5.644 | 5.987 | $4.516\times10^{-11}$ | 3.268 | 2.555 |
| 3 | $2.496\times10^{-6}$ | 2.501 | 4.367 | $1.090\times10^{-5}$ | 1.525 | 1.391 |
| 4 | $7.985\times10^{-4}$ | 1.914 | 2.731 | 0.002181 | 1.808 | 2.452 |
| 5 | $3.209\times10^{-14}$ | 5.156 | 5.639 | $1.811\times10^{-13}$ | 2.553 | 1.745 |
| 6 | $2.789\times10^{-5}$ | 3.788 | 4.938 | 0.0001377 | 1.722 | 2.16 |
| 7 | $7.985\times10^{-4}$ | 3.623 | 4.590 | 0.003665 | 1.733 | 1.834 |
| 8 | 0.7062 | 2.923 | 4.170 | 2.945 | 1.457 | 2.616 |
| 9 | 0.0011 | 1.317 | 2.618 | 0.00288 | 1.442 | 2.325 |
| 10 | $3.209\times10^{-14}$ | 0.5074 | 0.3431 | $1.102\times10^{-14}$ | 0.9556 | 1.022 |
| EW | 0.1970 | 1.189 | 1.809 | 0.3563 | | |
| PW | 0.7062 | 2.923 | 4.170 | 2.945 | | |

The first robustness check is to remove one expert at a time and recalculate the PW decision maker. As only one expert received weight in the original PW decision maker, removing any other expert does not affect the new PW decision maker's scores (Table 2). Removing expert 8 does lower the calibration and information of the PW decision maker, but it is still performs better than the original EW decision maker. The information of this new decision maker relative to the original is similar to that of an average expert relative to the EW decision maker (as seen in column 6 of Table 1).

Table 2. Robustness analysis on experts.

| | | Information score | | Information relative to original PW | |
|---|---|---|---|---|---|
| Expert removed | Calibration score | All variables | Calibration variables | All variables | Calibration variables |
| 1 | 0.7062 | 2.923 | 4.17 | 0 | 0 |
| 2 | 0.7062 | 2.922 | 4.169 | 0 | 0 |
| 3 | 0.7062 | 2.923 | 4.17 | 0 | 0 |
| 4 | 0.7062 | 2.923 | 4.17 | 0 | 0 |
| 5 | 0.7062 | 2.923 | 4.17 | 0 | 0 |
| 6 | 0.7062 | 2.923 | 4.17 | 0 | 0 |
| 7 | 0.7062 | 2.923 | 4.17 | 0 | 0 |
| 8 | 0.492 | 2.634 | 3.747 | 1.669 | 2.235 |
| 9 | 0.7062 | 2.729 | 4.096 | 0 | 0 |
| 10 | 0.7062 | 1.895 | 1.988 | 0 | 0 |
| None | 0.7062 | 2.923 | 4.170 | | |

The second robustness check is similar, removing one calibration variable at a time to see if any question has a large impact on the final decision maker. For each item removed, the PW decision maker assigned all weight to expert 8. Removing any one item has essentially no impact on the decision maker's performance (Table 3).

Table 3. Robustness analysis on items.

| Item removed | Calibration score | Information score | | Information relative to original PW | |
|---|---|---|---|---|---|
| | | All variables | Calibration variables | All variables | Calibration variables |
| 1 | 0.6827 | 2.919 | 4.274 | 0.0001736 | 0.0004634 |
| 2 | 0.6827 | 2.919 | 4.275 | 0 | 0 |
| 3 | 0.6827 | 2.907 | 4.201 | 0 | 0 |
| 4 | 0.6827 | 2.905 | 4.19 | 0 | 0 |
| 5 | 0.6827 | 2.891 | 4.105 | 0 | 0 |
| 6 | 0.6827 | 2.919 | 4.274 | 0 | 0 |
| 7 | 0.6827 | 2.886 | 4.077 | 0 | 0 |
| 8 | 0.6827 | 2.883 | 4.06 | 0 | 0 |
| 9 | 0.6827 | 2.895 | 4.129 | 0 | 0 |
| 10 | 0.5503 | 2.906 | 4.193 | 0 | 0 |
| 11 | 0.6827 | 2.888 | 4.091 | 0 | 0 |
| None | 0.7062 | 2.923 | 4.17 | | |

A final robustness check is to recalculate the PW scores without the top two experts: expert 1 and expert 8. These are the only two experts in this study with calibration scores larger than 0.05. Removing these two experts does degrade the calibration and information of the PW combination, but its calibration is still sufficiently high. Even with the two best experts removed, the PW decision maker outperforms the original EW decision maker with all ten experts included (Table 4).

Table 4. Expert and PW decision maker scores with expert 1 and expert 8 removed.

| | | Information score | | |
|---|---|---|---|---|
| Expert ID | Calibration score | All variables | Calibration variables | Normalized weight |
| 2 | $7.543\times10^{-12}$ | 5.668 | 5.987 | $5.09\times10^{-9}$ |
| 3 | $2.496\times10^{-6}$ | 2.5 | 4.367 | 0.00123 |
| 4 | 0.0007985 | 1.968 | 2.731 | 0.246 |
| 5 | $3.211\times10^{-14}$ | 5.156 | 5.639 | $2.04\times10^{-11}$ |
| 6 | $2.789\times10^{-5}$ | 3.787 | 4.938 | 0.0155 |
| 7 | 0.0007985 | 3.622 | 4.59 | 0.413 |
| 9 | 0.0011 | 1.316 | 2.618 | 0.325 |
| 10 | $3.211\times10^{-14}$ | 0.5067 | 0.3431 | $1.24\times10^{-12}$ |
| PW (without top two experts) | 0.492 | 1.868 | 2.843 | |

From the three robustness checks above, we can conclude that although the final PW decision maker contains only one expert, its performance is still quite robust to the exclusion of experts from the elicitation.