Cooke's Classical Model

Robustness and Discrepancy Analysis in Expert Judgment Studies

Puck de Nooy

July 14, 2025

TU Delft

Robustness and Discrepancy Analysis in Expert Judgment Studies

by

Puck de Nooy

to obtain the degree of Bachelor of Science
at the Delft University of Technology,
to be defended publicly on Tuesday July 15, 2025 at 10:00 AM.

Student number:	5543355	
Project duration:	April, 2025 – July, 2025	
Thesis committee:	Dr. ir. G.F. Nane,	TU Delft, supervisor
	Dr. ir. L.J.J. van Iersel,	TU Delft

This thesis is confidential and cannot be made public until July 15, 2025.

TU Delft

Laymen's Summary

Cooke's Classical Model is a method that can be used the make predictions when no historical data is available. It combines multiple predictions from experts to create one final prediction to answer the research question. This research analyses how stable and informative the outcomes of Cooke's Classical Model are. We examine how the results change when the input data is slightly altered, to see if the result is stable. Additionally, we compare each result of the model's with other results to see how informative they are and which result stays the most consistent.

We have found that the more experts are included in the study, the more stable the outcome becomes. In most cases, stability also improves when experts are asked to answer more questions. Among the models tested, the outcome that gives the same weight to all experts tends to produce the most informative results.

Summary

When researchers are left with important questions and no historical data is available, such as during the spread of a new virus, then Cooke's Classical Model (CM) of Structured Expert Judgment (SEJ) is one of the methods that can be used to make predictions. The model aggregates expert assessments into a single prediction, which we call a Decision Maker (DM). This bachelor thesis investigates the robustness and discrepancy of Cooke's Classical Model using a dataset of 49 different studies. Five types of DMs are analyzed: Equal Weight (EWDM), Global Weight (GWDM), Global Weight Optimized (GWDM_opt), Item Weight (IWDM), and Item Weight Optimized (IWDM_opt). Robustness is assessed by analyzing how calibration scores of DMs change when individual experts or calibration questions are removed. Furthermore, we look at the Robustness by Distribution Ratio (RDR). Discrepancy is analyzed by comparing the information score obtained from the uniform background measure and the information scores obtained using other DMs as background measures.

The analysis shows that the more experts there are in a study, the more robust the DMs become. For some DMs, robustness also improves with more calibration questions, while for others, no clear trend is observed. Overall, the IWDM, IWDM_opt, and GWDM_opt are found to be more robust, while the EWDM and the GWDM are the least discrepant. The thesis concludes with recommendations for choosing the best-performing, most robust, or least discrepant DM depending on the number of experts and calibration questions available in a study.

Contents

1 Introduction

Imagine a new virus is spreading across the world. Researchers are faced with important questions: How long does the infection last? How long is an infected person contagious? How long is someone immune after recovery? Since the virus is new, the researchers do not have official data to answer these questions with 100% certainty. This is where models such as Cooke's Classical Model (CM) of Structured Expert Judgment (SEJ), the SEIR model, the Delphi technique, or others become valuable. Cooke's Classical Model aggregates assessments from a group of experts to compute a final prediction. Therefore, when no historical data is available, Cooke's Classical Model is one of the methods that is very valuable for predicting answers to questions of interest (Hanea & Nane, 2021).

An example of a study using Cooke's Classical Model is from the World Health Organization (WHO), which investigated the role of different foods and exposure possibilities in preventing foodborne illnesses (Hoffmann et al., 2017). However, CM is not only limited to public health, the model has also been applied to studies from other fields such as climate change. One of them is a study published in PNAS that uses the model to predict sea level rise (Bamber, Oppenheimer, Kopp, Aspinall, & Cooke, 2019). Since many different fields of study use the CM (TU Delft OpenCourseWare, 2020), it is important to understand how well the model performs. Thus, this research focuses on a robustness and discrepancy analysis of Cooke's Classical Model in a expert judgment study. With robustness we refer to the stability of the solutions when the input data changes, such as outliers. Discrepancy measures how informative a model's solution is compared to a reference distribution.

We begin by explaining how the CM works and what experts assessments look like in section 2. This section also introduces the concept of Decision Makers (DMs) and the different methods to construct them. Furthermore, the scoring methods, calibration score and information score, are formulated and explained. In section 3, we explain what robustness means in the context of the CM. We analyze how DM scores change when individual experts or questions are removed from the data. We also introduce an additional robustness metric, the Robustness by Distribution Ratio (RDR), explained in section 3.3. Next, in section 4, we introduce the concept of discrepancy, focusing on the role of the background measure in computing the information scores. We compare the information scores of DMs using various background measures to assess how different assumptions affect the outcome. Finally, in section 5, we present recommendations on which DM to choose based on the number of experts and calibration questions available. These recommendations are based on the best scoring, most robust, and most discrepant DMs.

2 Cooke's Classical Model

Cooke's Classical Model (CM) is a model of Structured Expert Judgment (SEJ) that is used to combine multiple experts' opinions into a single prediction for a specific question of interest. An example of a question of interest is: What is the percentage of unvaccinated children in Europe in 2030? In order to answer this question with a proper prediction, you need to find and recruit multiple experts in the field of study of the question of interest. In selecting experts, you look at criteria such as publication records, previous attribution in studies, experience, and expertise. These experts may include epidemiologists, scientists, engineers, and professionals from government, academia, or non-governmental organizations (Beshearse, 2021). Each expert is asked to estimate the question of interest by giving three quantiles: 5%-, 50%$-$ and 95%$-$percentile. For instance, if an expert provides the answer 0.5%, 2.5% and 7%, for the percentiles respectively, then the expert predicts a 5% chance that the actual percentage of unvaccinated children in Europe will be below 0.5%. A 50% chance it will be below 2.5% and a 50% chance it will be above 2.5%. Lastly, he predicts a 5% chance that the percentage of unvaccinated children will exceed 7%.

However, how can we be certain that these experts give reliable predictions? This is where the Classical Model uses calibration questions. Alongside the question of interest, each study includes several calibration questions from the same field. These questions are based on data from official sources, often unpublished at the time of the study. Experts are asked to answer with 5%-, 50%$-$ and 95%$-$percentile estimates for each calibration question, just as they do for the question of interest. While the researcher knows the realizations of the calibration questions, the experts do not. This allows the researcher to objectively evaluate the accuracy and calibration of each expert's predictions against the known answers to the calibration questions. Each expert receives a calibration score based on their performance on these questions, this will be explained in more detail in Section 2.1. In addition to calibration, the informativeness of each expert's predictions is also evaluated, which is discussed in Section 2.2. Then, Section 2.3 introduces the combined score, which combines both calibration and informativeness. This combined score is used to create the final predictions, which will be further discussed in section 2.4. (Hanea & Nane, 2021).

But first, we will explain how the 5%-, 50%$-$ and 95%$-$percentile are used to compute the cumulative distribution function (CDF) of each expert, for each question. Suppose that we have N experts, denoted as $e_1, e_2, \ldots, e_N$. Let the assessment of expert i for a question be denoted as q_5^i, q_{50}^i and q_{95}^i for the 5%-, 50%$-$ and 95%$-$percentile, respectively. We define the range $[L, U]$ based on the percentiles and the realization of the question by:

$$L = \min_{1 \leq i \leq N} \left\{ q_5^i, \text{ realization} \right\} \tag{2.1}$$

$$U = \max_{1 \leq i \leq N} \left\{ q_{95}^i, \text{ realization} \right\} \tag{2.2}$$

Here L is the smallest value of all expert's 5%-percentiles and the realization. U is the largest value of all expert's 95%-percentiles and the realization of this question. From this interval, we define the intrinsic range as:

$$[L^*, U^*] = [L - k \cdot (U - L), U + k \cdot (U - L)] \tag{2.3}$$

The constant k is the overshoot, usually set to 10%. Each expert's distribution is then defined over this intrinsic range, and probability is assigned uniformly within each inter-percentiles interval. For example, if an expert provides the following assessment $q_5^i = 0.5$ and $q_{50}^i = 2.5$, then the probability that the true value is 1 is the same as the probability that this value is 2. This uniformity ensures that the (CDF) is piecewise linear. It starts at the lower bound of the intrinsic range, L^*, where the CDF value is zero. At the expert's 5%-quantile value, the CDF reaches probability 0.05. At the 50%-quantile, it increases to 0.5. At the 95%-quantile, it reaches probability 0.95. Finally, at the upper bound of the intrinsic range, U^*, the CDF reaches 1. Another way to describe this distribution function is (Hanea & Nane, 2021):

$$F_i(x) = \begin{cases} 0, & \text{for } x < L^* \\ \frac{0.05}{q_5^i - L^*} \cdot (x - L^*), & \text{for } L^* \leq x < q_5^i \\ \frac{0.45}{q_{50}^i - q_5^i} \cdot (x - q_5^i) + 0.05, & \text{for } q_5^i \leq x < q_{50}^i \\ \frac{0.45}{q_{95}^i - q_{50}^i} \cdot (x - q_{50}^i) + 0.5, & \text{for } q_{50}^i \leq x < q_{95}^i \\ \frac{0.05}{U^* - q_{95}^i} \cdot (x - q_{95}^i) + 0.95, & \text{for } q_{95}^i \leq x < U^* \\ 1, & \text{for } x \geq U^* \end{cases} \tag{2.4}$$

To explain it more graphically, we will use an example. Suppose that we have the assessments of three experts for a question with realization 30. The 5%-, 50%- and 95%-quantile are 10,36,50 for expert 1. For expert 2 and 3 we have 20,35,45 and 27,31,36 respectively. Then the CDF's are shown in figure 2.1

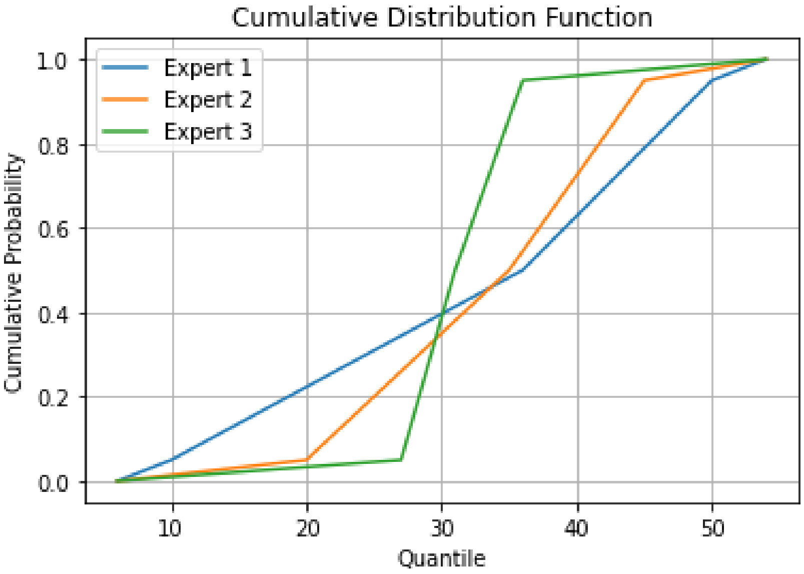


Figure 2.1: Cumulative distribution functions of three experts.

2.1 Calibration Scores

An expert's calibration score shows how well their predictions match real outcomes in the field related to both the calibration questions and the question of interest. In this section, we will explain how this score is computed.

Suppose that we have N experts, denoted as $e_1, e_2, \ldots, e_N$ and M calibration questions, denoted as $Q_1, Q_2, \ldots, Q_M$. Where we denote the prediction of expert i on question j as: $q_5^{i,j}$,$q_{50}^{i,j}$ and $q_{95}^{i,j}$ for the 5%-, 50%$-$ and 95%$-$percentile, respectively. These percentiles divide the prediction range into four inter-percentile intervals. We define the probability vector $p = (0.05, 0.45, 0.45, 0.05)$ that divides the prediction ranges. Next we will define the empirical distribution for expert i. Let x_j be the actual value, or realization, of calibration question Q_j. Then for each expert, we count the amount of realizations that fall into each of the four inter-percentile intervals of p. The normalized counts for each interval are given by the following formulas:

$$s_1(e_i) = \frac{|\{k \mid x_k \leq q_5^{i,j}\}|}{M} = \frac{1}{M}\sum_{k=1}^{M} \mathbf{1}_{\{x_k \leq q_5^{i,j}\}} \qquad \text{for } j \in \{1, \ldots, M\}, \tag{2.5}$$

$$s_2(e_i) = \frac{|\{k \mid q_5^{i,j} < x_k \leq q_{50}^{i,j}\}|}{M} = \frac{1}{M}\sum_{k=1}^{M} \mathbf{1}_{\{q_5^{i,j} < x_k \leq q_{50}^{i,j}\}} \qquad \text{for } j \in \{1, \ldots, M\}, \tag{2.6}$$

$$s_3(e_i) = \frac{|\{k \mid q_{50}^{i,j} < x_k \leq q_{95}^{i,j}\}|}{M} = \frac{1}{M}\sum_{k=1}^{M} \mathbf{1}_{\{q_{50}^{i,j} < x_k \leq q_{95}^{i,j}\}} \qquad \text{for } j \in \{1, \ldots, M\}, \tag{2.7}$$

$$s_4(e_i) = \frac{|\{k \mid q_{95}^{i,j} < x_k\}|}{M} = \frac{1}{M}\sum_{k=1}^{M} \mathbf{1}_{\{q_{95}^{i,j} < x_k\}} \qquad \text{for } j \in \{1, \ldots, M\}. \tag{2.8}$$

Where we use the indicator function $\mathbf{1}_{\{b \leq x \leq a\}}$, which is equal to 1 if the condition is true, and 0 otherwise.

$$\mathbf{1}_{\{b \leq x \leq a\}} = \begin{cases} 1, & \text{if } b \leq x \leq a, \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

Combining these functions, we obtain $s(e_i) = (s_1(e_i), s_2(e_i), s_3(e_i), s_4(e_i))$ which we call the empirical distribution vector for expert i. Now we can compare this distribution and the probability vector p using the Kullback-Leibler divergence: $I(s(e_i), p)$. The comparison is made in the following formula:

$$2MI\big(s(e_i), p\big) = 2M \sum_{l=1}^{4} s_l(e_i) \ln \frac{s_l(e_i)}{p_l} \tag{2.10}$$

This statistic is distributed as a chi-square random variable with 3 degrees of freedom under the null hypothesis H_{e_i}: the inter-percentile interval containing the actual value for each variable is drawn independently. Finally, the p-value of this statistic is used to score each expert. This score is called the calibration score with the following formula:

$$Cal(e_i) = \text{Prob}\{2MI(s(e_i), p) > r \mid H_{e_i}\} \tag{2.11}$$

Where the value r equal is to statistic 2.10, computed using the realizations of the calibration questions (Hanea & Nane, 2021).

Since the calibration score of an expert represents a probability, we know that each calibration score lies within the interval between zero and one (Grimmett & Welsh, 2014). A higher calibration score generally indicates a better calibrated expert. However, calibration scores can not always be directly compared between all experts. For instance, if we have two expert's: e_1 and e_2. In figure 2.2 we see the assessments of each expert for ten calibration questions. Each horizontal line corresponds to one calibration question. The left and right endpoint of a line represent the 5%- and 95%-percentile, respectively. The crosses indicate the realizations of the questions. If a realization falls within the $5\% - 95\%$ interval, then the cross is colored blue. However, if the realization falls outside this interval, then the cross is colored yellow. The blue circle marks the 50%-percentile. Notice that expert 1 and 2 give almost identical assessments. However, expert 1 has four times that their 50%-percentile coincides with the realization. To compute the calibration scores, we count the positions of the realizations compared to the percentile intervals. For expert 1 that is, for realization x_k we once have that $x_k \leq q_5^1$. We have six times that $q_5^1 < x_k \leq q_{50}^1$. Next, we have twice that $q_{50}^1 < x_k \leq q_{95}^1$. Finally, we have one time that $x_k > q_{95}^1$. This results in the vector $s(e_1) = (0.1, 0.6, 0.2, 0.1)$. Using the same counting method, we find for expert 2 the following $s(e_2) = (0.1, 0.4, 0.4, 0.1)$. These empirical distribution vectors give the calibration scores 0.39 for expert 1, and 0.83 for expert 2. One might expect expert 1 to have a higher calibration score, as their assessments appear to be better calibrated. However, this is not the case. This example shows that calibration scores are not always directly comparable. Therefore, you should look at a threshold of 0.05. If one expert's calibration score is below 0.05, and another expert's calibration score is above 0.05, then we can conclude that the second expert is better calibrated (Hanea & Nane, 2021).

