When are there too many experts?

By Structured Expert Judgment methods

Femke Heuff

Applied Mathematics, Delft, University of Technology

Thesis committee:

Tina Nane	TU Delft	Supervisor
Vandana Dwarka	TU Delft	Second Assessor

Lay Abstract

In some areas of science, like risk analysis or public health, we often don't have enough data to base decisions on. In such cases, we rely on experts. We create panels of experts and ask them to evaluate quintiles of interest. But, how many experts do we actually need? And when are there too many experts? These are the main questions of my thesis. To answer these questions I used the Classical Model (CM), which evaluates each expert based on two performance scores: the calibration score, which measures the expert's statistical accuracy, and the information score, which measures how precise the expert's uncertain assessments are. These scores are used to assign weights when combining expert opinions, and several weighting approaches are examined and compared in this thesis.
Using data from the National Institute for Public Health and the Environment (RIVM), this study analyzes how the number of experts in a panel affects the performance of the different Decision Makers. Special attention is given to the role of experts who consistently underestimate outcomes, and whether their uncertain assessments affect their individual performance and the performance of the aggregated distribution. This research contributes to understanding how to construct effective expert panels for crucial decisions.

Abstract

When data is missing, expert judgment becomes an essential tool for estimating uncertain outcomes. The Classical Model (CM), developed by Roger Cooke, provides a structured approach for evaluating and combining expert opinions. It does this by assigning each expert two performance scores: the calibration score and the information score. The calibration score reflects how well experts quantify uncertainty, and the information score measures the precision of their uncertain estimates. These scores are then used to assign weights to experts, which are then applied to combine the judgments of multiple experts into a single distribution. This combination is called a Decision Maker (DM). There are several ways to assign weights to the experts, each leading to a different kind of DM.

This thesis investigates how the size of an expert panel influences the quality of aggregated probabilistic judgments. Using data from the Dutch National Institute for Public Health and the Environment (RIVM), which includes estimates from 43 experts on 15 seed questions, 5 weighting methods were applied to panels of varying sizes. Seed questions are questions for which the true answers are known to the analyst but not to the expert. A sub-sampling analysis was conducted for panels of different sizes. The results indicate that performance is sensitive to expert selection in smaller panels, while larger panels tend to provide more stable results, though these are not necessarily more accurate. Notably, panels consisting of 6 to 8 experts achieved the best calibration scores.

This thesis also focuses on how experts estimate the 15 seed questions in terms of underestimation and overestimation. I examined how frequently underestimating or overestimating influences an expert's individual calibration score. In addition, I investigated how the presence of a high proportion of underestimators within a panel affects the calibration scores of different Decision Makers. The results show that all underestimators and overestimators tend to have low individual calibration scores (below the thresholds 0.05). Panels with a large proportion of underestimators most strongly affect the Equal Weight Decision Maker (EWDM), as this method gives each expert the same weight. However, overall panel performance is not determined by underestimation alone. The composition of the panel plays a critical role. Sometimes panels with many understimators can still perform well if they include strong-performing experts who compensate for other assessments.

The performance of the five different Decision Makers was compared across all sections. The results indicated that optimized approaches, which exclude experts with calibration scores below a specific threshold, consistently resulted in higher calibration scores, irrespective of the panel size.

This research contributes to understanding how expert panels should be constructed and evaluated. The findings provide practical guidance on selecting the number of experts and choosing appropriate aggregation methods to improve the reliability of expert-based opinions.

Contents

1 The Classical Model

1.1 Introduction

In many scientific fields, like risk analysis or environmental health, we often deal with situations where we don't have enough reliable data to work with. In such cases, expert judgment becomes an important source of information. The Classical Model (CM), developed by Roger Cooke, is a structured method that helps to mathematically combine expert opinions.
Experts are provided with a set of seed questions. These questions help assess how well each expert can estimate unknown values. The actual outcomes (realizations) of these seed questions are known to the analyst, though the experts do not have access to this information at the time of providing their estimations. Experts are evaluated based on their performance in assessing the uncertainty of seed questions. This evaluation consists of two scores: the calibration score and the information score. These scores are then used to assign weights to each expert, which are applied when combining their opinions. The final distribution is known as the combined expert opinion, or Decision Maker (DM). There are various types of Decision Makers, each employing a different weight distribution for combining experts' opinions. In this chapter, we explain how to calculate the calibration score and the information score, and how these scores can be utilized to create different types of Decision Makers.

1.2 Calibration Score

The calibration score measures how well an expert's uncertainty estimates match the actual outcomes. In other words, it evaluates whether the expert's predicted probabilities correspond with reality. During an expert elicitation, experts are asked to provide percentiles, typically the 5th, 50th, and 95th percentiles, for several seed questions. These percentiles divide the range of possible outcomes into four probability intervals:

1. until 5th percentile
2. from 5th percentile to 50th percentile
3. from 50th percentile to 95th percentile
4. above 95th percentile

If an expert is well calibrated, we expect the true value to fall into these ranges with the following probabilities:

- 5% chance for the first range
- 45% chance for the second range
- 45% chance for the third range
- 5% chance for the fourth range

To calculate the calibration score, we check in which interval the true value of each question falls, and we count how many times each interval is hit. Dividing these counts by the total number of questions gives the observed proportions, denoted as $S = (S_1, S_2, S_3, S_4)$.

Next, these observed proportions are compared to the expected probabilities $p = (0.05, 0.45, 0.45, 0.05)$. The difference between these two distributions is measured using the Kullback-Leibler divergence. It is calculated as:

$$l(S, p) = \sum_{k=1}^{4} S_k \ln\left(\frac{S_k}{p_k}\right) \tag{1}$$

where

- S_k is the observed proportion for $k = 1, 2, 3, 4$
- p_k is the expected proportion for interval k

Next, we calculate:

$$T = 2M \cdot l(S, p) \tag{2}$$

Where

- M is the number of seed questions.

The test statistic T is asymptotically distributed as a chi-square random variable with 3 degrees of freedom. Knowing this, a metric for the calibration score for an expert e can be defined. Namely:

$$Cal(e) = 1 - F(T) \tag{3}$$

Where:

- F is the cumulative distribution function corresponding to the Chi-square distribution with 3 degrees of freedom.

The calibration score represents the p-value under the null hypothesis that the expert's stated uncertainties are statistically correct. A high calibration score means the expert's uncertainty assessments are consistent with the true realizations.

- a high p-value (close to 1) suggests that the expert's uncertainty estimates match the true outcomes well.
- a low p-value (below 0.05) suggests that the expert's uncertainty estimates do not fit the actual results and are not statistically accurate.

Since the calibration score is a p-value, it will always be between 0 and 1. Thus, the calibration score gives a clear and statistical way to judge whether an expert can reliably quantify uncertainty in their assessments, and it's known as statistical accuracy.

1.3 Information Score

The information score measures how much useful information an expert provides through their uncertainty estimates. In simple words, it measures how sharp or narrow an expert's uncertainty ranges are. The narrower the range between the 5th and 95th percentile, the more confident the expert is, and the more information their answer contains.
To calculate this, we first define something called the intrinsic range for each question. The intrinsic range is built by looking at the smallest and largest values among all expert assessments and the true realization, and then slightly extending this range by a fixed

percentage (called the overshoot factor). Usually, an overshoot of 10% is used. The adjusted lower and upper bounds are calculated as:

$$L^* = L - k(U - L)$$

$$U^* = U + k(U - L)$$

where:

- L is the minimum of all expert percentiles and the realization
- U is the maximum of all expert percentiles and the realization
- k is the overshoot factor (typically 0.1)

We then assume a simple background distribution between L^* and U^* has a uniform distribution when the variable spans only a small range and has a log-uniform distribution if the variable spans several orders of magnitude (typically 4 orders of magnitude).

The information score now quantifies how "narrow" an expert's assessment is against the background measure. It is calculated as follows:

$$\begin{aligned} I(e,q) =& 0.05 \cdot \ln\left(\frac{0.05}{q_5 - L^*}\right) + 0.45 \cdot \ln\left(\frac{0.45}{q_{50} - q_5}\right) \\ &+ 0.45 \cdot \ln\left(\frac{0.45}{q_{95} - q_{50}}\right) + 0.05 \cdot \ln\left(\frac{0.05}{U^* - q_{95}}\right) + \ln(U^* - L^*) \end{aligned} \tag{4}$$

Where:

- e refers to the expert in question
- q is the question for which the score is calculated

This formula measures how much the expert's distribution differs from the background measure, which is typically assumed to be the uniform distribution. If the expert gives very wide uncertainty ranges, the expert's distribution will look similar to the background measure, and the information score will be low. If the expert gives sharp, narrow ranges, the information score will be high. Since each expert answers multiple seed questions, we average the information score across all questions to get the expert's overall score:

$$I(e) = \frac{1}{M} \sum_{j=1}^{M} I(e, q_j) \tag{5}$$

Where:

- M is the number of seed questions
- $I(e, q_j)$ the information score corresponding to question j.