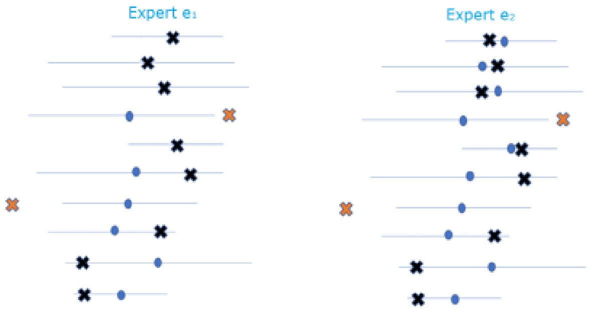


Figure 2.2: Assessments of two experts.

2.2 Information Scores

While the calibration of expert's is important, it is not the only quality that we look for. We also want the assessments from expert's to be informative. To assess the informativeness of an expert, we examine the expert's probability distribution and compare it to the background measure, which is usually the uniform distribution. In a uniform distribution, every outcome within an interval is equally likely, thus it has the same probability. For example, in the interval [1,3], the probability of picking 1 is the same as picking 3. In Cooke's Classical Model as background measure for calculating information scores is the uniform distribution used on the intrinsic range $[L^*, U^*]$ and the function:

$$U(x) = \frac{x - L^*}{U^* - L^*}, \text{ for } L^* \leq x \leq U^* \tag{2.12}$$

With this function we can calculate the probability that a value falls within each interpercentile interval of the background measure. For example, the probability that the realization of question j falls between L^* and the 5th percentile $q_5^{i,j}$ of expert i is:

$$r_1 = U(q_5^{i,j}) - U(L^*) = \frac{q_5^{i,j} - L^*}{U^* - L^*}, \qquad \text{for } x \in [L^*, q_5^{i,j}] \tag{2.13}$$

Similarly, for the remaining intervals we have:

$$r_2 = U(q_{50}^{i,j}) - U(q_5^{i,j}) = \frac{q_{50}^{i,j} - q_5^{i,j}}{U^* - L^*}, \qquad \text{for } x \in (q_5^{i,j}, q_{50}^{i,j}], \tag{2.14}$$

$$r_3 = U(q_{95}^{i,j}) - U(q_{50}^{i,j}) = \frac{q_{95}^{i,j} - q_{50}^{i,j}}{U^* - L^*}, \qquad \text{for } x \in (q_{50}^{i,j}, q_{95}^{i,j}], \tag{2.15}$$

$$r_4 = U(U^*) - U(q_{95}^{i,j}) = \frac{U^* - q_{95}^{i,j}}{U^* - L^*}, \qquad \text{for } x \in (q_{95}^{i,j}, U^*]. \tag{2.16}$$

Cooke's model assumes that the probability that the realization of the questions falls below the 5th percentil q_5 is 0.05. For the other inter-quantile intervals we have the probabilities 0.45,0.45,0.05, respectively. This reflects in the probability vector p. Therefore, we also know that the distribution functions for all experts should have these values at the corresponding quantiles. Which we can also see directly from figure 2.1. Let $F(\cdot)$ be the expert's cumulative distribution function, then we have:

$$f_1 = F(q_5^{i,j}) - F(L^*) = 0.05, \tag{2.17}$$

$$f_2 = F(q_{50}^{i,j}) - F(q_5^{i,j}) = 0.45, \tag{2.18}$$

$$f_3 = F(q_{95}^{i,j}) - F(q_{50}^{i,j}) = 0.45, \tag{2.19}$$

$$f_4 = F(U^*) - F(q_{95}^{i,j}) = 0.05. \tag{2.20}$$

To compute the information score of expert i on question j, we then compare the observed probabilities r_1, r_2, r_3, r_4 with the expected ones f_1, f_2, f_3, f_4 in the following formula:

$$I_j(e_i) = \sum_{k=1}^{4} f_k \ln \frac{f_k}{r_k} \tag{2.21}$$

We can then rewrite this to:

$$I_j(e_i) = 0.05 \ln \frac{0.05}{q_5^{i,j} - L^*} + 0.45 \ln \frac{0.45}{q_{50}^{i,j} - q_5^{i,j}} + 0.45 \ln \frac{0.45}{q_{95}^{i,j} - q_{50}^{i,j}} + 0.05 \ln \frac{0.05}{U^* - q_{95}^{i,j}} + \ln(U^* - L^*) \tag{2.22}$$

From equation 2.22 we can see that a higher information score for a calibration question results from a more concentrated distribution from an expert. This indicates more informativeness in the assessment of the expert.

Finally, to get the overall information score of expert i, we simply take the average of the information scores for all questions (Hanea & Nane, 2021):

$$I(e_i) = \frac{1}{M} \sum_{j=1}^{M} I_j(e_i) \tag{2.23}$$

2.3 Combined Scores

In sections 2.1 and 2.2 we have introduced two different methods to objectively score experts: the calibration score and the information score. The calibration score determines the accuracy of an expert, while the information score assesses how informative the assessment of experts are. However, we want experts to be both accurate and informative. To achieve this, we have defined the combined score. The combined score of an expert is their calibration score multiplied by their information score. This score includes a cutoff level α for the calibration score. If the calibration score of an expert is smaller than the cutoff value α, then their combined score is zero. The combined score of expert e_i is calculated as follows:

$$CS(e_i) = Cal(e_i) \cdot I(e_i) \cdot \mathbf{1}_\alpha(Cal(e_i)) \tag{2.24}$$

Where $\mathbf{1}_\alpha(Cal(e_i))$ is the indicator function that is equal to one if $Cal(e_i) \geq \alpha$, and equal to zero otherwise. We note that if there is a significant change in the number of calibration questions, then the value of the calibration score changes quickly. However, the information score is the average of all questions. So a change in the number of questions has less influence on the overall information score of an expert. Therefore, the calibration score will have more influence on the combined score. Cooke's model compares experts by their combined scores, so the model prioritizes calibration scores. Since the model finds it more valuable to listen to experts who are better calibrated than experts who are more informative (Hanea & Nane, 2021).

2.4 Decision Maker

At the beginning of section 2 we explained how each expert's cumulative distribution function is constructed. We now want to aggregate these cumulative distribution functions from all experts for a given question to create a single final distribution function called the Decision Maker (DM) for that question. To do this, weights are assigned to each experts. However, there are several possible methods for assigning these weights. There are Equal Weight, Global Weight and Item Weight Decision Makers. These methods will be discussed in the following subsections.

2.4.1 EWDM

In the Equal Weight Decision Maker (EWDM), all experts receive the same weight. If there are N experts, then each expert's distribution is given a weight of $1/N$. The cumulative distribution function of the EWDM is then defined as:

$$F_{EWDM}(x) = \sum_{i=1}^{N} w_i \cdot F_i(x) \tag{2.25}$$

Since each expert has the same weight, we can rewrite this to:

$$F_{EWDM}(x) = \sum_{i=1}^{N} \frac{1}{N} F_i(x) \tag{2.26}$$

In the previous example, there were three experts. So in the EWDM, each expert is assigned a weight of 1/3. The EWDMs cumulative distribution function is then defined as:

$$F_{EWDM}(x) = \frac{1}{3}F_1(x) + \frac{1}{3}F_2(x) + \frac{1}{3}F_3(x) \tag{2.27}$$

In figure 2.3 the cumulative distribution function of the EWDM is shown.

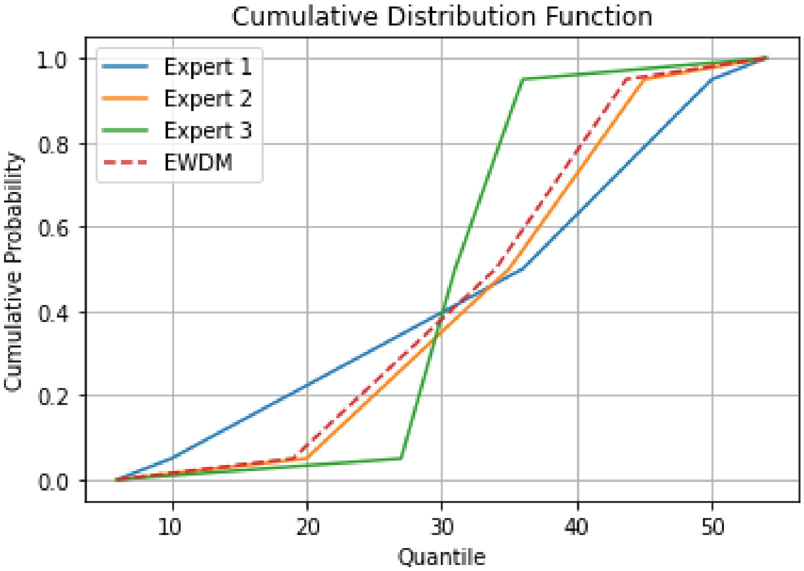


Figure 2.3: Cumulative distribution functions of three experts and the Equal Weight Decision Maker.

2.4.2 GWDM

In the Global Weight Decision Maker (GWDM) the performance of experts on the calibration questions play an important role. The weight that each experts gets assigned is based on their combined score. The weight assigned to expert i is defined as:

$$w_i = \frac{CS(e_i)}{\sum_{k=1}^{N} CS(e_k)} \tag{2.28}$$

This means the better an expert performs on the calibration questions, the higher the weight they receive. Therefore, their assessment have more influence on the DM. The formula for the cumulative distribution function of the GWDM then becomes:

$$F_{GWDM}(x) = \sum_{i=1}^{N} w_i \cdot F_i(x) \tag{2.29}$$

Note that the CDF of the GWDM depends on the cutoff value α. If an expert's calibration score is lower than this cutoff value, then they are assigned a weight of zero. However, the GWDM usually uses $\alpha = 0$. So all experts contribute to the DM (Hanea & Nane, 2021).

The GWDM CDF for the previous example with three experts and $\alpha = 0$ is shown in figure 2.4.

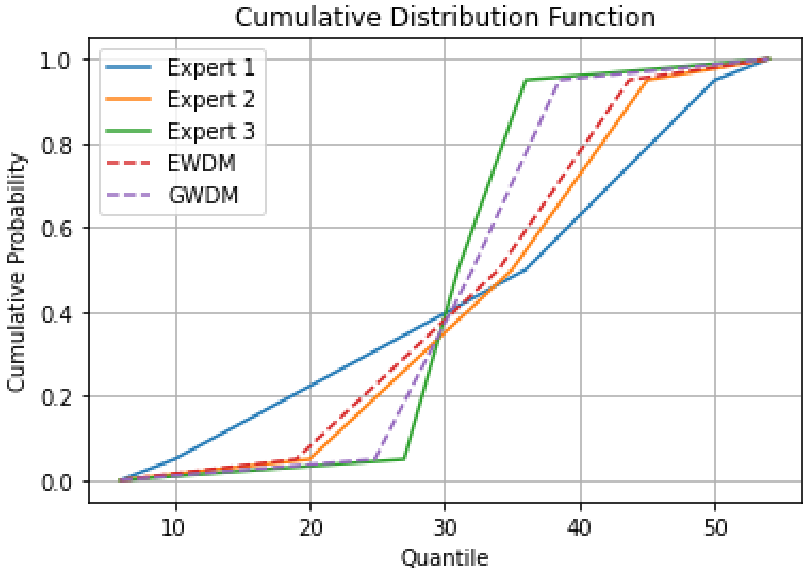


Figure 2.4: Cumulative distribution functions of three experts, the Equal Weight Decision Maker, and the Global Weight Decision Maker.

2.4.3 GWDM Optimized

As mentioned in section 2.4.2, the Global Weight Decision Maker (GWDM) typically uses the cutoff value $\alpha = 0$. However, there are different values for α that can be chosen to change the influence of experts on the DM. For example, if α is equal to the second lowest calibration score, then the worst calibrated expert is excluded from computing the GWDM. Or if α is equal to the highest calibration score, then only the best calibrated expert is left over. Thus, the GWDM becomes identical to the distribution of the best calibrated expert (Hanea & Nane, 2021).

Generally, there are three common ways to choose α. First we have a statistical threshold where α is chosen what is considered to be an acceptable calibration value. Typically this statistical threshold would be $\alpha = 0.05$ or $\alpha = 0.01$. Next we have the inclusive threshold where α is low enough such that all experts contribute to the DM. Thus, α is set lower than the smallest calibration score. The last option is the optimized threshold. Where α is chosen such that the combined score of the GWDM is maximized. This DM is called the GWDM optimized (Cooke, n.d.).

2.4.4 IWDM

As explained in section 2.2, each expert receives an information score for every question they answer. The overall information score of an expert is then the average of all individual information scores. This overall information score is then used to determine the combined score of an expert. Which is then used to construct the GWDM.

However, we can also compute a combined score for each question individually. The combined score of expert i for calibration question j is computed using the following formula:

$$CS_j(e_i) = Cal(e_i) \cdot I_j(e_i) \cdot \mathbf{1}_\alpha(Cal(e_i)) \tag{2.30}$$

The weight assigned to expert i for question j is then determined by:

$$w_i^j = \frac{CS_j(e_i)}{\sum_{k=1}^{N} CS_j(e_k)} \tag{2.31}$$

These weights are also known as item weights. In the GWDM each, expert gets assigned one weight. In contrast, the Item Weight Decision Maker (IWDM) assigns a vector of weights to each expert. Each weight in the vector representing one calibration question. Item weights are useful when experts know more about certain calibration questions, and less about other questions. This allows experts to be up-weighted or down-weighted for certain questions (Cooke, n.d.).

The cumulative distribution function for calibration question j, where $F_i^j(x)$ is the CDF of expert i for that question, is defined as:

$$F_{IWDM}^j(x) = \sum_{i=1}^{N} w_i^j \cdot F_i^j(x) \tag{2.32}$$

To determine the cumulative distribution function for a question of interest l, the information score for that question is first calculated for each expert. The information score of expert i for question l is denoted as $I_l(e_i)$. Using this score and the calibration score of the calibration questions, the combined score for question l is computed by:

$$CS_l(e_i) = Cal(e_i) \cdot I_l(e_i) \cdot \mathbf{1}_\alpha(Cal(e_i)) \tag{2.33}$$

With these new combined scores, the weights of the experts for question l are computed using equation 2.31. Finally, the cumulative distribution function of the question of interest l, using IWDM, is then defined as:

$$F_{IWDM}^l(x) = \sum_{i=1}^{N} w_i^l \cdot F_i^l(x) \tag{2.34}$$

2.4.5 IWDM Optimized

Similar to GWDM Optimized, the IWDM can also be optimized by varying the cutoff value α. The DM with α value such that the combined score of the IWDM is maximized, is called the IWDM optimized.

3 Robustness

Cooke's Classical Model (CM) generates Decision Makers (DMs) based on datasets consisting of assessments from multiple experts. Each DM also receives a calibration score and an information score. But what happens to the DM and its scores when outliers are present in the dataset? How can we determine whether the model's results are robust? Robustness in a model refers to the ability to keep stable results when there is uncertainty in the data, such as outliers (Mia Hubert, 2004). For CM, robustness means that the calibration score of a DM is relatively immune to changes in the dataset.

Let's consider an example. Suppose the Global Weight Decision Maker (GWDM) has a calibration score of 0.75 when using dataset A. However, when using dataset B, the calibration score of the GWDM becomes 0.2. Now consider the Item Weight Decision Maker (IWDM). Suppose that it has a calibration score of 0.1 with dataset A and 0.15 with dataset B. In this case, the IWDM has lower calibration scores when both datasets are used respectively. However, the difference of the two scores for the IWDM is smaller than that of the GWDM. This suggests that the IWDM, while performing worse than the GWDM, is less sensitive to changes in the used dataset. Thus, the IWDM would be more robust than the GWDM. This example also shows that a robust solution is not always the optimal or best solution (Mehdi, 2022).

In this study, we analyze a dataset consisting of 49 different studies. Each study includes between 4 and 48 experts, and between 7 and 21 calibration questions. All analyzes were performed on a Windows 10 computer, with R version 4.5.1 and RStudio version 2025.5.1.513. The model used in this analysis is developed by Tina Nane, associate professor of Applied Probability at Delft University of Technology, in October 2022, to reproduce Cooke's Classical Model. We then adapted and extended this model in RStudio to generate the results presented in this research.

First, the calibration scores for all experts in each study are calculated. These are presented in a boxplot in figure 3.1, with the study name on the x-axis and the calibration scores on the y-axis. Next, the DMs are computed and scored. The calibration scores of each DM across all studies are shown in figure 3.1 as colored dots.

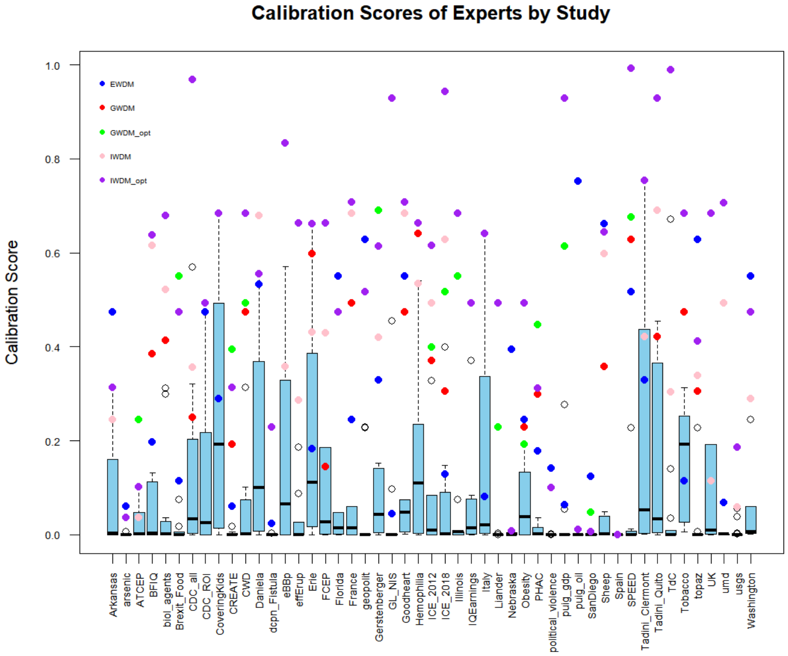


Figure 3.1: Boxplots of the calibration scores of all experts in a study. The calibration scores of all Decision Makers for each study are represented as the colored dots.

To assess robustness, we apply the following algorithm, which returns the calibration scores after excluding individual items from the dataset. This allows us to analyze the scores changes.

Algorithm 1 Computing Calibration Scores Excluding Items for Robustness Analysis

```
Input: Original dataset $D$
Output: Calibration scores for each Decision Maker (DM) after excluding each item once
Step 1: Compute the Decision Makers (DMs) using the original dataset $D$
Step 2: Take an item $i$ in $D$, and do the following:
    a. Temporarily remove the item from the dataset, creating a modified dataset $D' \leftarrow D \setminus \{i\}$
    b. Recompute the DMs using the modified dataset $D'$
    c. Calculate the calibration scores for the new DMs
    d. Store the calibration scores for the modified dataset $D'$
    e. Add the excluded item $i$ back into the dataset
Step 3: Repeat Step 2 until each item has been excluded once
Step 4: Return the set of calibration scores to assess the robustness of the DMs across all exclusions
```

3.1 Removing Experts

To analyze the robustness of the model, we modify the original dataset by excluding individual items. In this section we will focus on excluding experts from the original data. For each study, we remove one expert from the dataset and then recompute the DMs and their corresponding calibration score using the same model. Once the scores are computed, the expert is added back into the study in the dataset, and the next expert is removed from this study. This process is repeated until every expert in every study has been excluded once. As a result, we obtain a list of calibration scores for each DM in each study, with the length of the total number of experts in that study in the original dataset. Algorithm 1 is then adapted to algorithm 2.

Algorithm 2 Computing Calibration Scores Excluding Experts for Robustness Analysis

```
Input: Original dataset $D$
Output: Calibration scores for each Decision Maker (DM) after excluding each expert once
Step 1: Compute the Decision Makers (DMs) using the original dataset $D$
Step 2: Take an expert $e_i$ in $D$, and do the following:
    a. Temporarily remove the expert from the dataset, creating a modified dataset $D' \leftarrow D \setminus \{e_i\}$
    b. Recompute the DMs using the modified dataset $D'$
    c. Calculate the calibration scores for the new DMs
    d. Store the calibration scores for the modified dataset $D'$
    e. Add the excluded expert $e_i$ back into the dataset
Step 3: Repeat Step 2 until each expert has been excluded once
Step 4: Return the set of calibration scores to assess the robustness of the DMs across all exclusions
```

A boxplot is then created where the calibration scores are visualized for each study in figure 3.2. The calibration score from the original dataset is indicated by a red dot. The white dots represent outliers in the calibration scores after expert removal. On the x-axis, the study names are shown, along with the number of experts in each study in parentheses. The studies are order by increase in number of experts.

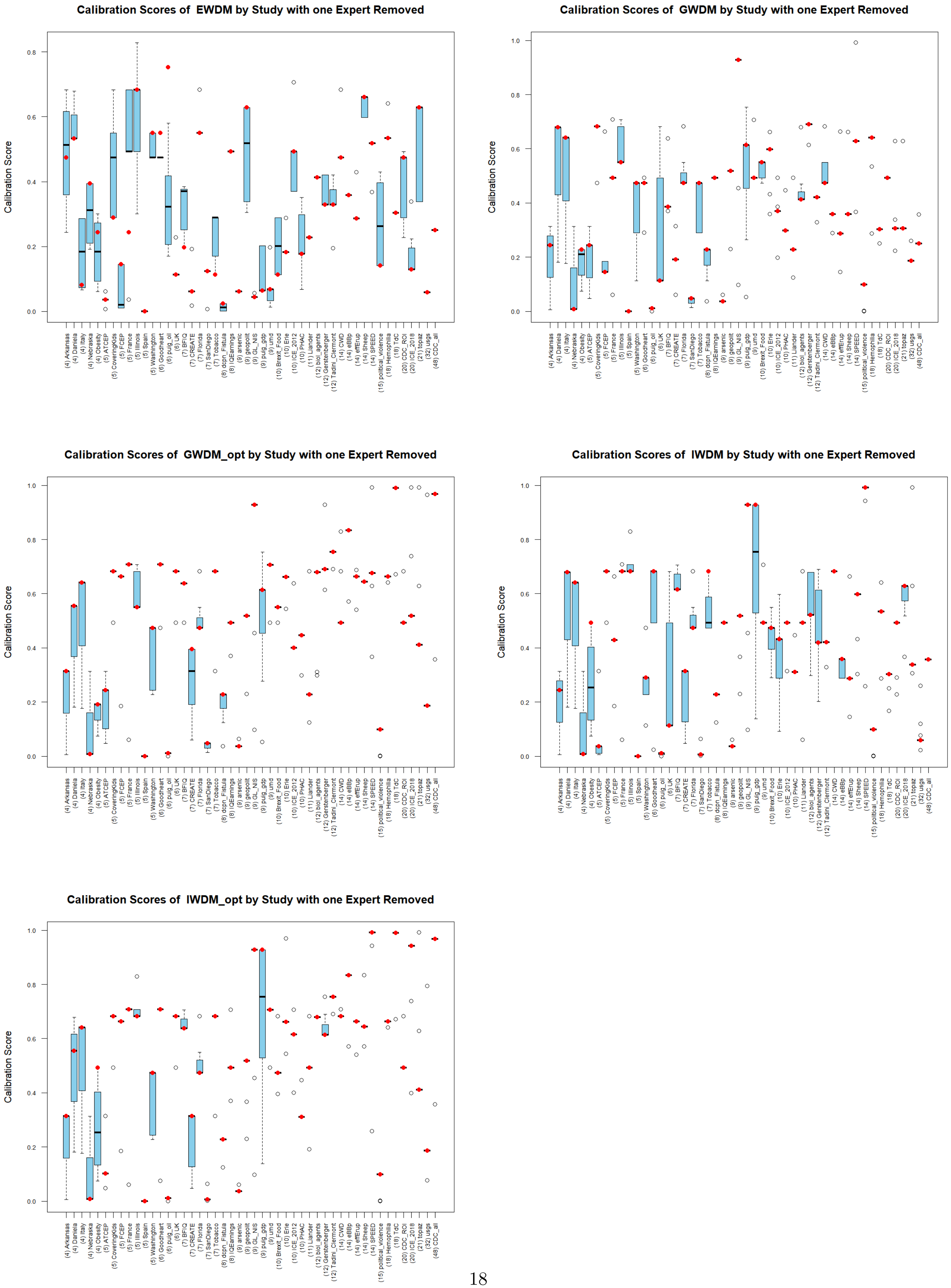


Figure 3.2: Calibration scores of Decision Makers for all studies with expert removal. The red dots represent the calibration score of the original dataset. The white dots are outliers of the calibration scores with expert removal.

Figure 3.2 shows that for the GWDM, GWDM_opt, IWDM and IWDM_opt, the more experts there are in the original dataset, the smaller the boxes become. This means that the calibration scores vary less. Additionally, the boxes are closer to the red dot, representing the original calibration score. This indicates that a larger number of experts leads to a more robust DM. For example, in the study CDC_all, the boxplots for these four DMs consists only of the median, or 50% quantile, which is equal to the calibration score from the original dataset. This suggests that for any expert you remove, the calibration score of the DM remains the same as the original calibration score. This is a clear sign of robustness. However, this trend does not necessarily hold for the EWDM. For studies with 14 to 21 experts, the boxes remain relatively large, indicating greater variability. Only in studies with 32 or 48 experts does the trend of more robustness with more experts clearly hold. In these cases, we again see that the boxplots only consist of the 50% quantile, which is equal to the original calibration scores.

Another way to assess robustness is to only look at the number of experts, rather than analyzing individual studies. Figures 3.3 and 3.4 present boxplots of the calibration scores for all DMs in the original dataset in blue. The green boxplots represent the calibration scores after removing each expert once. The black dots are the outliers of the calibration scores for the given dataset.

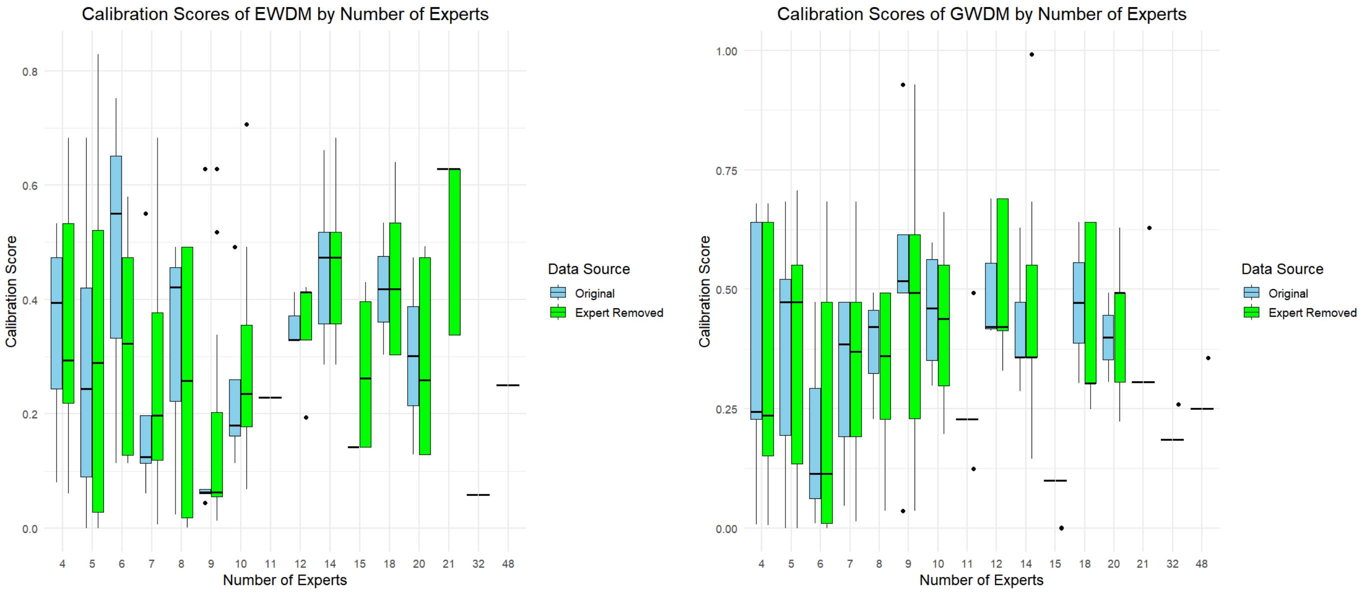


Figure 3.3: Blue boxplots are the calibration scores of Decision Makers from the original dataset. Green boxplots are the calibration scores of Decision Makers from the datasets with experts removed. Black dots represent the outliers in the calibration scores.

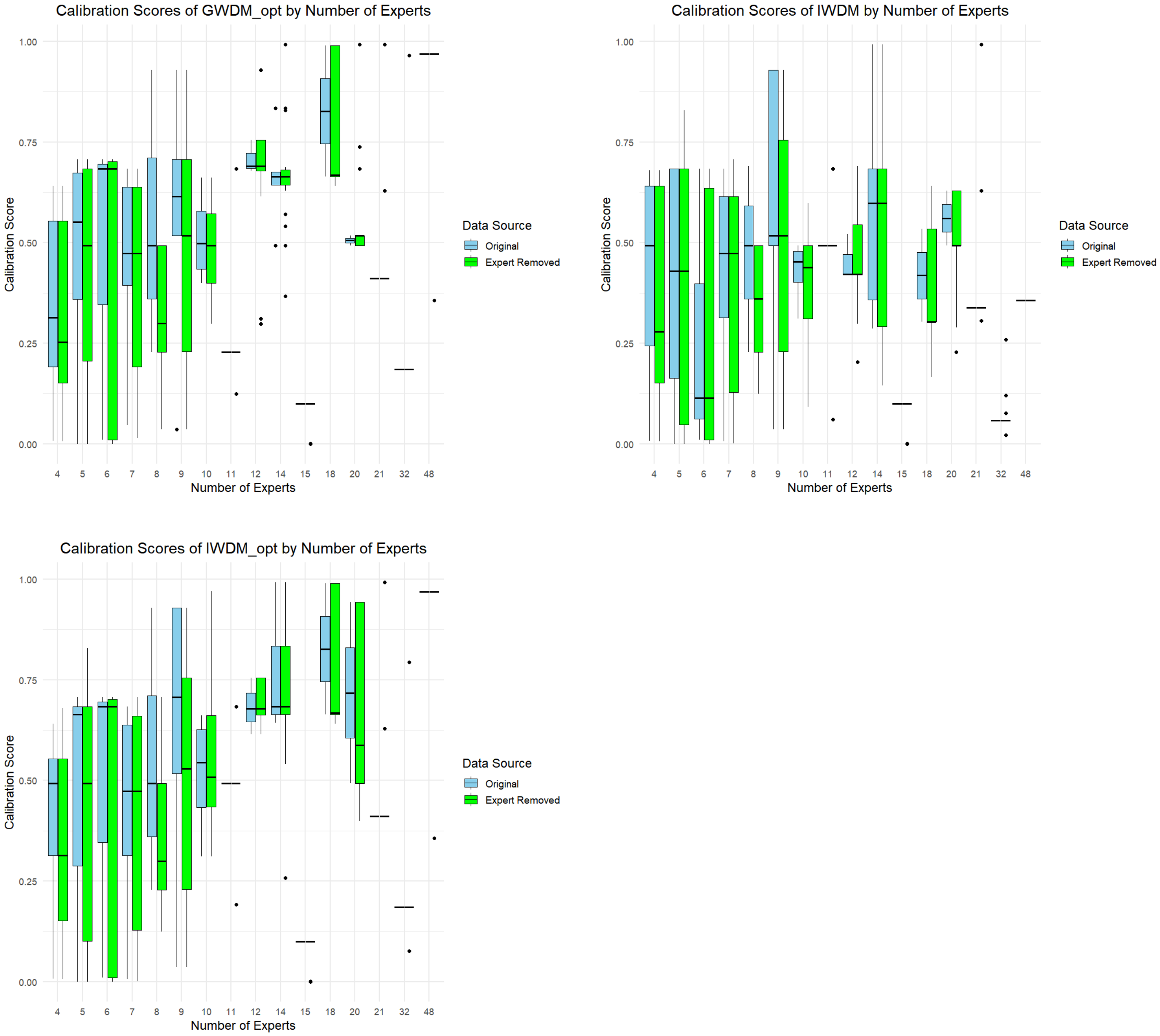


Figure 3.4: Blue boxplots are the calibration scores of Decision Makers from the original dataset. Green boxplots are the calibration scores of Decision Makers from the datasets with experts removed. Black dots represent the outliers in the calibration scores.