Unlike the calibration score, the information score is not a p-value and can be larger than 1. The information score changes more slowly than the calibration score, which means the calibration score often has more influence on the final performance weight. However, to be considered a good expert, it is important to score well on both the calibration score and the information score

1.4 Decision Makers (DMs)

The final goal in the Classical Model is to combine the judgments of multiple experts into a single distribution. This combination is called a Decision Maker (DM). The DM is a mathematical combination of all experts' opinions, created using a method called linear pooling. This means the final distribution is a weighted average of all expert distributions. There are several ways to assign weights to the experts, each leading to a different kind of DM. The DM can also be evaluated like an expert and it gets its own calibration and information scores (by applying it to the seed questions). Below, we explain the most common ones.

- **Equal weights DM (EWDM)**: The simplest approach is to give every expert the same weight, regardless of how well they performed. This means each expert contributes equally to the final distribution. In formula form:

$$f_i = \frac{1}{N} \sum_{j=1}^{N} f_{j,i}$$

  Where:

  - f_i denotes the DMs probability distribution for question i
  - $f_{j,i}$ is the probability distribution from expert j for question i
  - N is the total number of experts.

- **Global Weights DM (GWDM)**: In this method, each expert gets a weight based on how well they performed overall. We calculate a combined score for each expert by multiplying their calibration and information scores:

$$\mathrm{W}(e_i) = \frac{\mathrm{Cal}(e_i) \cdot I(e_i)}{\sum_{k=1}^{N} \mathrm{Cal}(e_k) \cdot I(e_k)}$$

  Where:

  - $W(e_i)$ = weight of expert i
  - $Cal(e_i)$ = Calibration score of expert i
  - $I(e_i)$ = Information score of expert i

  These weights are then used to create the DM by:

$$f_i = \sum_{j=1}^{N} \mathrm{W}(e_i) \cdot f_{j,i}$$

- **Item Weights DM (IWDM)**: This approach is similar to Global Weights, but it focuses more on how confident and informative an expert is for each individual question. Instead of using the average information score, we use the expert's score only for that question:

$$W_i^j = \frac{\mathrm{Cal}(e_i) \cdot I_j(e_i)}{\sum_{k=1}^{N} \mathrm{Cal}(e_k) \cdot I_j(e_k)}$$

  Where:

– W_i^j is the weight of expert i of question j

Then the DM is created as:

$$f_i = \sum_{j=1}^{N} W_i^j \cdot f_{j,i}$$

This method allows the final distribution to reflect the fact that some experts may be more reliable for specific topics.

- **Optimized GWDM and Optimized IWDM (GWDM_Opt and IWDM_Opt)**: In some cases, it makes sense to only include experts who are statistically reliable. This is done by defining an $\alpha > 0$ and not weighting an expert if his or her calibration score falls below this threshold. For example, we might only include experts with a calibration score above 0.05. α can also be found by iteratively repeating the process described below:

  – define a set of α's over which to iterate

  – for each α, construct a decision maker. In this decision maker only the experts with a calibration score higher than α are used to calculate the weights.

  – Calculate what combined score would the decision maker receive if it were added as an actual expert to the study.

  Where, The Combined Score = The Calibration Score × The Information Score. Now the α for which the decision maker has the greates combined score is chosen as the cut-off value for the final decision maker.

2 Performance insights

To answer the main questions of my thesis: How many experts do you need in a panel? And when are there too many experts?, I used two datasets, both provided by the Dutch National Institute for Public Health and the Environment (RIVM).

The first part of the analysis is based on the first dataset, which contains uncertainty assessments from 43 experts. Although 47 experts were initially invited, experts numbered 7, 21, 39, and 41 did not participate in the elicitation process. Each of the 43 experts provided estimates for 15 seed questions by determining the 5th percentile (q_5), the 50th percentile (q_{50}), and 95th percentile (q_{95}) for each question. The actual outcomes of these questions, referred to as realizations, were also included in the dataset.

2.1 Individual Performance Scores

Using the q_5, q_{50}, and the q_{95} of the 15 different seed questions from 43 experts, the two performance scores from the Classical Model were calculated for each expert:

- The **calibration score**, which tells us how statistically accurate the experts' judgments are.
- The **information score**, which measures how precise the ranges are.

To explore how the number of experts in a panel impacts the average calibration score and information score, a sub-sampling method was employed. Panels of sizes $k \in \{5, 10, 15, 20, 25\}$ were repeatedly created by randomly selecting experts from the first dataset of 43 experts. For each panel size, 100 random subsets were generated. Within each subset, the average calibration and information scores were computed based on all 15 seed questions. In Section 3.2, the results were presented using box plots to illustrate how performance varies with different panel sizes. It's important to note that in this part of the analysis, the different decision makers were not considered; instead, I focused only on the average calibration score and information score. This approach allows us to understand how the average performance of a group behaves with different group sizes. Later, this will help in evaluating the added value of applying decision makers.

2.2 Evaluating Decision Makers Across Different Panel Sizes

In the previous section, the calibration score and information score for each of the 43 experts were calculated. These scores are then used to construct various weighted combinations of expert judgments, known as Decision Makers (DMs). In this thesis, the following DMs are considered:

- EWDM: Equal-weights Decision Maker
- GWDM: Global-performance-weighted Decision Maker
- GWDM_opt: Optimized global performance-weighted Decision Maker
- IWDM: Item-based performance-weighted Decision Maker
- IWDM_opt: Optimized item-based Decision maker

It's important to note that a Decision Maker behaves like a new expert. This means that once a DM is formed using a group of experts, its performance can be evaluated in the same way by calculating the calibration score using the 15 seed questions.

In Section 3.3, the performance of the five different DMs is evaluated across different panel sizes $k \in \{5, 10, 15, 20, 25\}$. For each group size, 100 random subsets of experts were created from the 43 experts. Within each subset, the calibration score and information score of each expert is calculated. Based on these scores, the five different DMs were constructed: EWDM, GWDM, GWDM_opt, IWDM, and IWDM_opt. Each DM represents a weighted combination of the selected experts' probability distributions. After constructing these DMs, their calibration scores were computed by applying them to the original set of 15 seed questions. The results are shown in box plots, which makes it easier to compare the performance of different DMs and to determine how many experts we need in a panel. This analysis was later extended to smaller panel sizes, specifically $k \in \{2, 4, 6, 8, 10\}$. This was done to see how the calibration score acts in small groups of experts.

2.3 Pathogen-Specific Panel Evaluation

In Section 3.4, the second dataset is used, which consists of real panels of experts who estimated the spread of 26 different pathogens. The experts in these panels are the same 43 individuals from my first dataset. The number of experts per pathogen panel ranged from 4 to 22. For each panel, calibration scores were calculated for all DMs listed above. The results were visualized in graphs where I plotted the calibration score of the different DMs against the number of experts in the panels. This was done to determine if the analyzes of before corresponds with the real expert panels.

2.4 Behavioral Profiling of Experts

To clarify the results of the different pathogen panels, the experts involved in each panel were closely examined in Section 3.5. For each expert and seed question, the realization r was compared to their predicted 90% credible interval $[q_5, q_{95}]$. The classification was performed as follows:

- if $r \in [q_5, q_{95}]$, the estimate was labeled as Good
- if $r < q_5$, the estimate was classified as an Overestimate
- if $r > q_{95}$, the estimate was classified as an Underestimate
- difference = #overestimates - #underestimates

From these classifications, the proportion of Good, Overestimated, and Underestimated answers were computed for each expert. The difference between the numbers of overestimates and underestimates was used to characterize each expert's tendency.