In figures 3.3 and 3.4 it is clear that for the GWDM, GWDN_opt, IWDM and IWDM_opt once the number of expert reaches 21 or more, both boxes become very small, indicating low variabillity in calibration scores. Additionally, the two boxes, representing the original dataset and the altered dataset, become almost identical. Which suggests that the expected calibration scores of the DM remain the same. Thus, the calibration score is stable. This indicates robustness. A similar pattern is observed for the EWDM, however in this case robustness is seen from 32 experts onward.

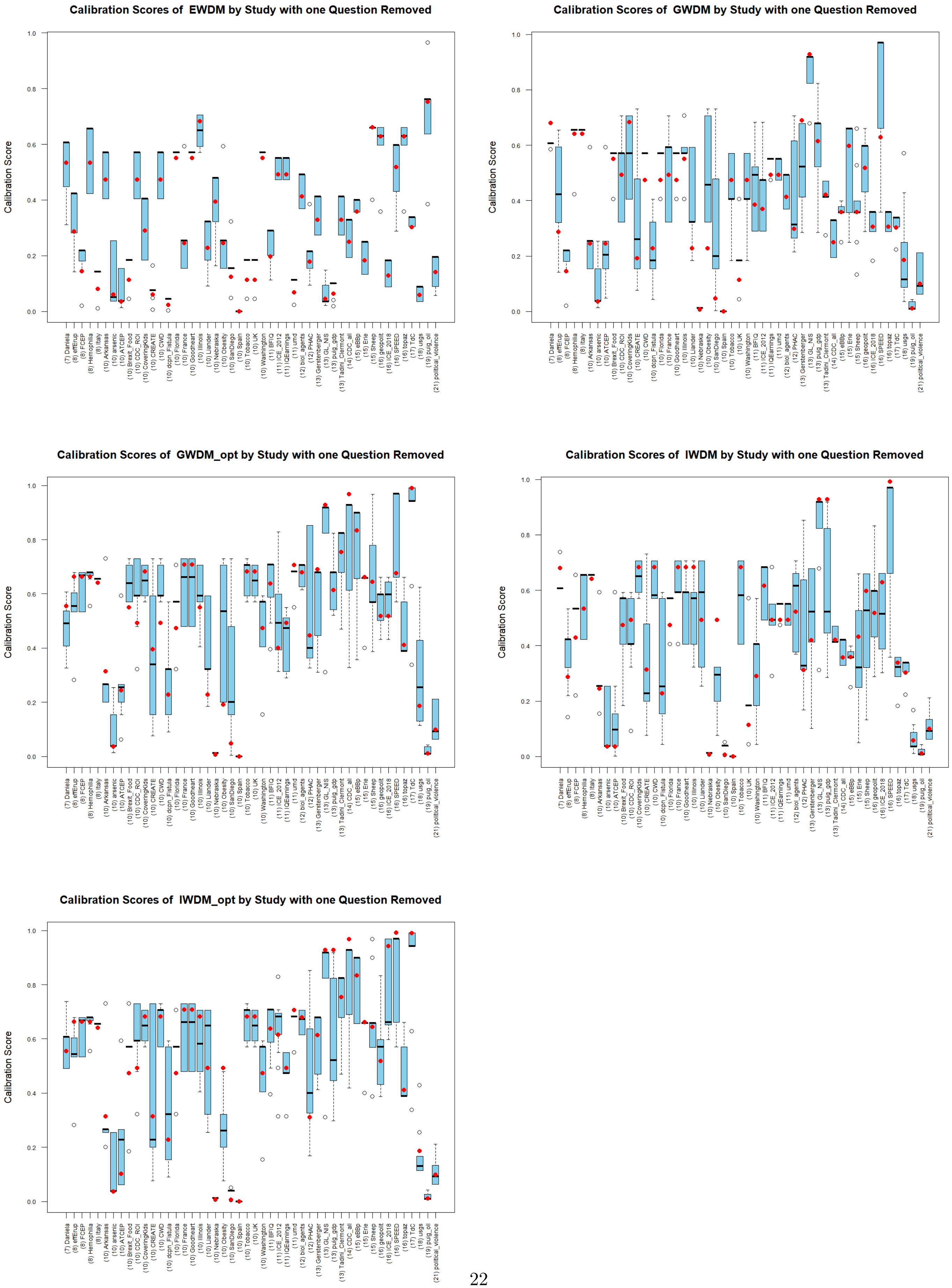
3.2 Removing Calibration Questions

The total number of experts in a study is not the only variable that can influence robustness. In this section, we analyze the influence of the number of calibration questions on the robustness of the model. That is done in a similar way to the expert removal process used in section 3.1. Instead of removing all experts once, each calibration question is removed once from the dataset. Starting with the original dataset, one calibration question is removed. Then the calibration scores of the DMs computed using the modified dataset. After that, the removed calibration question is added back to the dataset, and the next calibration question is removed. This is repeated until every calibration question in each study has been removed once. Algorithm 1 is then adapted to algorithm 3.

Algorithm 3 Computing Calibration Scores Excluding Calibration Questions for Robustness Analysis

```
Input: Original dataset $D$
Output: Calibration scores for each Decision Maker (DM) after excluding each calibration score once
Step 1: Compute the Decision Makers (DMs) using the original dataset $D$
Step 2: Take an calibration question $i$ in $D$, and do the following:
    a. Temporarily remove the question from the dataset, creating a modified dataset $D' \leftarrow D \setminus \{i\}$
    b. Recompute the DMs using the modified dataset $D'$
    c. Calculate the calibration scores for the new DMs
    d. Store the calibration scores for the modified dataset $D'$
    e. Add the excluded question $i$ back into the dataset
Step 3: Repeat Step 2 until each calibration question has been excluded once
Step 4: Return the set of calibration scores to assess the robustness of the DMs across all exclusions
```

The resulting boxplots of the calibration scores for all DMs are shown in figure 3.5. On the x-axis, the names of the studies are shown with the number of calibration question in parentheses. The studies are arranged in increasing order of calibration questions. The red dots represent the calibration score of each study and DM in the original dataset. The white dots represent outliers in the calibration scores after question removal.



Calibration Scores of EWDM by Study with one Question Removed

Calibration Scores of GWDM by Study with one Question Removed

Calibration Scores of GWDM_opt by Study with one Question Removed

Calibration Scores of IWDM by Study with one Question Removed

Calibration Scores of IWDM_opt by Study with one Question Removed

Figure 3.5: Calibration scores of Decision Makers for all studies with calibration question removal. The red dots represent the calibration score of the original dataset. The white dots are outliers of the calibration scores with calibration question removal.

When removing experts, it was clear that in studies with the largest number of experts, the boxplots for the modified dataset were significantly smaller, indicating greater robustness. However, when removing calibration questions, we see less effect. For the IWDM and IWDM_opt, we observe in figure 3.5 that when there are 17 or more calibration questions, the boxes become smaller and align more with the red dot. However, for the EWDM, GWDM and GWDM_opt we do not see a significant trend. In figures 3.6 and 3.7 we have the boxplots of the calibration scores against the number of calibration questions for both the original dataset and the modified dataset. Once again, for the IWDM and IWDM_opt, the boxplots become smaller and more concentrated around the red dot from 17 questions onward. For the EWDM however, we see this same trend, boxes become more narrow from 17 questions onward compared to studies with fewer calibration questions. For the GWDM and GWDM_opt no clear trend is seen in these new figures.

Thus, we can conclude that a higher number of calibration questions improves robustness in the IWDM and IWDM_opt, and to some extent also for the EWDM. However, the number of experts appears to have a greater influence on the robustness of the model than the calibration questions. This is demonstrated by the narrower boxplots resulting from expert removal, compared to the subtle changes in the boxplots when calibration questions are removed.

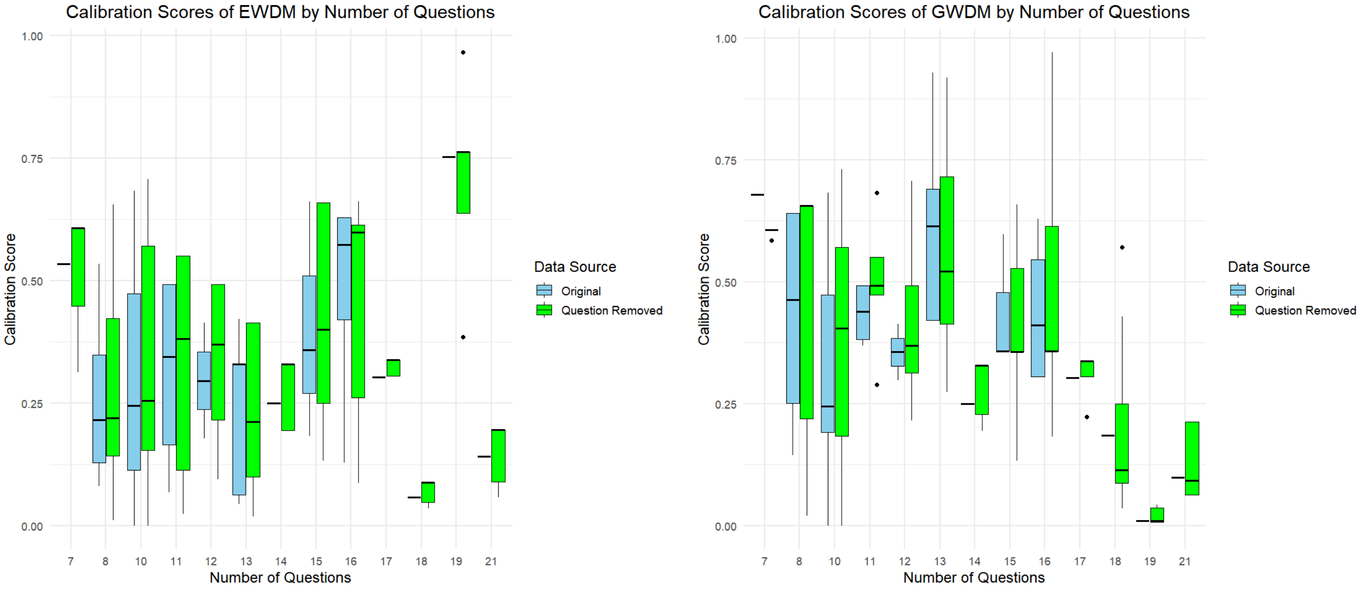


Figure 3.6: Blue boxplots are the calibration scores of Decision Makers from the original dataset. Green boxplots are the calibration scores of Decision Makers from the datasets with calibration questions removed. Black dots represent the outliers in the calibration scores.

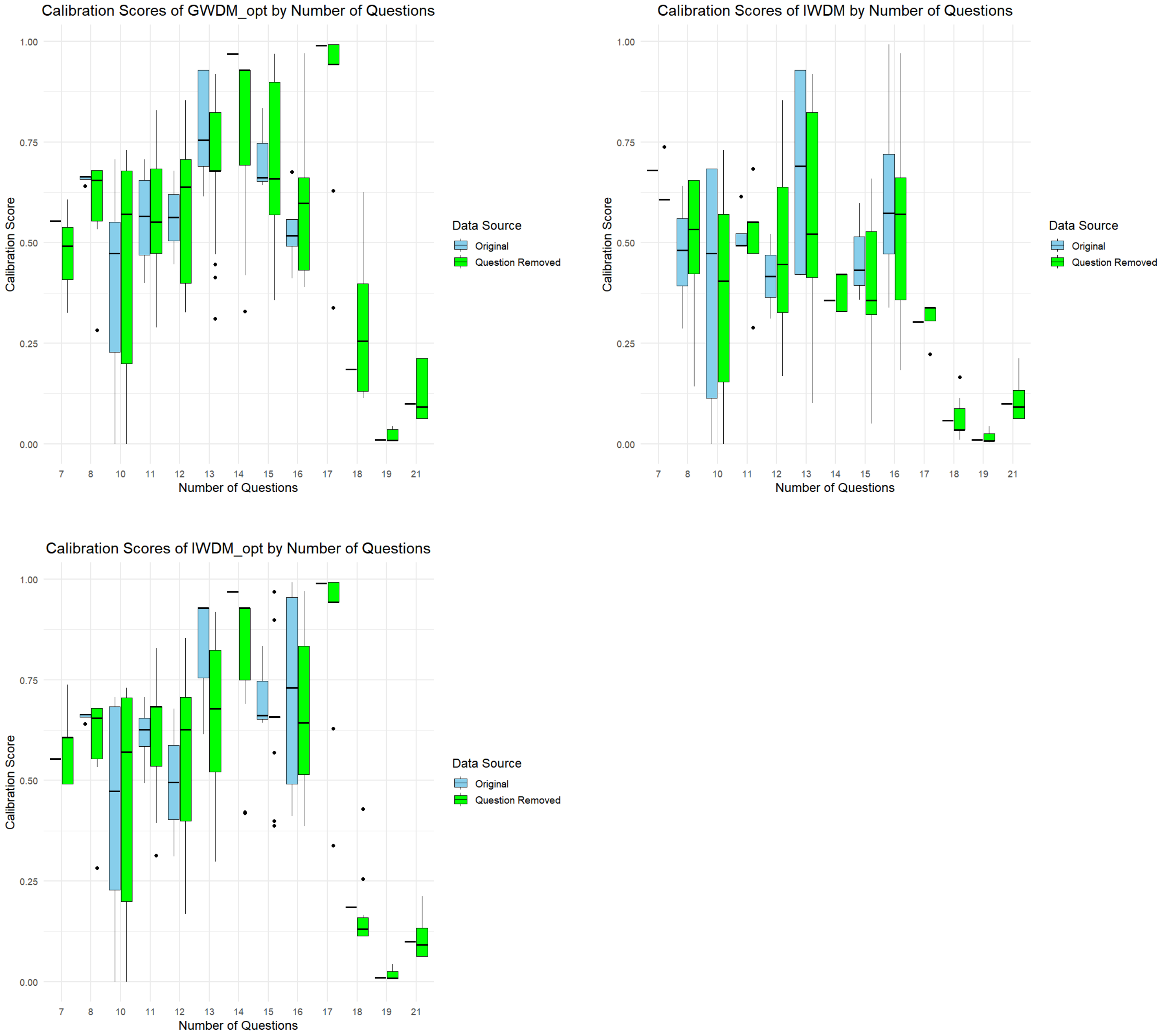


Figure 3.7: Blue boxplots are the calibration scores of Decision Makers from the original dataset. Green boxplots are the calibration scores of Decision Makers from the datasets with calibration questions removed. Black dots represent the outliers in the calibration scores.

3.3 Robustness by Distribution Ratio Metric

In the previous sections, we examined the variation in calibration scores and the expected calibration score when removing expert or questions, compared to the original calibration scores. In this section, we will introduce a different method to assess robustness: the Robustness by Distribution Ratio metric.

The Robustness by Distribution Ratio metric is also known as the RDR metric. This metric evaluates the robustness of a model by combining standardized performance metrics into a single ratio. The metrics used in RDR are: R2, Root Mean Squared Error (RMSE), and Dynamic Time Warping (DTW). R2 quantifies the proportion of variance that is caused by the predictor variable. The RMSE calculates the average distance between predicted and observed values of the model. A downside is that this metric is sensitive to outliers and large errors. The DTW can compare datasets with different lengths (Sarkar, 2023).