In this analysis, an expert was classified as an underestimator if their difference score was less than or equal to -3, and as an overestimator if their difference score was larger than or equal to 3. This threshold was chosen to avoid misclassifying experts with only a small difference, likely due to coincidence. A score of -3 or lower, or a score of 3 or higher,

indicates a more consistent pattern of underestimation or overestimation across the 15 seed questions. To better understand how estimation bias affects expert performance, a table was made in section 3.5 listing all experts identified as underestimators or overestimators, along with their individual calibration scores. This allowed for a better view of how systematic bias affects the individual calibration score. In addition, a mathematical explanation was given of how underestimators influence the individual calibration score, supported by an example.
Since my data contains significantly more underestimators than overestimators, the focus of the rest of my analysis will be primarily on how underestimation affects calibration scores.

After examining how underestimation affects individual calibration scores, the analysis was extended to assess the impact of underestimators on the performance of different Decision Makers. To do this, five scatter plots (one for each DM) were created showing how the calibration score changes with the percentage of underestimators in each panel. This allowed me to observe how each DM responds to increasing levels of underestimation within a panel.

To further investigate these trends, I selected three panels for detailed examination: the BRUCL, CRYPT, and LEPTO panels. These three panels were selected since they clearly illustrate how underestimators can influence the calibration score, while also showing that other factors, such as the composition of the panels, may play a role in affecting the results. This also allowed for a comparative analysis of how different DMs perform under varying panel conditions.

3 Results

In the first part of my analysis, the first dataset provided by the Dutch National Institute for Public Health and the Environment (RIVM) was used. This data contains uncertainty assessments from 43 experts, each of whom provided estimates in the form of three percentiles (5th, 50th, and 95th) for 15 seed questions. The actual realizations of these questions were also included in the dataset. Although 47 experts were originally invited to participate, experts numbered 7, 21, 39, and 41 did not complete the elicitation process and were therefore excluded from the final analysis.

3.1 Individual Performance Scores

Using the first dataset, both the individual calibration score and information score were calculated for each of the 43 experts. This provides an overview of how each expert performed independently in terms of both statistical accuracy and informativeness. The results are presented in Table 1 below.

ExpertID	Cal Score	Inf Score
1	3.48e-05	3.60
2	0.0115	3.31
3	1.60e-10	3.58
4	0.0485	2.99
5	0.6432	2.19
6	0.5710	3.22
8	1.05e-04	3.55
9	1.48e-06	3.48
10	1.48e-06	3.68
11	0.0158	3.07
12	0.0222	3.21
13	0.0842	3.11
14	3.28e-06	3.73
15	0.0005	3.77
16	0.0031	3.58
17	1.48e-06	4.62
18	0.0017	3.22
19	0.5710	1.15
20	2.70e-06	4.12
22	3.12e-05	4.19
23	0.0842	3.00
24	1.19e-05	3.52
25	0.0115	3.52
26	0.0158	3.61
27	0.0047	3.57
28	0.0346	2.92
29	6.35e-05	3.51
30	2.25e-05	3.07
31	0.0007	3.32
32	0.0178	3.39
33	1.26e-09	4.53
34	0.3869	3.19
35	0.0467	3.21
36	2.25e-05	3.73
37	0.0008	3.59
38	0.0842	3.57
40	2.96e-05	3.59
42	0.0003	3.67
43	1.08e-07	2.85
44	0.0002	3.81
45	0.1781	2.38
46	0.0239	2.91
47	0.4314	2.26

Table 1: Calibration and Information Scores of the 43 experts

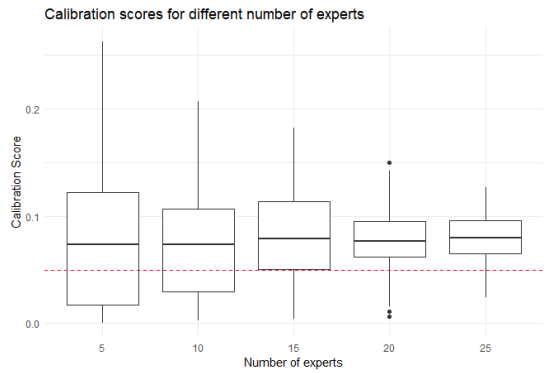
The results show a wide variation in calibration scores and information scores among the experts. While some experts, such as Expert 5 and Expert 6, have high calibration scores, others, like Expert 29 and Expert 33, show extremely low scores. This discrepancy indicates a poor alignment between their predicted intervals and the realizations. The Classical Model uses a threshold of 0.05 to determine whether an expert is considered

statistically accurate. As shown in Table 1, 34 out of the 43 experts have a calibration score below this threshold. This means that, based on their responses to the 15 seed questions, the majority of the experts do not meet the accuracy criterion. However, it's important to note that these results refer only to individual calibration scores and do not yet reflect the performance of decision makers, in which expert judgments are combined to produce a final result.

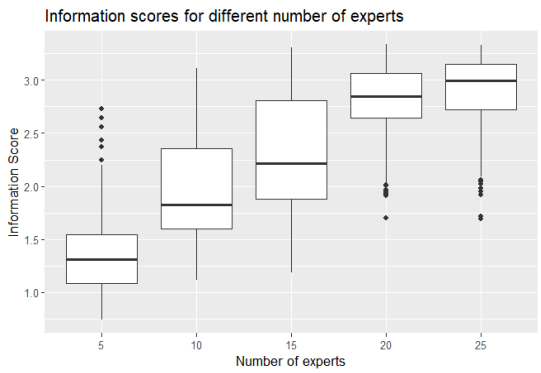
As shown in Table 1, the information score also varies across experts. Expert 19 has the lowest information score of 1.15, indicating that this expert provided relatively wide uncertainty intervals when estimating the 15 seed questions. In contrast, Expert 17 has the highest information score of 4.62, meaning their intervals were the narrowest and thus more precise. However, it is important to keep in mind that a high information score is only meaningful if the expert is also statistically accurate. An expert who is very precise but estimated the questions completely wrong (reflected by a low calibration score) does not contribute reliable input. Therefore, the calibration score is considered more important than the information score.

3.2 Effect of Panel Size on Average Performance

In this part of my analysis, I wanted to investigate how calibration and information scores behave across different panel sizes. Therefore, a sub-sampling analysis was performed. For each selected panel size (5, 10, 15, 20, and 25 experts), experts from the first dataset of 43 experts were randomly selected. Then, the average calibration and information score for each group was calculated. This process was repeated 100 times for each panel size. The results were visualized in the box plots shown in Figure 1 below.



(a) Average calibration scores depending on the number of experts in a panel of size 5,10,15,20, and 25



(b) Average information scores depending on the number of experts in a panel of size 5,10,15,20, and 25

Figure 1: Two box plots of the average calibration score and information score depending on the number of experts in a panel of size 5, 10, 15, 20, and 25.

Figure 1 shows how both the average calibration score and information score change as the number of experts in the panel increases. In Figure 1a, the red dashed line represents the threshold value $\alpha = 0.05$, which is commonly used in the Classical Model. According to the CM, a calibration score above this threshold is considered statistically acceptable, meaning that it's well-calibrated.

As shown in Figure 1b, the information score increases as the number of experts in a panel becomes larger. This suggests, when we only look at the average information score, that larger panel sizes have better performance. However, as discussed earlier in Chapter 1, a high information score is not meaningful if the estimates themselves are inaccurate, that is, if the calibration score is low. For that reason, the calibration score is more important than the information score. Therefore, throughout this thesis, the main focus will be on how the calibration score changes with different panel sizes.

Looking at the calibration score in Figure 1a, panels with 5 or 10 experts show a lot of variation. The boxes are wide, and there are many outliers, indicating that some groups of experts perform very well while other groups score poorly. This means that in smaller panels, the overall result depends heavily on which experts happen to be included.

As the panel size increases to 15 or 20 experts, the spread in calibration scores becomes smaller and therefore more stable, which suggests more stable average calibration scores. Also, most groups have an average calibration score above the threshold 0.05 and therefore are defined as statistically accurate groups. However, when increasing the number to 25 experts, the results remain roughly the same. This suggests that adding more than 20 experts does not lead to noticeable improvements and therefore is not useful anymore.

Based on this analysis, if we only look at the average calibration scores of the experts used, panels consisting of around 15 to 20 experts strike a good balance. These groups tend to produce stable results with scores that are generally above the threshold of 0.05. In other words, they are large enough to reduce the influence of individual outliers, but not so large that additional experts stop contributing to better performance. However, smaller panels with only 5 experts can also perform well. In many cases, the average calibration score for such small groups is above the threshold of 0.05. Nevertheless, smaller panels are more sensitive to expert selection. One group of 5 experts may perform very well, while another group may perform poorly. This makes the outcome less predictable.