3.3.1 Methodology

The RDR (Robustness by Distribution Ratio) metric assesses the robustness of a model by analyzing the relative distribution of calibration scores in the modified dataset compared to the original dataset. In this study, the original data point is the calibration score of a Decision Maker (DM) from a given study. We then compare this to the calibration scores obtained when experts or calibration questions are removed from that study. This creates a new dataset consisting of both the original and modified calibration scores. Each calibration score in this data set is ranked according to its value. The lowest value receives rank 1, the second lowest value rank 2, etc. If there is a tie, then average rank is used. These ranks are then divided by the rank of the original calibration score to obtain the RDR values. The RDR metric is then computed as the average of the absolute difference between the RDR values and the RDR value of the original calibration score. A lower RDR metric indicates that the calibration scores from the modified dataset are consistently close to the original score and evenly distributed around it. Thus, a model with low RDR metric is considered to be more robust (Sarkar, 2023).

In other words, we can describe the RDR metric with the following formulas. Where $S = \{s_1, s_2, \ldots, s_n, s_{org}\}$ is the dataset with the calibration scores of the DM if there are n experts in the study from the original dataset. Each s_i represents the removal of one expert for $i \in \{1, \ldots, n\}$, and s_{org} is the original calibration score with all experts. Then define $R(s_j)$ to be the rank of s_j in this dataset with $j \in \{1, \ldots, n, org\}$. The RDR value of s_j is then:

$$RDR(s_j) = \frac{R(s_j)}{R(s_{org})} \tag{3.1}$$

The RDR metric for this DM and this study is then defined as:

$$\begin{aligned} RDR \text{ metric } &= \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{R(s_i)}{R(s_{org})} - \frac{R(s_{org})}{R(s_{org})} \right| \\ &= \frac{1}{n} \cdot \sum_{i=1}^{n} |RDR(s_i) - 1| \qquad (3.2) \end{aligned}$$

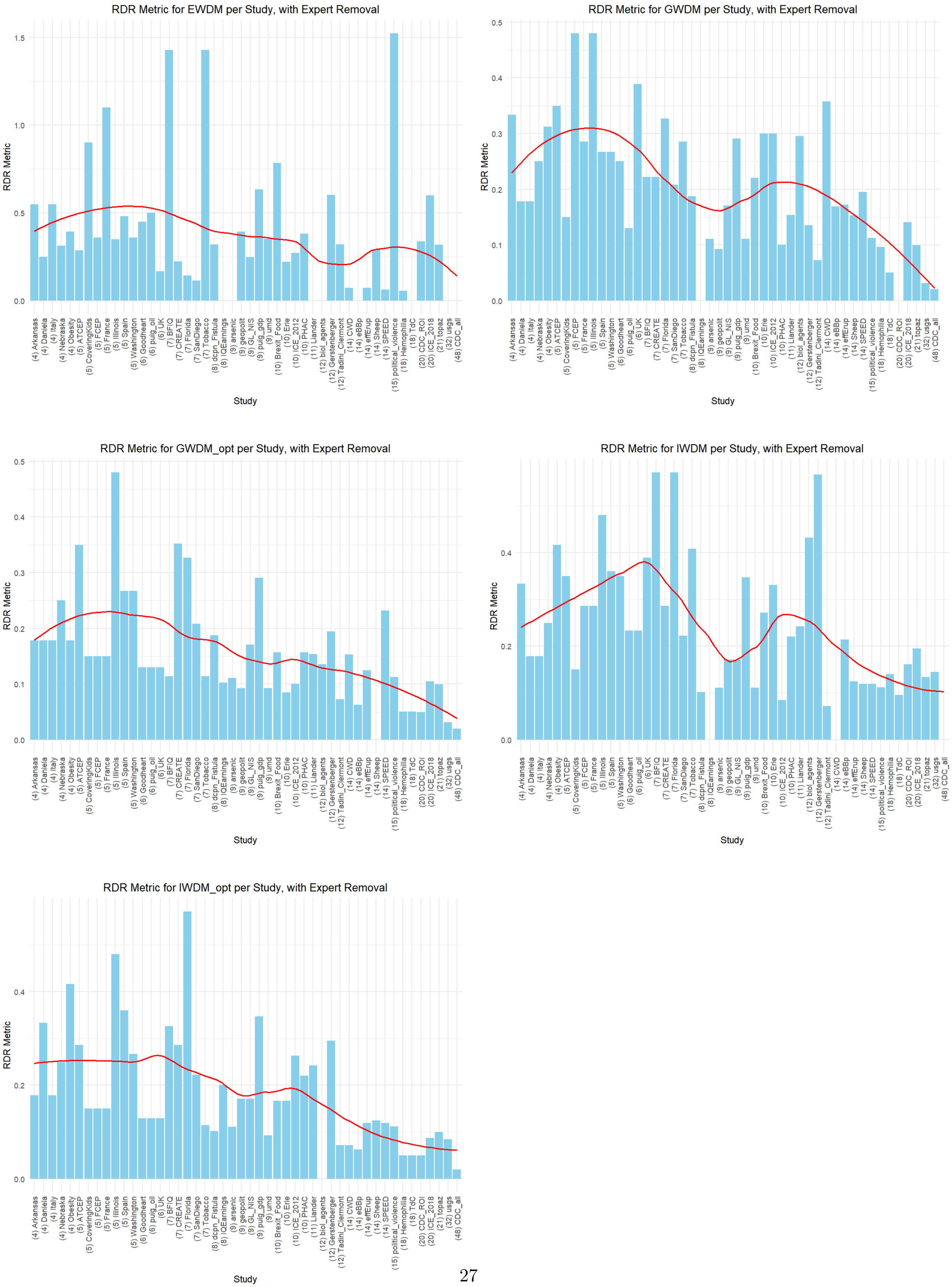
These formulas are based on experts. However, it is also possible to do this for the calibration questions. The vector S then has the calibration scores when calibration questions are removed, and n will become the total number of calibration questions in that study.

To make it more clear, we will give an example. Suppose that a study has 4 experts and the calibration score of the original GWDM is 0.75. Suppose that when removing each expert once, the calibration scores of the GWDM become 0.7,0.9,0.6,0.95, respectively. Then we have $S = \{0.7, 0.9, 0.6, 0.95, 0.75\}$ and $R = \{2, 4, 1, 5, 3\}$. For the RDR metric we have:

$$\begin{aligned} RDR \text{ metric } &= \frac{1}{n} \cdot \sum_{i=1}^{n} |RDR(s_i) - 1| && (3.3) \\ &= \frac{1}{4} \cdot \sum_{i=1}^{4} \left| \frac{R(s_i)}{R(s_{org})} - 1 \right| && (3.4) \\ &= \frac{1}{4} \cdot \sum_{i=1}^{4} \left| \frac{R(s_i)}{3} - 1 \right| && (3.5) \\ &= \frac{1}{4} \cdot (\left| \frac{2}{3} - 1 \right| + \left| \frac{4}{3} - 1 \right| + \left| \frac{1}{3} - 1 \right| + \left| \frac{5}{3} - 1 \right|) && (3.6) \\ &= \frac{1}{4} \cdot (\frac{1}{3} + \frac{1}{3} + \frac{2}{3} + \frac{2}{3}) && (3.7) \\ &= 0.5 && (3.8) \end{aligned}$$

3.3.2 Expert Removal

In this section, we examine the RDR metric in context of expert removal. As previously mentioned, the original dataset is the calibration score of the Decision Maker (DM) with all experts included. When an expert is removed, the calibration scores of that same DM are added to the dataset where the ranks are given. Figure 3.8 presents the RDR metrics for each DM. On the x-axis displays the study names, with the number of experts in parentheses. The studies are ordered by increasing number of experts. The y-axis represents the RDR metric values. The red line represents the LOESS curve, which uses local fitting to find a smooth polynomial that captures the underlying trend in the data.



RDR Metric for EWDM per Study, with Expert Removal

RDR Metric for GWDM per Study, with Expert Removal

RDR Metric for GWDM_opt per Study, with Expert Removal

RDR Metric for IWDM per Study, with Expert Removal

RDR Metric for IWDM_opt per Study, with Expert Removal

Figure 3.8: Robustness by Distribution Ratio metric for each Decision Maker for all studies with expert removal. The red lines are fitted smooth curves representing the underlying trends of the data.

The red line in figure 3.8 shows a clear downward trend in the RDR metric for all DMs as the number of experts increases. This indicates that including more experts leads to a more robust model for each DM. To identify the DM that yields the most robust model, we look at figure 3.9. This figure allows us to easily compare different DMs in one study. We can find the optimal DM for a study or a given number of experts. Optimal DM would be the DM with the lowest RDR metric, and thus the highest robustness. For example, in the Arkansas study with 4 experts, the EWDM has the highest RDR metric, indicating the least robust model. In contrast, the GWDM_opt and IWDM_opt have the smallest RDR metrics, this makes them the most robust models of this study.

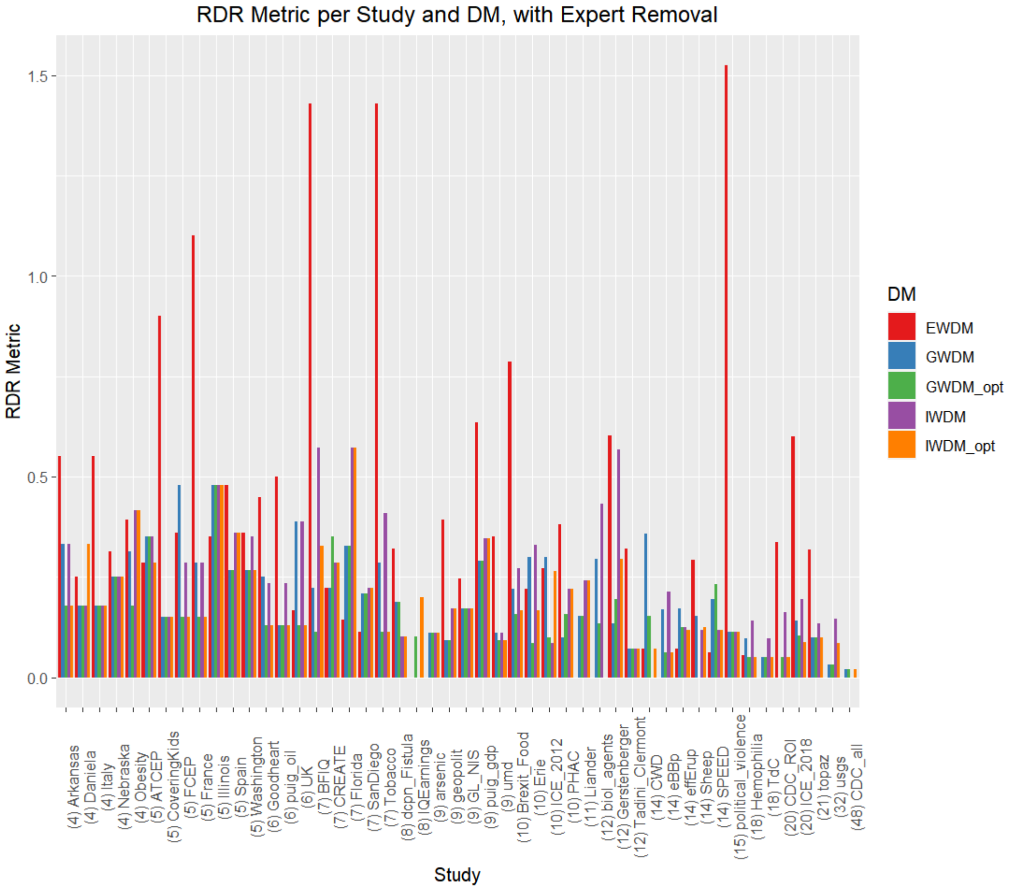


Figure 3.9: Robustness by Distribution Ratio metric with expert removal.

To assess overall robustness, not only individual studies, we evaluate each DM using three statistics: mean, median and standard deviation of the RDR metric across all studies. The mean represents the average RDR metric, thus the sum of all metric values divided by the total number of studies. The median is the middle value when all RDR metrics are sorted in ascending order. The standard deviation calculates the variability of the metric values by taking the square root of the average squared difference from the mean. Let

$(RDR \text{ metric})_j$ be the RDR value for study j, and we have 49 studies. Then formulas are:

$$\text{mean } RDR = \frac{1}{49} \cdot \sum_{j=1}^{49} (RDR \text{ metric})_j \tag{3.9}$$

$$\text{Standard Deviation } RDR = \sqrt{\frac{1}{49-1} \sum_{j=1}^{49} \Big((RDR \text{ metric})_j - \text{mean } RDR \Big)^2} \tag{3.10}$$

To explain the median value in more detail we will give an example. Suppose we have the RDR metrics: 0.5,0.55,0.7,0.85, and 0.88, the median value is 0.7. These three statistics are used to evaluate the RDR metric of each DM. The results are shown in table 3.1. From this table it is clear that if a researcher is looking for a robust model with a low mean RDR, the IWDM_opt is the best choice. If the goal is a low median RDR, then GWDM_opt, IWDM or IWDM_opt are all good choices. Finally, for a low standard deviation GWDM_opt would be the most robust choice.

Decision Maker	**Mean**	**Median**	**Standard Deviation**
EWDM	0.3792	0.3185	0.3738
GWDM	0.2907	0.2485	0.1150
GWDM_opt	0.2457	0.1786	0.0943
IWDM	0.2433	0.1786	0.1457
IWDM_opt	0.2312	0.1786	0.1207

Table 3.1: Mean, Median, Standard Deviation of the RDR metric for all Decision Makers with expert removal.

Another method to evaluate the performance of each DM is by calculating the percentage of studies in which that DM is most robust, meaning it has the lowest RDR metric in that study. For each study, the DM with the lowest RDR metric is identified. Figure 3.9 shows that some studies have multiple DMs with the lowest RDR metric. In those cases, the study is counted for each of the DMs with the lowest RDR metric. Because of this, the sum of the percentages of table 3.2 is not exactly 100%. This table shows that the GWDM_opt is most often identified as the most robust DM. Interestingly, the IWDM appears in table 3.2 to perform less robust than the other DMs. However, table 3.1, which presents the mean, median, and standard deviation of the RDR scores, shows that the IWDM is not the least robust DM, as its values for these statistics are not the highest. This indicates that the IWDM usually performs quite well, but is likely slightly outperformed by another DM in many studies.

Decision Maker	**Percentage Most Robust (%)**
EWDM	31.25
GWDM	37.50
GWDM_opt	52.08
IWDM	25.00
IWDM_opt	41.67

Table 3.2: Percentage of studies where the Decision Maker is identified as the most robust Decision Maker based on expert removal.

3.3.3 Calibration Question Removal

In the previous section, we examined the RDR metric for expert removal. In this section, we examine the RDR metric for calibration question removal, following a similar approach. Figure 3.10 displays the RDR values for each DM across all studies. The x-axis displays the studies, sorted in ascending order based on the number of calibration questions. Unlike the results from expert removal, these red lines in the plots do not show a consistent downward trend for all DMs. However, there is a noticeable dip or low value around 13 calibration questions for all DMs. This suggests that models with 13 calibration questions are generally more robust. Notably, for the GWDM_opt and the IWDM_opt we again observe that the RDR values decrease when the number of calibration questions increase. This indicates that the robustness improves when there are more questions available.