3.2.1 Mathematical explanation of the Box plots

The findings from the box plots can also be understood through mathematical explanations.
Suppose we have N experts (in our panel $N = 43$), each with their own calibration score C_j, where $j = 1, 2, .., N$. These scores vary across experts as you could see at the beginning of this chapter in Table 1. We form a group of k experts for $k = 5, 10, 15, 20, 25$, and we calculate the group calibration score $C_k^{(r)}$ for a randomly selected group. This is repeated 100 times so $r = 1, .., 100$. The group calibration score is an average of the individual scores:

$$C_k^{(r)} = \frac{1}{k} \sum_{j=1}^{k} C_j^{(r)}$$

Where $C_j^{(r)}$ is the calibration score of the j-th expert in the r-th sample of k experts. The variance of the average calibration score for a group of size k is given by:

$$Var(C_k^{(r)}) = Var(\frac{1}{k}\sum_{j=1}^{k} C_j^{(r)})$$

If we assume that the individual calibration scores C_j are independent and identically distributed with variance σ^2, then:

$$Var(C_k^{(r)}) = Var(\frac{1}{k}\sum_{j=1}^{k} C_j^{(r)}) = \frac{1}{k^2}Var(C_1^{(r)} + C_2^{(r)} + ... + C_j^{(r)} = \frac{k\sigma^2}{k^2} = \frac{\sigma^2}{k}$$

This shows that as group size k increases, the variance of the group score decreases. This is why you can see in the box plot that for small groups the variance is high and therefore we have more outliers. For larger k the variance is smaller and therefore the box plots are narrower.

Table 2 shows the variance of the calibration scores across 100 randomly selected expert panels for each group size. The results clearly demonstrate that the variance decreases as the number of experts in a panel increases.

Panel Size	**Variance of Calibration Score**
5	0.00522
10	0.00223
15	0.00119
20	0.00057
25	0.00046

Table 2: Variance of the calibration score for different panel sizes (based on 100 subsamples).

3.3 Evaluating Decision Maker Performances Across Different Panel Sizes

Previously, I examined how the average calibration and information scores of experts behave across different panel sizes. In this part of my analysis, I will focus on how the different decision makers perform when the number of experts in a panel increases. Since the calibration score is more important than the information score, I will focus solely on how the calibration score changes with panel size. As mentioned before, a decision maker can be seen as a new expert, which means that we can evaluate their performance by calculating their calibration score. The following 5 different DMs were used in this analysis:

- EWDM: who assigns equal weight to each expert
- GWDM: who assigns more weight to better-performing experts
- IWDM: who assigns more weight to better-performing experts per question
- GWDM_opt: Uses the same method as GWDM and also uses optimization

- IWDM_opt: uses the same method as IWDM ans also uses optimization

So to evaluate how different DMs perform under varying panel sizes, the calibration scores were calculated for each DM using expert subsets of sizes 5,10,15,20, and 25 experts. The results are shown in box plots in Figure 2 to Figure 6 below. Each box represents the distribution of calibration scores for a given number of experts and a specific DM method. The red line was added to show the average pattern as the panel size increases. The horizontal dashed line at 0.05 marks the common threshold used in the Classical Model to judge whether it's statistically well-calibrated. As shown in Figure 2 to Figure 6, a very small percentage of the DMs are actually below the 0.05 threshold.

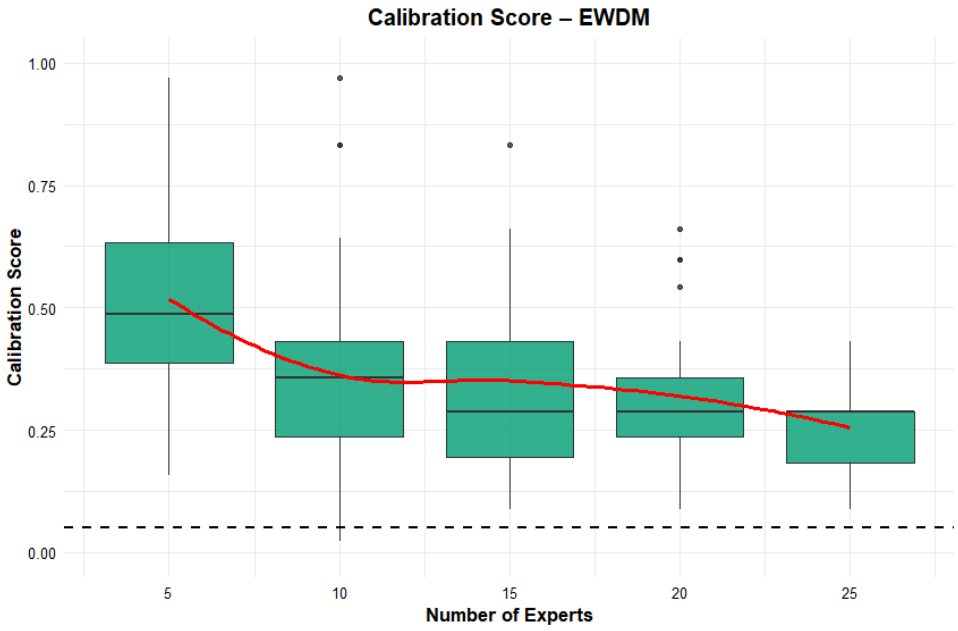


Figure 2: Calibration score of different panel sizes of the EWDM

Figure 2 illustrates how the calibration score of the EWDM changes across different panel sizes. As shown in the figure, the calibration score tends to decrease as the number of experts increases, suggesting that smaller groups tend to perform better in terms of calibration. At the same time, the variation in the calibration scores is noticeably wider for smaller panels, while larger groups show more concentrated and consistent results.

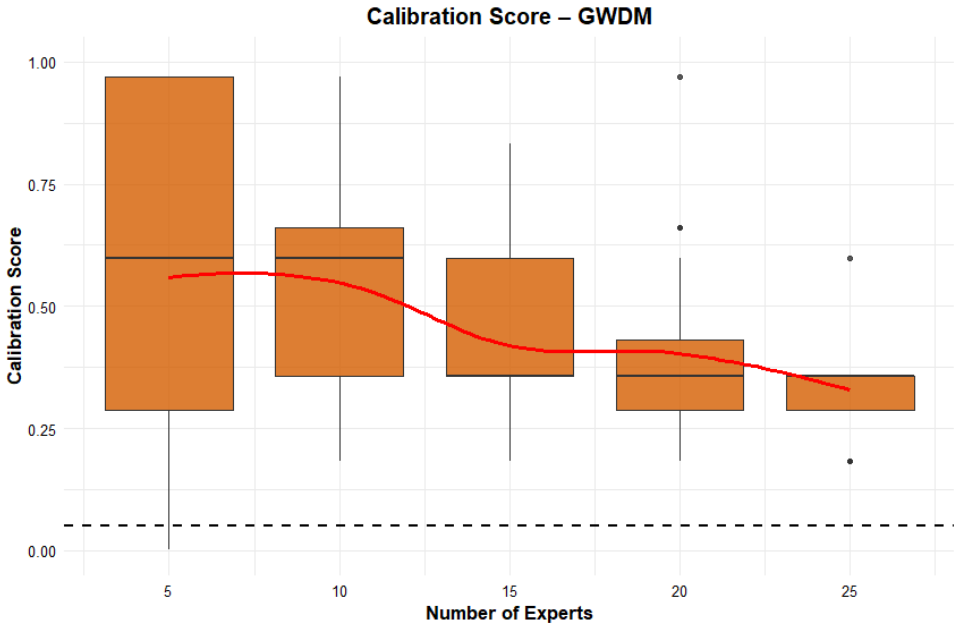


Figure 3: Calibration score of different panel sizes of the GWDM

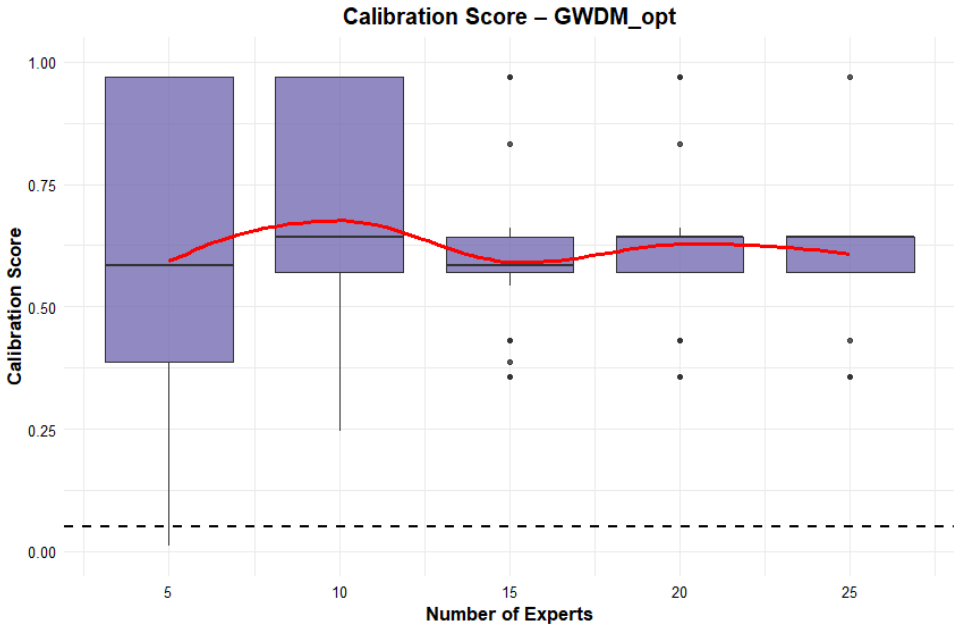


Figure 4: Calibration score of different panel sizes of the GWDM_opt

Figures 3 to 6 show how the calibration scores of the GWDM, GWDM_opt, IWDM, and IWDM_opt change with increasing panel sizes. In all cases, we observe high variability in calibration scores for smaller groups, particularly panels of 5 experts. This variation decreases as the panel size grows, indicating more stable performance in larger groups. All DMs show a peak in calibration performance around a panel size of 10 experts. For the GWDM and IWDM in Figures 3 and 5, the calibration score clearly starts to decline after this point. While the scores remain above the 0.05 threshold, meaning the DM is still statistically accurate, smaller groups appear to perform better in terms of calibration. In contrast, the GWDM_opt and IWDM_opt in Figures 4 and 6 also show a peak around

10 experts, but the calibration scores remain relatively stable afterwards. This suggests that larger groups still perform well, but do not necessarily add significant value beyond what smaller panels provide.

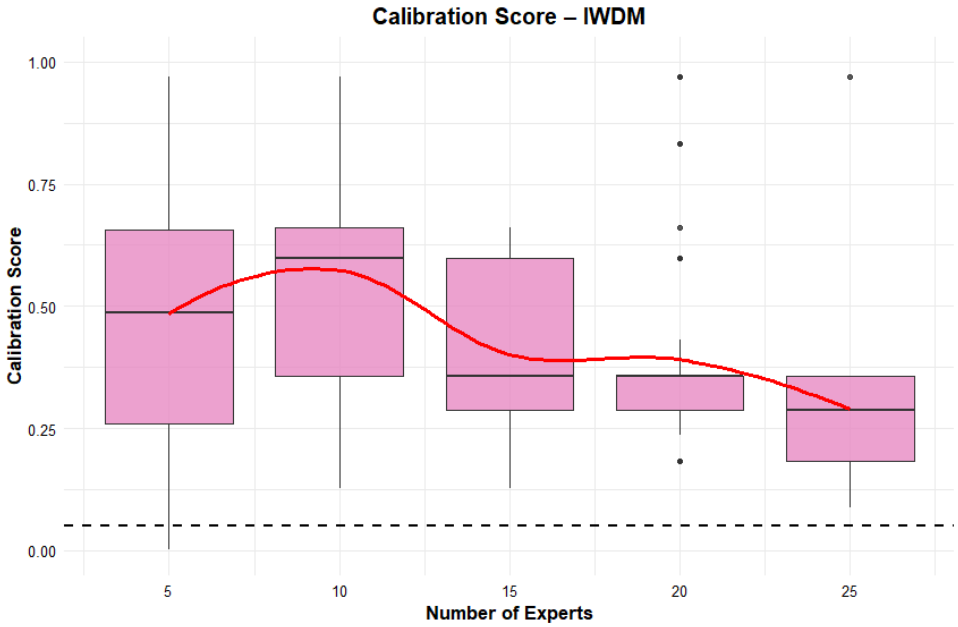


Figure 5: Calibration score of different panel sizes of the IWDM

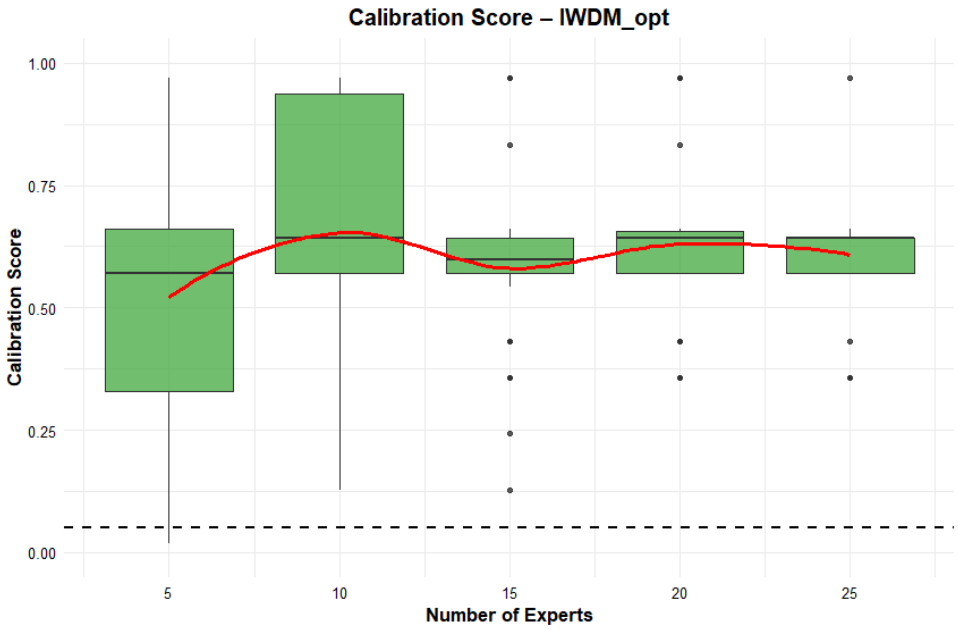


Figure 6: Calibration score of different panel sizes of the IWDM_opt

3.3.1 Comparing The Different Decision Makers

When comparing the different Decision Makers, a clear difference in performance can be observed. Among all methods, the EWDM of Figure 2 performs the worst. Its calibration scores are generally lower across all panel sizes, indicating that giving all experts

the same weight can lead to poorer performance when less well-calibrated experts are included. However, the overall calibration scores are higher than the threshold of 0.05 and therefore this DM is still useful to use.
Both the GWDM and IWDM of Figure 3 and Figure 5 show better results than the EWDM. These methods assign more weight to better performing experts, which lead to better calibration scores.
The optimized versions, GWDM_opt and IWDM_opt of Figure 4 and Figure 6, perform the best overall (especially for large groups). You can see this by looking at the median (the thick black line) for each group size. They consistently achieve high calibration scores across all panel sizes and are more robust to variation in panel composition. These DMs clearly outperform the non-optimized versions, making them the most reliable methods in this comparison.

In Figure 1a, we saw how the average calibration score changes with panel sizes. When only averaging the calibration scores for different group sizes, many groups fail to reach the threshold of 0.05. In contrast, when using DMs, nearly all panels achieve calibration scores above this threshold. This demonstrates the value of applying weighted combinations.

3.3.2 Determining The Optimal Amount Of Experts In A Panel

In Figure 2 to 5, we can also see how the calibration scores of the different decision makers change with different panel sizes. For the EWDM in Figure 2, the calibration score tends to decline as the number of experts increases. This suggests that simply adding more equally weighted experts can reduce overall accuracy, especially if poorly calibrated experts start to dominate the panel.
For the GWDM and IWDM in Figure 3 and Figure 5, a similar trend is visible. Both methods show relatively high calibration scores for smaller groups, with the best performance occurring around 10 experts. In these cases, the scores remain consistently above the 0.05 threshold, suggesting that panels of 10 experts may be a good target for panel size. Beyond this point, calibration scores tend to drop slightly, but still remain acceptable. Therefore, we could say that after 10 experts is not useful anymore to add more experts, but it still gives a good calibration score.
The optimized DMs in Figure 4 and 6 are less affected by changes in panel size. Even as more experts are added, the calibration scores remain high and stable. However, also with these DMs the performance tends to peak around 10 experts. This indicates that while these methods are more tolerant of larger groups, increasing the panel beyond 10 experts offers little additional value in terms of calibration. Base on these results, a panel size of approximately 10 experts appears to be sufficient for achieving strong calibration performance.