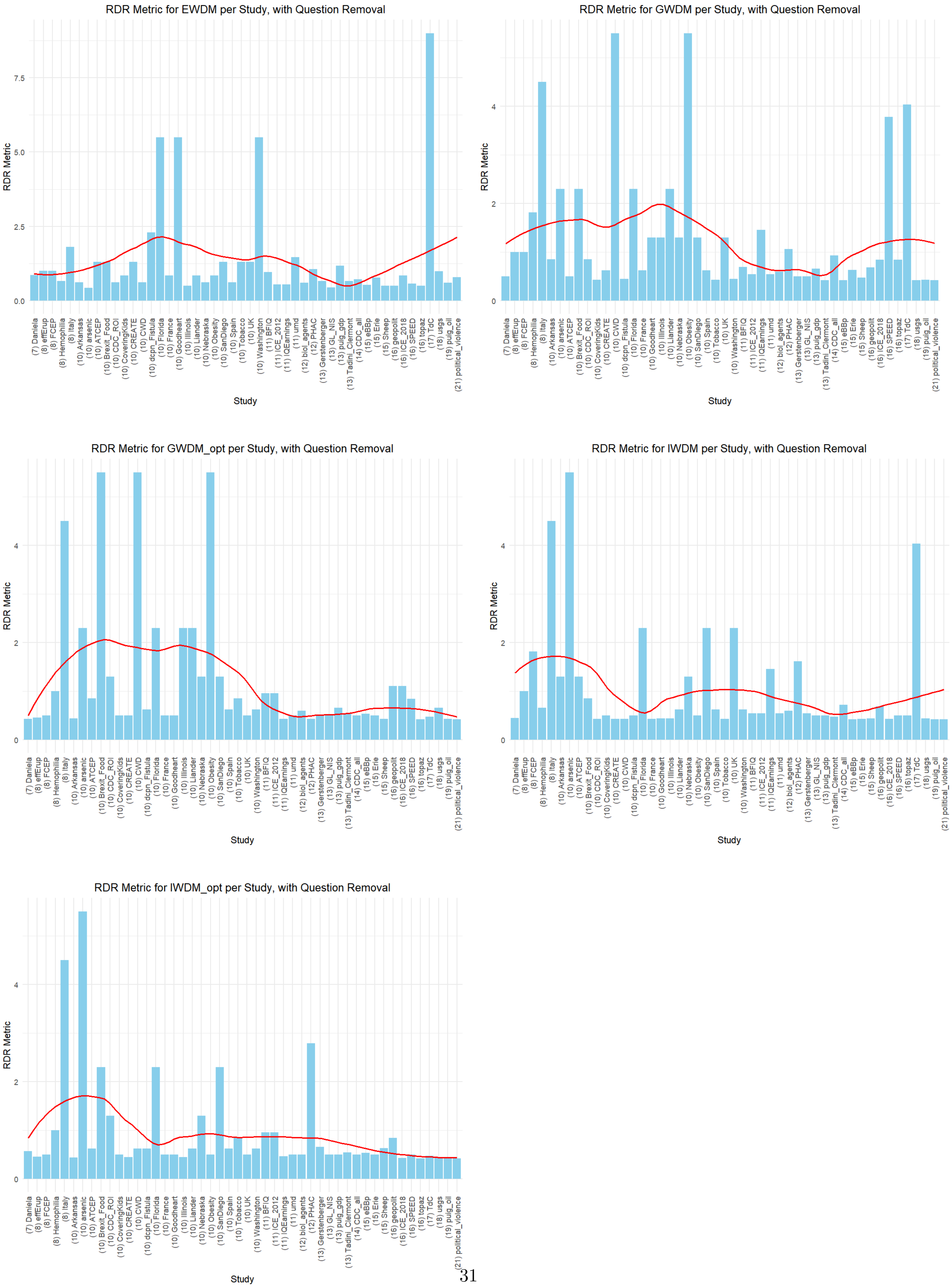


Figure 3.10: Robustness by Distribution Ratio metric for each Decision Maker for all studies with calibration question removal. The red lines are fitted smooth curves representing the underlying trends of the data.

We now return to an overall view of the RDR metric to identify the most robust model for each study, based on the lowest RDR value when removing calibration question. All RDR metrics are shown in figure 3.11. For the Daniela study, which includes 7 calibration questions, the EWDM has the highest RDR metric. Which demonstrates the least robust model. In contrast, the GWDM_opt achieves the lowest RDR value, indicating it is the most robust model for this study.

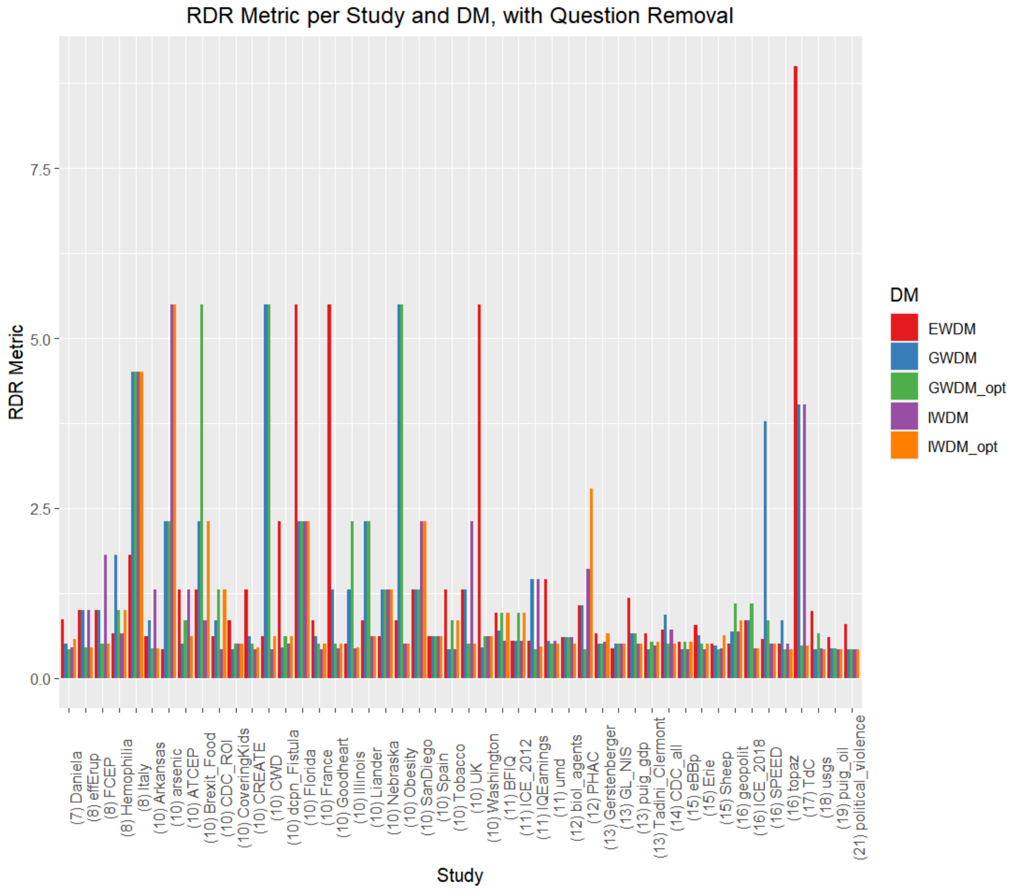


Figure 3.11: Robustness by Distribution Ratio metric with calibration question removal.

The overall robustness of question removal will be assessed once again by the three statistics: mean, median, and standard deviation of the RDR metric. These results are presented in table 3.3. From this table, it can be concluded that the IWDM_opt achieves the lowest values across all three statistics. Thus, this DM performs the best across all three statistics. This demonstrates that, when calibration questions are removed, the IWDM_opt is the most robust model. Additionally, the EWDM has the highest values across all three statistics, which suggests that it is the least robust of the DMs.

Decision Maker	**Mean**	**Median**	**Standard Deviation**
EWDM	1.3278	0.8469	1.6432
GWDM	1.3117	0.8438	1.2992
GWDM_opt	1.2663	0.6847	1.3517
IWDM	1.2031	0.6407	1.1005
IWDM_opt	1.1494	0.6200	1.0237

Table 3.3: Mean, Median, Standard Deviation of the RDR metric for all Decision Makers with calibration question removal.

We once again look at the percentage of studies in which each DM is identified the most robust, based on the lowest RDR metric value. The results are presented in table 3.4. If a study has multiple DMs with the smallest RDR metric, then each of them is counted as the most robust DM for that study. Therefore, the sum of the percentages in this table is not equal to 100%. From table 3.4, it appears that the IWDM is the most robust DM, since it has the highest percentage of studies in which it performs best. However, the three statistics in table 3.3 suggest that the IWDM is the second best DM, right after the IWDM_opt. This suggests that both the IWDM and the IWDM_opt perform quite robust, with the IWDM slightly outperforming the IWDM_opt in a few more studies. In both table 3.3 and table 3.4, the EWDM consistently scores the least robust, since it has the highest values for the three statistics and the lowest percentage of studies where it is considered most robust. Lastly, the GWDM_opt is preferred over the GWDM, as it has a higher mean and median score, and is favored in a larger proportion of studies.

Decision Maker	**Percentage Most Robust (%)**
EWDM	18.75
GWDM	25.00
GWDM_opt	33.33
IWDM	41.67
IWDM_opt	35.42

Table 3.4: Percentage of studies where the Decision Maker is identified as the most robust Decision Maker based on calibration question removal.

4 Discrepancy

In section 2.2, we explained what the information score of an expert is and how it is calculated. The information score of a Decision Maker (DM) is calculated in the same way. In brief, the observed probabilities from the cumulative distribution function of the DM are compared to those from the background measure: the uniform distribution (0.05,0.45,0.45,0.05). The information score quantifies how discrepant the DMs distribution is from the uniform distribution (Hanea & Nane, 2021). However, what if we compare the DMs distribution not to the uniform distribution, but to the distribution of another DM? The metric that compares the difference between an 'ideal' solution and an actual solution, is called discrepancy (Matousek, 1999). In this context, discrepancy refers to the difference between the information score of a DM based on the uniform distribution and the information score of this DM calculated with another DM as background measure. In this section, we analyze the discrepancy of the original dataset.

For each DM, we computed the information score across all studies, using each of the five possible background measures: EWDM, GWDM, GWDM_opt, IWDM, or IWDM_opt. The results are visualized in figures 4.1 to 4.5. On the x-axis, each study is labeled along with the number of experts in parentheses, and the studies are ordered in ascending order based on the number of experts. The y-axis represents the information score. The black dots indicate the information scores of the DMs when the original uniform background measure is used.

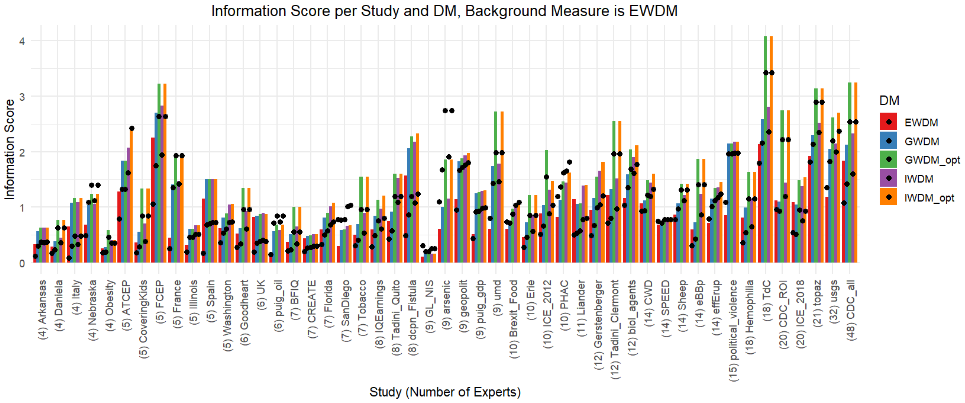


Figure 4.1: Information Scores of the Decision Makers with the EWDM as background measure. The black dots are the information scores of the Decision Makers with the uniform distribution as background measure.

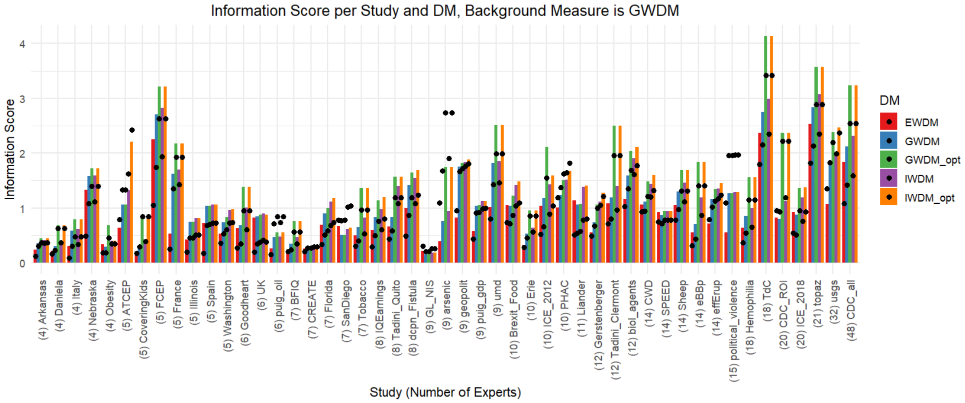


Figure 4.2: Information Scores of the Decision Makers with the GWDM as background measure. The black dots are the information scores of the Decision Makers with the uniform distribution as background measure.

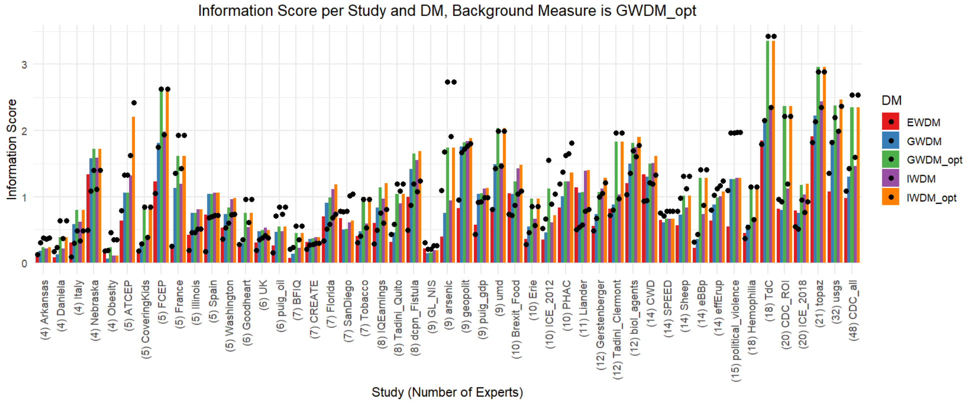


Figure 4.3: Information Scores of the Decision Makers with the GWDM_opt as background measure. The black dots are the information scores of the Decision Makers with the uniform distribution as background measure.

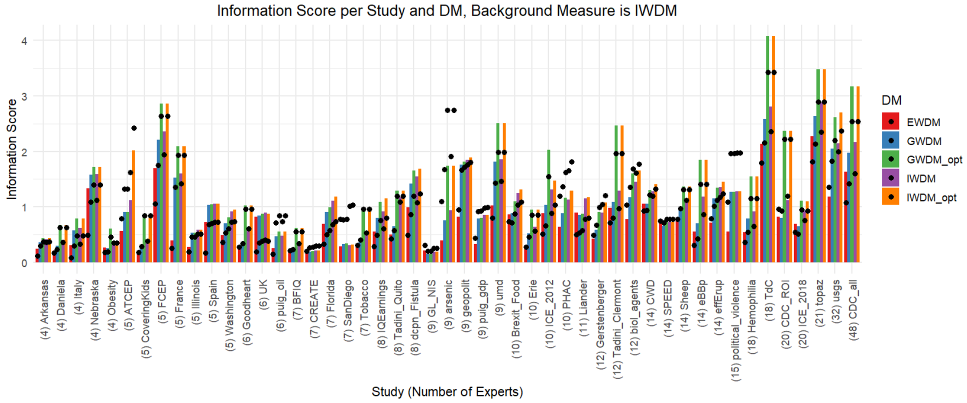


Figure 4.4: Information Scores of the Decision Makers with the IWDM as background measure. The black dots are the information scores of the Decision Makers with the uniform distribution as background measure.

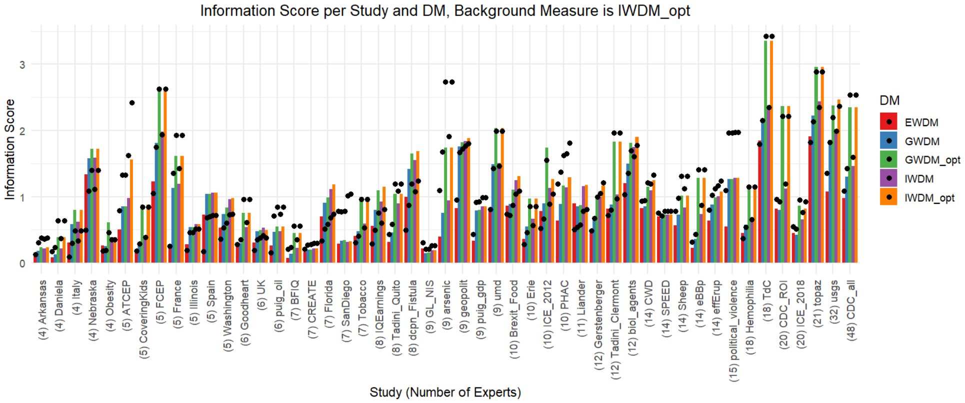


Figure 4.5: Information Scores of the Decision Makers with the IWDM_opt as background measure. The black dots are the information scores of the Decision Makers with the uniform distribution as background measure.