Since the earlier box plots suggested that a panel size of around 10 experts may be optimal, I decided to take a closer look at smaller panels and smaller differences in size. Panels consisting of 2 to 10 experts were analyzed to better understand how performance changes within this range. Below in Figure 7 to 11, you can find the different box plots for the different Decision Makers.

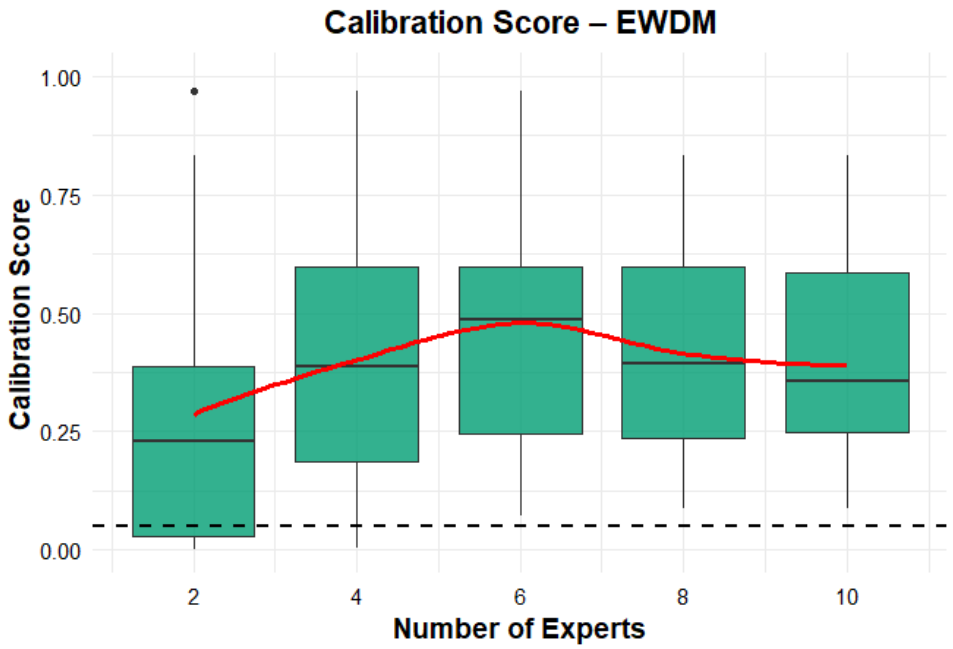


Figure 7: Calibration score of different panel sizes of the EWDM for panels of size 2, 4, 6, 8, and 10

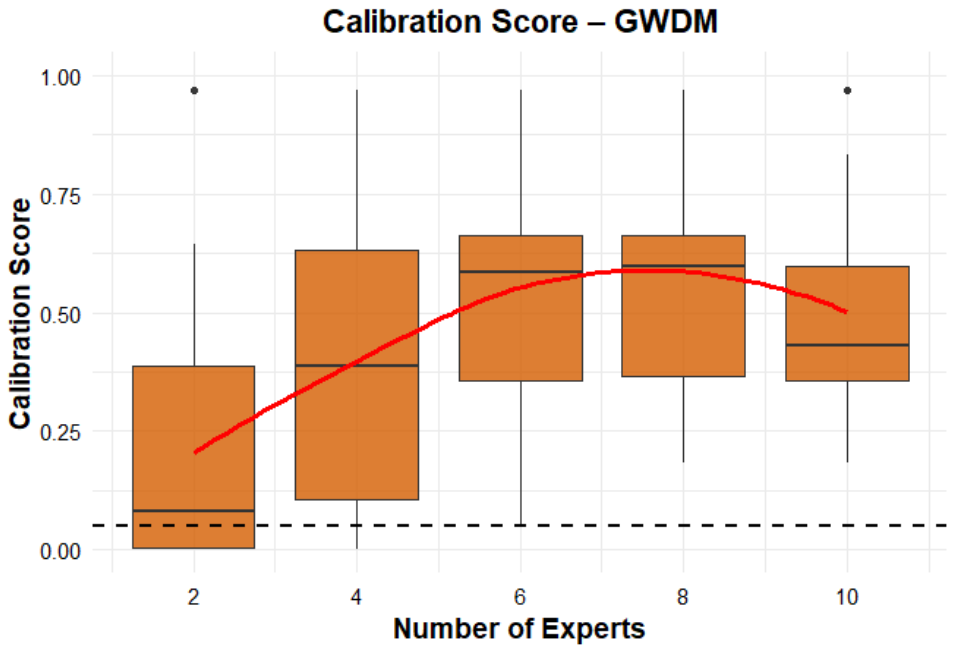


Figure 8: Calibration score of different panel sizes of the GWDM for panels of size 2, 4, 6, 8, and 10

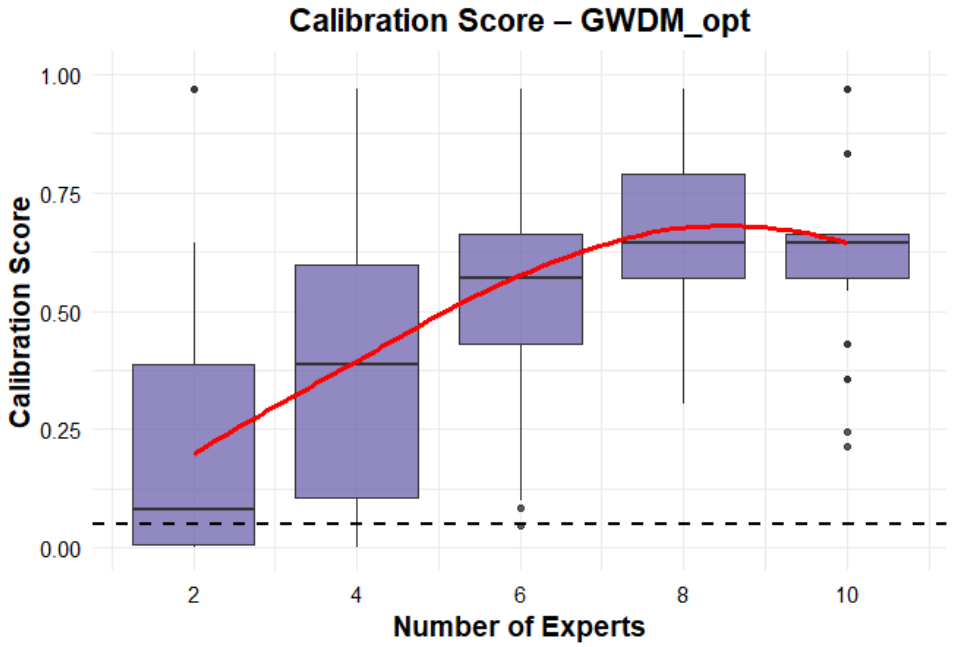


Figure 9: Calibration score of different panel sizes of the GWDM_opt for panels of size 2, 4, 6, 8, and 10

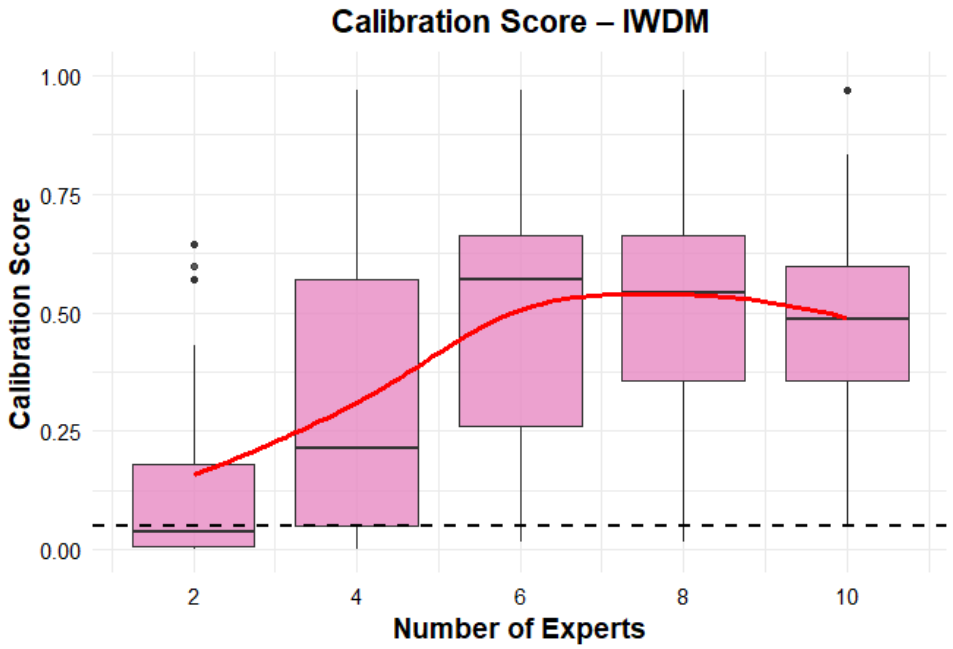


Figure 10: Calibration score of different panel sizes of the IWDM for panels of size 2, 4, 6, 8, and 10