For example, figure 4.5 shows that the information scores of the DMs in the Nebraska study are higher when using the IWDM_opt as the background measure compared to the uniform measure. This indicates that the DM's distributions are more closely aligned with the uniform distribution than with the IWDM_opt. The difference in information scores between the IWDM_opt background measure and uniform background measure is smallest for the GWDM_opt and IWDM_opt, suggesting that these two DMs have the least discrepancy when evaluated against the IWDM_opt background measure. In contrast, the EWDM has the largest difference between the two background measures, indicating the highest discrepancy. In the Daniela study we see the opposite, the information scores are higher when the uniform background measure is used than when using IWDM_opt. In this case, the EWDM and the GWDM have the smallest differences in information scores between the two measures, suggesting lower discrepancy compared to the other DMs.

To summarize the overall performance of each background measure, we calculate the mean, median, and standard deviation of the information scores across all DMs and studies. The results are presented in table 4.1. For example, DMs evaluated using the EWDM as background measure have a higher mean and median information score compared to those evaluated using the GWDM as background measure. This means that the DMs are more closely aligned with the GWDM than with the EWDM based on the mean and median. Conversely, the standard deviation of the information scores are lower with the EWDM background measure than with the GWDM background measure, suggesting that the DMs assessments are more closely aligned with the EWDM than with the GWDM based on this statistic. The EWDM has the mean information scores 0.8072, 0.6514, 0.6936, 0.6099, and 0.5996 with the GWDM, GWDM_opt, IWDM, IWDM_opt, and uniform as background measures, respectively. The smallest mean occurs with the uniform distribution as background measure. Similar, the median is smallest for the uniform measure as well, indicating that the EWDM is least discrepant from the uniform distribution. However, the smallest standard deviation comes from the IWDM_opt measure, indicating that the EWDM is least discrepant from the IWDM_opt based on this statistic. Similar, the GWDM shows it is least discrepant from the IWDM_opt based on both the mean and the standard deviation. Based on the median, the GWDM is most closely aligned to the uniform distribution. Furthermore, the GWDM_opt aligns most closely with the IWDM_opt in terms of mean and standard deviation, and with the uniform distribution based on the median. The IWDM is the least discrepant from the IWDM_opt based on the mean, median, and standard deviation. Lastly, the IWDM_opt shows that is the least discrepant from the GWDM_opt and uniform distribution.

Decision Maker	**Background Measure**	**Mean**	**Median**	**Standard Deviation**
EWDM	EWDM	0.8242	0.6827	0.4920
GWDM	EWDM	1.1333	1.0380	0.6060
GWDM_opt	EWDM	1.5690	1.4625	0.8270
IWDM	EWDM	1.3080	1.2433	0.6253
IWDM_opt	EWDM	1.6089	1.4727	0.8291
EWDM	GWDM	0.8072	0.6733	0.5332
GWDM	GWDM	1.0556	0.8782	0.6324
GWDM_opt	GWDM	1.4754	1.3428	0.8513
IWDM	GWDM	1.2300	1.1240	0.6552
IWDM_opt	GWDM	1.5284	1.4036	0.8451
EWDM	GWDM_opt	0.6514	0.5779	0.4262
GWDM	GWDM_opt	0.8763	0.7534	0.5366
GWDM_opt	GWDM_opt	1.2372	1.0647	0.7116
IWDM	GWDM_opt	1.0268	1.0052	0.5573
IWDM_opt	GWDM_opt	1.2953	1.1854	0.7297
EWDM	IWDM	0.6936	0.5700	0.4786
GWDM	IWDM	0.9513	0.8009	0.6060
GWDM_opt	IWDM	1.3745	1.1004	0.8580
IWDM	IWDM	1.1100	1.1110	0.6265
IWDM_opt	IWDM	1.4184	1.2844	0.8513
EWDM	IWDM_opt	0.6088	0.5325	0.4136
GWDM	IWDM_opt	0.8419	0.7915	0.5334
GWDM_opt	IWDM_opt	1.2145	1.0102	0.7247
IWDM	IWDM_opt	0.9875	0.9639	0.5501
IWDM_opt	IWDM_opt	1.2520	1.1504	0.7231
EWDM	Uniform	0.5996	0.4901	0.4224
GWDM	Uniform	0.8586	0.7136	0.5478
GWDM_opt	Uniform	1.2492	0.9968	0.7557
IWDM	Uniform	1.0247	0.9687	0.5657
IWDM_opt	Uniform	1.3049	1.0895	0.7663

Table 4.1: Mean, median, and standard deviation of the information scores for all Decision Makers and all possible background measures.

Table 4.2 also presents the DMs to which each DM is least discrepant from.

Decision Maker	Mean	Median	Standard Deviation
EWDM	Uniform	Uniform	IWDM_opt
GWDM	IWDM_opt	Uniform	IWDM_opt
GWDM_opt	IWDM_opt	Uniform	IWDM_opt
IWDM	IWDM_opt	IWDM_opt	IWDM_opt
IWDM_opt	GWDM_opt	Uniform	GWDM_opt

Table 4.2: Show for all Decision Makers to which other Decision Maker they are least discrepant from, based on the mean, median, and standard deviation of the information scores.

To quantify overall discrepancy, we compare the information score of each DM using other DMs as background measures to the information score obtained using the uniform distribution as the background. This comparison is made using two metrics: Mean Absolute Error (MAE), and Mean Squared Error (MSE). Let $I(DM)_i^j$ represent the information score of a DM in study j, using background measure i. For example, $I(EWDM)_{IWDM}^3$ refers to the information score of the EWDM in the third study, with IWDM as background measure. The formulas for MAE and MSE are as follows:

$$\text{MAE}(DM)_i = \frac{1}{49} \cdot \sum_{j=1}^{49} \left| I(DM)_i^j - I(DM)_{uniform}^j \right| \tag{4.1}$$

$$\text{MSE}(DM)_i = \frac{1}{49} \cdot \sum_{j=1}^{49} \left(I(DM)_i^j - I(DM)_{uniform}^j \right)^2 \tag{4.2}$$

The results are presented in tables 4.3 and 4.4. For example, the MAE of the GWDM when using the IWDM as background measure is 0.2512, while the MSE is 0.1038. Table 4.3 that the EWDM consistently has the lowest MAE across almost all background measures, indicating that it is the least discrepant DM by the MAE. The GWDM typically has the second lowest MAE for almost all background measures, followed by the IWDM, indicating the GWDM is less discrepant than the IWDM. Notably, the optimized DMs, GWDM_opt and IWDM_opt, have the highest MAE's, indicating the most discrepancy. One possible explanations is that, since these are optimized DMs, there is a cutoff value α. If an expert's combined score is smaller than this cutoff value, the expert is assigned a weight of zero, and thus does not contribute to the DM. However, non optimized DMs still assign a small weight to these experts, and therefore still contribute to the DM. As a result comparing GWDM_opt to the GWDM for example, then there are contributions of some experts in the GWDM, but not in the GWDM_opt, which could lead to a higher error. In table 4.4, we observe a similar trend for the MSE, both the EWDM and GWDM perform best, though in this case, GWDM does outperform the EWDM. Indicating that the GWDM and the EWDM are the least discrepant DMs based on the MSE. Unlike with the MAE, the IWDM does not consistently outperform the GWDM_opt and the IWDM_opt with the MSE. This suggests that the GWDM_opt, IWDM, and IWDM_opt are the most discrepant DMs based on the MSE.

	Decision Maker				
Background Measure	**EWDM**	**GWDM**	**GWDM_opt**	**IWDM**	**IWDM_opt**
EWDM	0.3241	0.3299	0.3867	0.3515	0.3828
GWDM	0.3063	0.3081	0.3363	0.3244	0.3381
GWDM_opt	0.2043	0.2096	0.2298	0.2260	0.2355
IWDM	0.2438	0.2512	0.2801	0.2628	0.2852
IWDM_opt	0.1957	0.1954	0.2105	0.2124	0.2268

Table 4.3: Mean Absolute Error (MAE) for each combination of Decision Maker and background measure. The error is based on the difference between the information scores of a Decision Maker using another Decision Maker as the background measure, and their information scores when using the uniform distribution as the background measure.

	Decision Maker				
Background Measure	**EWDM**	**GWDM**	**GWDM_opt**	**IWDM**	**IWDM_opt**
EWDM	0.1724	0.1692	0.2043	0.1808	0.2048
GWDM	0.1559	0.1446	0.1565	0.1571	0.1609
GWDM_opt	0.0771	0.0774	0.0854	0.0870	0.0904
IWDM	0.1050	0.1038	0.1252	0.1164	0.1320
IWDM_opt	0.0744	0.0739	0.0787	0.0881	0.0986

Table 4.4: Mean Squared Error (MSE) for each combination of Decision Maker and background measure. The error is based on the difference between the information scores of a Decision Maker using another Decision Maker as the background measure, and their information scores when using the uniform distribution as the background measure.

From tables 4.3 and 4.4, we can identify for each DM, which other DM it has the smallest discrepancy with. For example, the EWDM has the smallest MAE and MSE when compared to the IWMD_opt, suggesting that the EWDM has the smallest discrepancy with the IWDM_opt. This is also the case for the GWDM and the GWDM_opt. The IWDM has the least discrepancy with IWDM_opt based on the MAE, but according to the MSE, it has the least discrepancy with the GWDM_opt. Finally, the IWDM_opt shows the least discrepancy with the GWDM_opt. These findings are summarized in table 4.5.

Decision Maker	MAE	MSE
EWDM	IWDM_opt	IWDM_opt
GWDM	IWDM_opt	IWDM_opt
GWDM_opt	IWDM_opt	IWDM_opt
IWDM	IWDM_opt	GWDM_opt
IWDM_opt	GWDM_opt	GWDM_opt

Table 4.5: Background measure for the smallest discrepancy for each Decision Maker based on Mean Absolute Error (MAE) and Mean Squared Error (MSE).

Similar as in section 3, we examine each study individually to determine the least and most discrepant DM. For each DM across all studies, we determine its MAE and MSE, following a similar approach as in formulas 4.1 and 4.2. However, this time the study is fixed, and the information score are compared across the different DMs as background measures. The adjusted formulas are:

$$\text{MAE}(DM)^j = \frac{1}{5} \cdot \sum_{i \in BG} \left| I(DM)_i^j - I(DM)_{uniform}^j \right| \tag{4.3}$$

$$\text{MSE}(DM)^j = \frac{1}{5} \cdot \sum_{i \in BG} \left(I(DM)_i^j - I(DM)_{uniform}^j \right)^2 \tag{4.4}$$

$$BG = \{EWDM, GWDM, GWDM\_opt, IWDM, IWDM\_opt\} \tag{4.5}$$

Here, BG is the set of DMs that are used as background measures. For each study, we identify the DM with the smallest and largest MAE and MSE values. If one study has multiple DMs who share the same smallest or largest values, then each DM is counted. This explains why the sum of the percentages in tables 4.6 and 4.7 is not equal to 100%. From these tables, it is clear that the EWDM is least discrepant overall, as it has the highest percentage of studies where it is the least discrepant DM based on both the MAE and MSE. Following the EWDM, we have the GWDM as least discrepant, as it also has high percentage of studies as least discrepant. Additionally, both the EWDM and the GWDM are in a small percentage of studies identified as the most discrepant DM. The IWDM_opt however, has the highest percentage of studies in which it is identified as the most discrepant DM, and a low percentage as the least discrepant DM. This indicates that this DM is overall the most discrepant. The GWDM_opt also shows a high percentage as the most discrepant DM, and low percentage as least discrepant DM, making it the second most discrepant DM. Interestingly, the IWDM has low percentages in both the most and least discrepant DMs, suggesting that in most studies there is typically at least one other DM performing more and less discrepant than the IWDM.

Decision Maker	**Percentage Most Discrepant (MAE) (%)**	**Percentage Most Discrepant (MSE) (%)**
EWDM	16.67	18.75
GWDM	2.08	4.17
GWDM_opt	35.42	35.42
IWDM	2.08	0.00
IWDM_opt	45.83	41.67

Table 4.6: Percentage of studies where the Decision Maker is identified as the most discrepant Decision Maker based on MAE and MSE.

Decision Maker	**Percentage Least Discrepant (MAE) (%)**	**Percentage Least Discrepant (MSE) (%)**
EWDM	41.67	45.83
GWDM	12.50	10.42
GWDM_opt	8.33	6.25
IWDM	8.33	8.33
IWDM_opt	4.17	6.25

Table 4.7: Percentage of studies where the Decision Maker is identified as the least discrepant Decision Maker based on MAE and MSE.

5 Selecting the Appropriate Decision Maker

Suppose you are conducting a study similar to those in the dataset, but you are unsure which Decision Maker (DM) to use when you are aiming for the best performance, the most robust, or the highest discrepancy model. Imagine that the only information you have is the number of experts and the number of calibration questions in your study. In this case, it would be helpful to have a clear overview of which DM performs most robust under different conditions, based on the number of experts and questions. In the following sections, we present the recommended DM for each of the three objectives.

5.1 Best Calibrated Selection

To recommend the best calibrated DM, we analyzed the DMs across all studies. For each study, we recorded the number of experts and the number of calibration questions. We then identified the DM with the highest calibration score, and recorded both the DM and its score with the study information. This process is repeated for every study. If multiple studies have the same number of experts and calibration questions, then only the DM with the highest calibration question among them is stored. In figure 5.1 the results are presented. For example, if your study includes 20 experts and 10 calibration questions, then the DM with the highest calibration score is recommended to be the GWDM. However, if your study has 20 experts and 16 questions, then the IWDM_opt is recommended for the highest calibration.

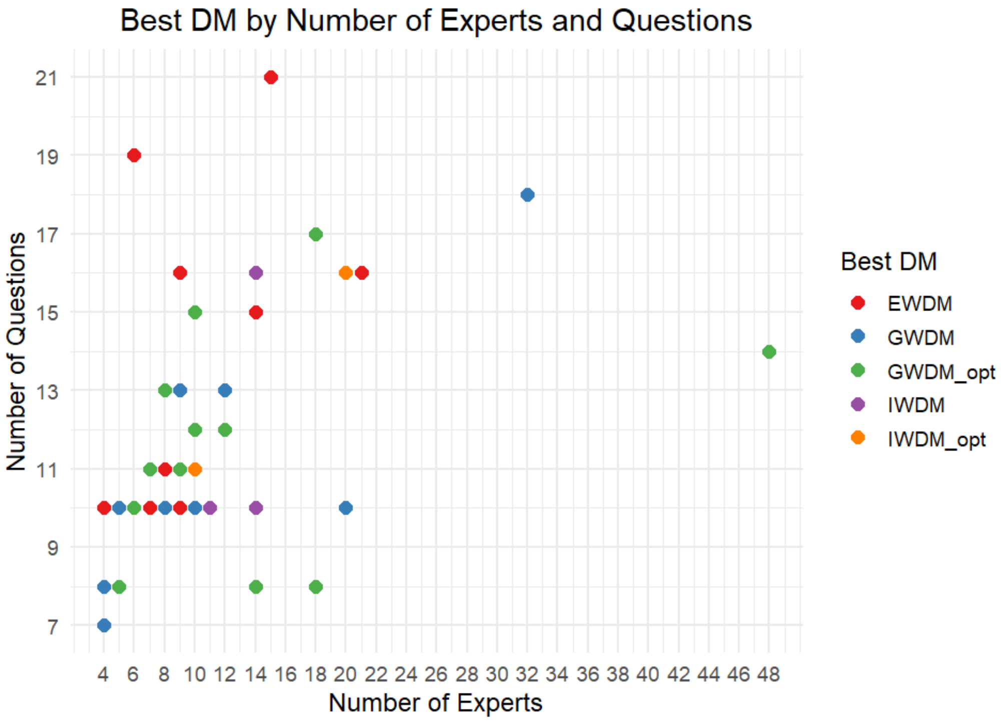
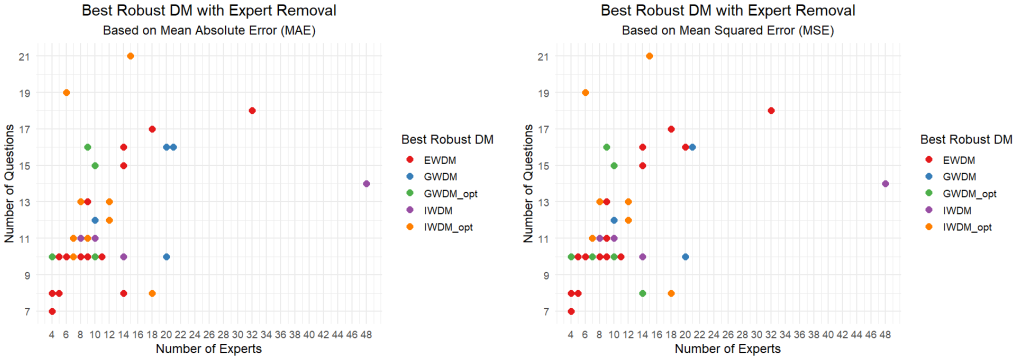


Figure 5.1: Highest calibrated Decision Makers for a given number of experts and calibration questions.