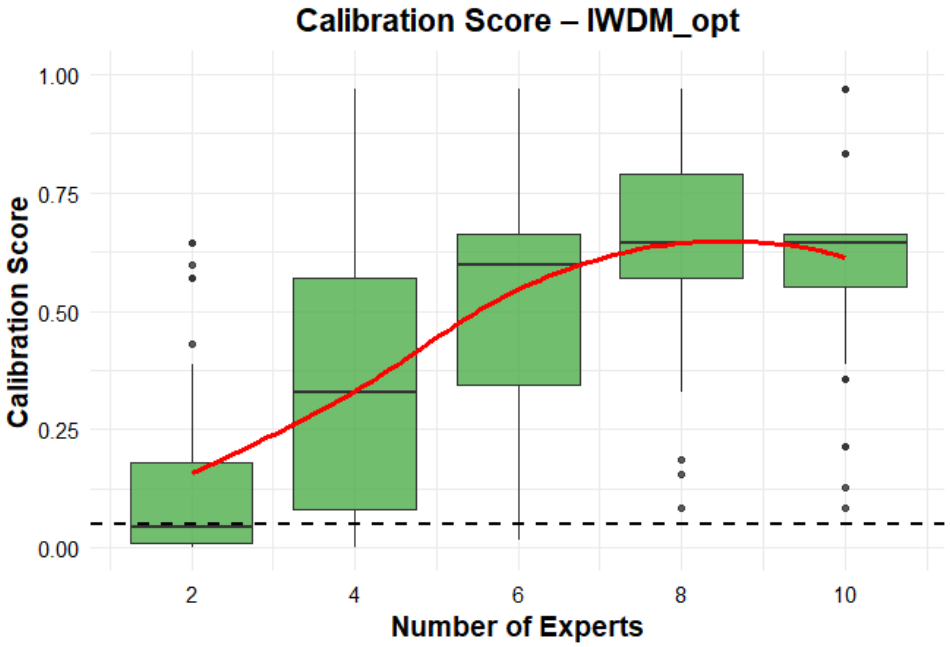


Figure 11: Calibration score of different panel sizes of the IWDM_opt for panels of size 2, 4, 6, 8, and 10

The box plots in Figure 7 to Figure 11 show that the calibration scores of the different Decision Makers peak around 6 to 8 experts. After this, a slight decline is visualized. This suggests that the optimal number of experts in a panel is 6 to 8 experts.
In many studies, a minimum of 4 experts is commonly recommended for structured expert judgment panels. In my analysis, panels of 4 experts often performed reasonably well in terms of calibration score. However, the variance at this size is high, and some panels fall below the 0.05 threshold, indicating inconsistent reliability. Therefore, based on these findings, I would recommend using at least 6 to 8 experts in a panel. This range appears to offer a good balance between maintaining high calibration scores and ensuring more stable and consistent performance. However, it's important to note that even within the 6 to 8 expert range, there is still considerable variation in performance. While the average calibration scores of the DMs tend to peak in this range, some groups still perform noticeably worse than others.

3.4 Performance DMs for the different panels for different pathogens.

To assess how well the different DMs perform in practice and to see if my first analysis corresponds to real panels, the second dataset provided by the RIVM was analyzed. This dataset consists 26 separate expert panels, each focused on estimating the uncertainty around the spread of a specific pathogen. The experts involved in these panels are the same individuals as those used in the first part of the analysis. In Figures 12 and 13, you can find the second dataset.

BRUCL	CAMPY	LISTM	LEGIO	LEPTO	THDSA	SAENT	SATYP	OTSAL	STECO	STECN	OTECO	YERSI
Exp09	Exp09	Exp09	Exp09	Exp09	Exp09	Exp17	Exp17	Exp17	Exp17	Exp17	Exp17	Exp01
Exp33	Exp12	Exp12	Exp01	Exp44	Exp11	Exp12	Exp12	Exp12	Exp11	Exp11	Exp02	Exp06
Exp15	Exp11	Exp11	Exp19	Exp06	Exp01	Exp11	Exp11	Exp11	Exp02	Exp02	Exp01	Exp15
Exp8	Exp02	Exp02	Exp45	Exp38	Exp06	Exp02	Exp02	Exp02	Exp01	Exp01	Exp47	Exp46
Exp38	Exp20	Exp01	Exp38	Exp40	Exp30	Exp05	Exp05	Exp01	Exp47	Exp47	Exp28	
	Exp05	Exp27	Exp37	Exp43	Exp15	Exp01	Exp01	Exp47	Exp28	Exp28	Exp31	
	Exp01	Exp28	Exp40	Exp36	Exp37	Exp47	Exp47	Exp10	Exp31	Exp31	Exp30	
	Exp10	Exp25	Exp46		Exp46	Exp10	Exp10	Exp06	Exp30	Exp30	Exp18	
	Exp06	Exp31				Exp06	Exp06	Exp31	Exp26	Exp18	Exp26	
	Exp27	Exp30				Exp27	Exp31	Exp30	Exp15	Exp26	Exp15	
	Exp24	Exp45				Exp31	Exp30	Exp18	Exp34	Exp15	Exp14	
	Exp25	Exp26				Exp30	Exp18	Exp26	Exp37	Exp34	Exp43	
	Exp31	Exp15				Exp18	Exp26	Exp15	Exp14	Exp37	Exp46	
	Exp18	Exp34				Exp26	Exp15	Exp22	Exp43	Exp14		
	Exp26	Exp16				Exp15	Exp22	Exp16	Exp46	Exp43		
	Exp15	Exp38				Exp22	Exp16	Exp37		Exp46		
	Exp16	Exp37				Exp16	Exp14	Exp14				
	Exp38	Exp14				Exp37	Exp43	Exp43				
	Exp40	Exp43				Exp14		Exp46				
	Exp14	Exp46				Exp43						
	Exp43	Exp23				Exp46						
	Exp46											
5	22	21	8	7	8	21	18	19	15	16	13	4

Figure 12: Panels for the first 13 pathogens

VIBRI	BCTOX	CLPET	STAUT	ASTRV	HEPAV	HEPEV	ROTAV	NOROV	SAPOV	CRYPT	GIARD	TOXP
Exp12	Exp28	Exp02	Exp02	Exp32	Exp32	Exp32	Exp32	Exp32	Exp32	Exp01	Exp01	Exp01
Exp18	Exp31	Exp25	Exp20	Exp06	Exp01	Exp01	Exp01	Exp01	Exp06	Exp44	Exp44	Exp44
Exp15	Exp26	Exp31	Exp28	Exp29	Exp25	Exp04	Exp06	Exp06	Exp29	Exp6	Exp06	Exp04
Exp46	Exp43	Exp26	Exp31	Exp13	Exp29	Exp29	Exp29	Exp25	Exp35	Exp03	Exp03	Exp03
	Exp46	Exp34	Exp34	Exp43	Exp34	Exp35	Exp35	Exp31	Exp13	Exp19	Exp19	Exp25
	Exp36	Exp40	Exp37		Exp35	Exp13	Exp37	Exp29		Exp8	Exp8	Exp8
		Exp43	Exp43		Exp37	Exp43		Exp34		Exp42	Exp42	Exp42
			Exp36		Exp40	Exp46		Exp35		Exp23	Exp23	Exp43
					Exp46			Exp13				Exp46
								Exp37				Exp23
								Exp43				
								Exp46				
4	6	7	8	5	9	8	6	12	5	8	8	10

Figure 13: Panels for the last 13 pathogens

In the dataset, there are 26 different pathogens. A pathogen is any organism or agent that can cause disease. This includes microscopic organisms such as viruses, bacteria, fungi, and parasites. The dataset includes many pathogens of which are also evaluated in the expert elicitation study by Havelaar et al. (2008). These pathogens were selected based on their public health relevance in the Netherlands, either due to high disease burden, potential for foodborne transmission, or the need for improved source attribution.
The RIVM studies these pathogens to better understand how people get infected and how much of the illness is caused by food. This is important for making good food safety policies. As Havelaar et al. (2008) explain, it is difficult to know the exact source of every infection because pathogens can spread in different ways. Not only through food, but also through the environment, animals, people, or travel. That is why expert judgment was used in their study to estimate how much each pathway contributes to the disease.
So the pathogens in my dataset are relevant for the RIVM to better understand the risks and improve food safety in the Netherlands.