5.2 Robust Selection

In section 3, we discussed which models are more robust. To determine the most robust DM for a given number of experts and calibration questions, we analyze the calibration scores obtained after removing individual experts or questions again. For each study, we first record the number of experts and calibration questions. Using the list of calibration scores generated by removing one expert or one question at a time, we calculate the Mean Absolute Error (MAE) and Mean Squared Error (MSE) for each DM. These are determined by comparing the modified calibration scores to the calibration scores of the original dataset, following the same approach of formulas 4.1 and 4.2 from section 4. For each combination of numbers experts and questions, we find the DMs with the lowest MAE and MSE, and store both the errors and the corresponding DMs. If multiple studies share the amount of experts and questions, then only the DM with the lowest error is stored. The resulting recommendations, based on the MAE and MSE for the expert removal, are visualized in figures 5.2a and 5.2b. For example, a study with 14 experts and 8 calibrations questions, the MAE suggest using the EWDM as the most robust model. However, the MSE suggests using the GWDM_opt as the most robust model.

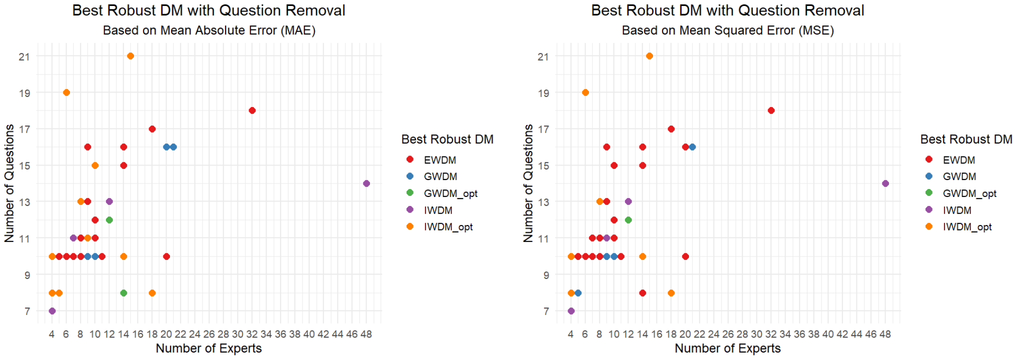


(a) MAE

(b) MSE

Figure 5.2: Robust recommendations based on MAE and MSE of Decision Makers calibration scores, with expert removal.

We repeated this process to evaluate the robustness of question removal. The resulting recommendations are presented in figures 5.3a and 5.3b. In this case, for a study with 14 experts and 8 calibration questions, we see that the MAE recommends the GWDM_opt. However, the MSE recommends the EWDM as most robust DM.

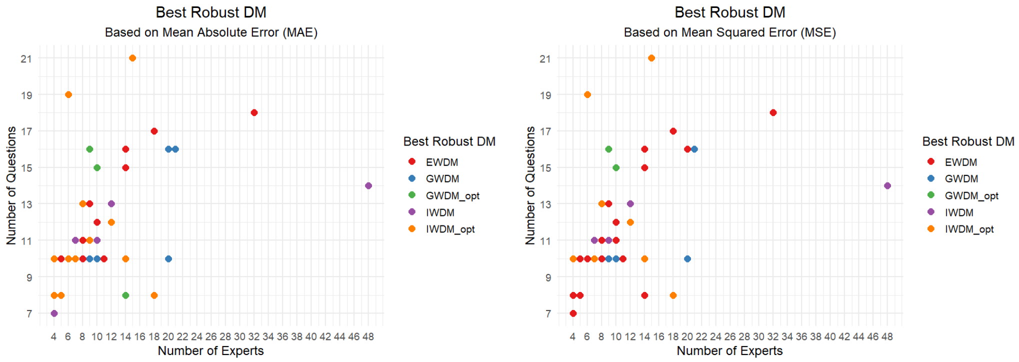


(a) MAE

(b) MSE

Figure 5.3: Robust recommendations based on MAE and MSE of Decision Makers calibration scores, with calibration question removal.

To combine both sets of recommendations, we calculate the sum of the MAE values for each DM, from expert and question removal. We then find the DM with the lowest combined error and record this DM along with the corresponding error, number of experts, and number of calibration questions. The same process is applied to the MSE values. The final recommendations are visualized in figures 5.4a and 5.4b. The final recommendation in a study with 8 experts and 14 questions, the GWDM_opt is the most robust DM according to the MAE, while the EWDM is preferred by the MSE.



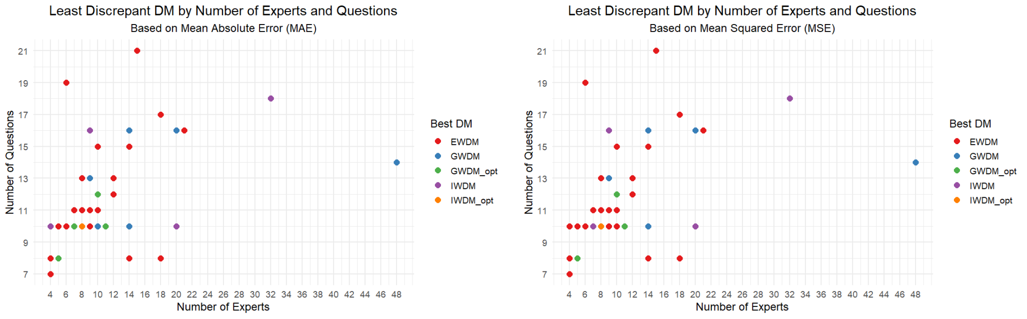
(a) MAE

(b) MSE

Figure 5.4: Robust recommendations based on MAE and MSE of Decision Makers calibration scores

5.3 Discrepant Selection

In section 4, it is discussed which DMs are more discrepant than the others. It showed that the EWDM and the GWDM are, overall, the least discrepant DMs across all studies. That section also discussed the percentage of studies in which each DM was identified to be the least discrepant. Using the same Mean Absolute Error (MAE) and Mean Squared Error (MSE) calculations from formulas 4.3 and 4.4, we identify the most and least discrepant DMs for all 49 studies. For each study we have a specific number of experts and calibration questions, along with the smallest and largest values for the MAE and MSE. The DMs with the smallest MAE and smallest MSE are selected for that number of experts and questions as the least discrepant choice of DM. If multiple studies share the same number of experts and questions, then the DM with the lowest MAE is kept for the MAE. This selection process is also done for the most discrepant DMs, using the highest MAE and MSE. The final recommendations based on the least discrepant DMs are visualized in figures 5.5a and 5.5b. For example, in a study with 10 experts and 10 calibration questions, the GWDM is recommended to be the least discrepant based on the MAE. However, in this same study, the EWDM is preferred to be the least discrepant by the MSE.

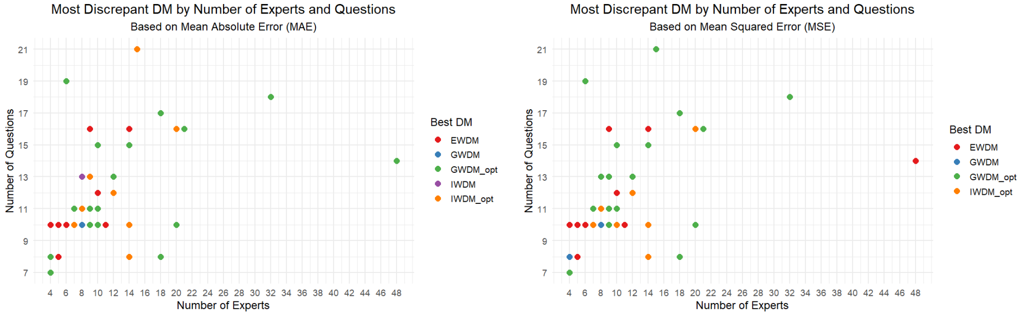


(a) MAE

(b) MSE

Figure 5.5: Recommendations for least discrepant Decision Maker based on Mean Absolute Error (MAE) and Mean Squared Error (MSE) of Decision Makers. The error is based on the difference between the information scores of a Decision Maker using another Decision Maker as the background measure, and their information score when using the uniform distribution as the background measure.

For the recommendations based of the most discrepant DMs we have figures 5.6a and 5.6b. For example, in a study with 10 experts and 10 calibration questions, the GWDM_opt is recommended to be the most discrepant based on the MAE. However, in this same study, the IWDM_opt is the most discrepant by the MSE.



(a) MAE

(b) MSE

Figure 5.6: Recommendations for most discrepant Decision Maker based on Mean Absolute Error (MAE) and Mean Squared Error (MSE) of Decision Makers. The error is based on the difference between the information scores of a Decision Maker using another Decision Maker as the background measure, and their information score when using the uniform distribution as the background measure.

6 Conclusion

This report presented an analysis of the robustness and discrepancy of Cooke's Classical Model (CM) based on Structured Expert Judgment (SEJ). To evaluate robustness, section 3 analyzes the calibration scores of each Decision Maker (DM) across all 49 studies. To test the stability of these scores under data uncertainty, each expert and each calibration question was removed once from the data. Based on the boxplots of the calibration scores, it can be concluded that for the GWDM, GWDM_opt, IWDM, and IWDM_opt, increasing the number of experts led to more robust models, since the variation of the scores decreased and the scores were more closely centered around the original calibration scores. More specifically, for these DMs, the calibration scores remain constant when the number of experts reaches 21 or more. For the EWDM, robustness increased in studies with 32 experts or more.

When assessing robustness by the number of calibration questions, it can be concluded that for IWDM and IWDM_opt the calibration scores became more stable with more calibration questions. Since the variety of the scores decrease when the questions increase. In contrast, there was no clear trend found for GWDM, and GWDM_opt. For the EWDM it was noticeable that the variability in calibration scores decreased when the number of calibration questions reached 17 or more. This suggests that the IWDM and IWDM_opt become increasingly robust as more questions are added, while the EWDM shows improved robustness when at least 17 questions are available.

The Robustness by Distribution Ratio (RDR) further supports the conclusion that an increase in number of experts generally leads to greater robustness of the models. For all 49 studies, the GWDM_opt, IWDM, and IWDM_opt were the most robust DMs when evaluated using the mean, median, and standard deviation of the RDR metric under expert removal. Additionally, the GWDM_opt and the IWDM_opt have the highest percentage of studies in which they are identified as the most robust models. For question removal, only the GWDM_opt and the IWDM_opt showed a consistent decrease in the RDR metric as the number of questions increased. Interestingly, the EWDM, GWDM, and IWDM showed smallest RDR metric, and thus best robustness, at 13 calibration question based on the fitted RDR curves. Overall, the IWDM_opt was found to be the most robust with question removal based on the three statistics. Furthermore, the IWDM, IWDM_opt, and GWDM_opt have the highest percentages of studies where this DM is identified to be the most robust DM. Thus, based on expert and calibration question removal, we can conclude that the IWDM, IWDM_opt, and GWDM_opt are overall the most robust DMs.

In section 4, discrepancy was analyzed by comparing the Mean Absolute Error (MAE) and the Mean Squared Error (MSE) of the information scores of each DM against alternative background measures. These information scores showed that EWDM had the lowest MAE across all background measures, followed by GWDM and IWDM, making them the least discrepant DMs. In contrast, the GWDM_opt and IWDM_opt had the highest MAE, making them the most discrepant. This is also shown in the percentage of studies where each DM is identified as the least discrepant DM. The EWDM had the highest percentage, followed by the GWDM, then the IWDM, GWDM_opt, and IWDM_opt. For the MSE, the GWDM and EWDM had the lowest scores, making them the least discrepant DMs. The IWDM, GWDM_opt and IWDM_opt score the highest on MSE, and thus are more discrepant by this measure. This is further supported by the percentage of studies where each DM is identified as the least discrepant DM by the MSE. The EWDM and the GWDM appeared most often as the least discrepant DM, while the GWDM_opt, IWDM, and IWDM_opt appeared less often as the least discrepant DM. Thus overall, the EWDM and the GWDM are suggested to be the least discrepant DMs.

7 Discussion

This analysis is based on a dataset of 49 studies. Some of these studies focus on topics such as vaccinations, political violence, and obesity, or represent only certain countries. Therefore, the findings of this report may not fully generalize to Cooke's Classical Model (CM) across other domains, such as climate change or different geographical regions where this model is also used. Additionally, the dataset has only a few studies with 21 or more experts and 17 or more calibration questions. This does limit the strength of the conclusion taken in this report that an increase in the number of experts leads to a more robust model. Similar for the influence of the number of calibration questions on the robustness of a Decision Maker (DM).

For further research, a wider dataset should be used. Including studies from other domains, and a wider range of number of experts and questions.

References

Bamber, J. L., Oppenheimer, M., Kopp, R. E., Aspinall, W. P., & Cooke, R. M. (2019). Ice sheet contributions to future sea-level rise from structured expert judgment. *Proceedings of the National Academy of Sciences*, *116*(23), 11195–11200.

Beshearse, E. (2021). Attribution of Illnesses Transmitted by Food and Water to Comprehensive Transmission Pathways Using Structured Expert Judgment, United States. *Emerging Infectious Diseases*, *27*(1), 182–195. doi: https://doi.org/10.3201/eid2701.200316

Cooke, R. (n.d.). *Supplementary information for structured expert judgment.* Retrieved from `https://rogermcooke.net/rogermcooke_files/SEJ%20-%20SI%20June%2022%202022.pdf`

Grimmett, G., & Welsh, D. (2014). *Probability an introduction.* Oxford University Press.

Hanea, A. M., & Nane, G. F. (2021). An in-depth perspective on the classical model. *Expert Judgement in Risk and Decision Analysis*, 225–256.

Hoffmann, S., Devleesschauwer, B., Aspinall, W., Cooke, R., Corrigan, T., Havelaar, A., ... others (2017). Attribution of global foodborne disease to specific foods: Findings from a world health organization structured expert elicitation. *PloS one*, *12*(9), 1.

Matousek, J. (1999). *Geometric discrepancy: An illustrated guide* (Vol. 18). Springer Science & Business Media.

Mehdi, E. R. E. (2022). *Beyond classical optimization paradigms: Robustness, fairness others.* Retrieved from `https://erraqabielmehdi.medium.com/beyond-classical-optimization-paradigms-robustness-fairness-others-dce754836c0d`

Mia Hubert, P. J. R. . S. v. A. (2004). Robustness. *Encyclopedia Of Actuarial Science*, *3*, 1515–1529. doi: https://wis.kuleuven.be/stat/robust/papers/2004/hubertrousseeuwvanaelst-robustness-encyactsciences.pdf

Sarkar, R. (2023). *Unveiling model robustness in value prediction, causal inference, and forecasting: An in-depth rdr analysis.* Retrieved from `https://medium.com/@sticktorick/unveiling-model-robustness-in-value-prediction-causal-inference-and-forecasting-an-in-depth-rdr-ab7b6ef542c2`

TU Delft OpenCourseWare. (2020). *Examples of sej studies using cooke's method.* Retrieved from `https://ocw.tudelft.nl/course-readings/1-4-2-examples-of-sej-studies-using-cookes-method/`

Parts of the writing in this report were improved and rephrased using ChatGPT, a language model developed by OpenAI.