For every panel, the calibration scores of the 5 different DMs were calculated. In Table 3, you can find the calibration scores of the different panels for the different DMs.

Pathogen	NE	EWDM	GWDM	GWDM_opt	IWDM	IWDM_opt
BRUCL	5	0.0428	0.0485	0.0842	0.0485	0.0842
CAMPY	22	0.2880	0.5976	0.6432	0.6608	0.6608
LISTM	21	0.3579	0.3579	0.3869	0.3579	0.5976
LEGIO	8	0.5976	0.4314	0.6608	0.2357	0.5976
LEPTO	7	0.4314	0.9698	0.9698	0.9698	0.9698
THDSA	8	0.2880	0.9698	0.5710	0.9698	0.9698
SAENT	21	0.2880	0.3579	0.6432	0.3579	0.6608
SATYP	18	0.2880	0.3579	0.6432	0.3579	0.6608
OTSAL	19	0.2880	0.3579	0.5710	0.3579	0.5710
STECO	15	0.3579	0.3579	0.4314	0.2880	0.4314
STECN	16	0.3579	0.3579	0.4314	0.2880	0.4314
OTECO	13	0.5976	0.2880	0.4314	0.3579	0.4314
YERSI	4	0.6608	0.5710	0.5710	0.5710	0.5710
VIBRI	4	0.5976	0.5976	0.5976	0.9698	0.9698
BCTOX	6	0.5433	0.9698	0.9698	0.5433	0.5433
CLPET	7	0.1824	0.5976	0.5710	0.5433	0.5433
STAUT	8	0.2880	0.4314	0.6608	0.2357	0.6608
ASTRV	5	0.4314	0.5976	0.9698	0.5976	0.9698
HEPAV	9	0.2880	0.5976	0.3869	0.5976	0.5710
HEPEV	8	0.5433	0.8331	0.9698	0.5433	0.5710
ROTAV	6	0.4314	0.5976	0.9698	0.5976	0.9698
NOROV	12	0.1824	0.2880	0.5710	0.5976	0.8331
SAPOV	5	0.4314	0.5976	0.9698	0.6608	0.9698
CRYPT	8	0.0301	0.3579	0.3579	0.3579	0.3579
GIARD	8	0.0301	0.3579	0.3579	0.3579	0.3579
TOXP	10	0.2357	0.9698	0.9698	0.5710	0.8331

Table 3: Calibration scores for all five decision makers across 26 pathogen panels. Here, NE means the number of experts in the panel.

Table 3 shows variation among the different pathogen panels. For instance, the BRUCL panel had significantly lower calibration scores compared to all the other panels. In contrast, the THDSA and LEPTO panels achieved calibration scores close to 0.97 across multiple DMs, indicating excellent performance.

The results are also visualized below in Figure 14 to 18 for the different DMs. Each blue point represents the calibration score of a panel, and the red line shows the average calibration score of the DMs per panel size. These plots provide insight into how panel size might influence the quality of combined expert judgments.

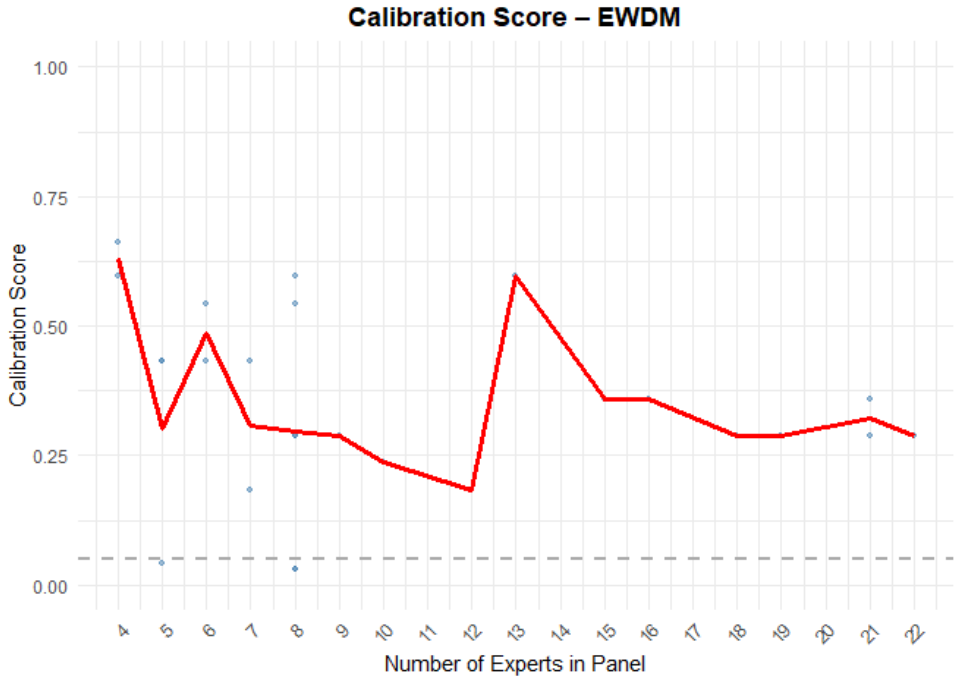


Figure 14: Calibration scores of the panels for various pathogens using EWDM.

In Figure 14, we observe the variation in the calibration scores of the EWDM across different panel sizes. Generally, smaller panels exhibit greater variability. At times, their calibration scores perform well, while at other times, they fall below the 0.05 threshold. As the panel size increases, the average calibration score tends to decline, but it consistently remains above the 0.05 threshold.

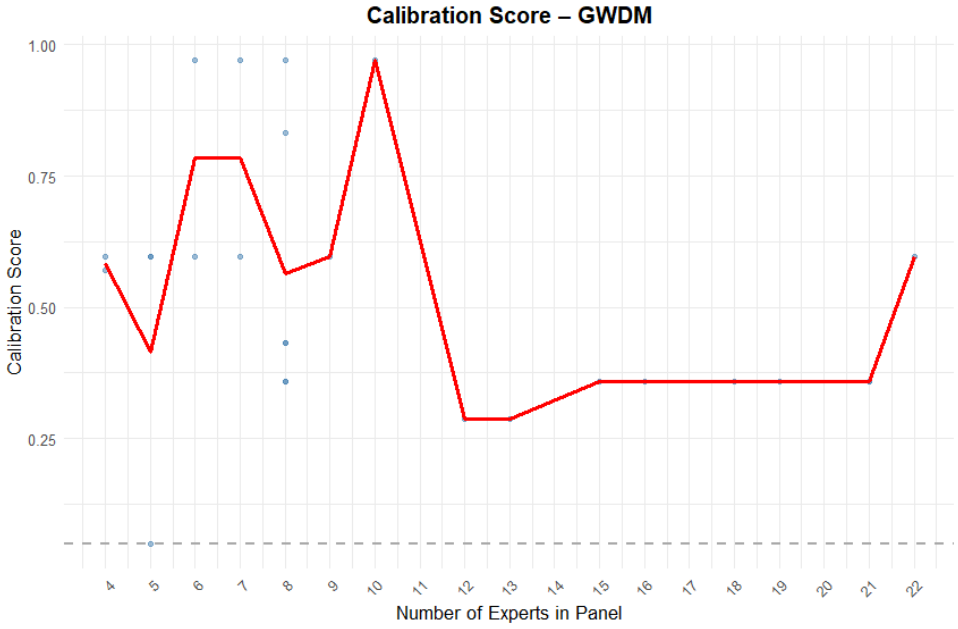


Figure 15: Calibration scores of the panels for various pathogens using GWDM.

In Figure 15, the calibration scores for the GWDM show the highest values for panels sized between 6 and 10. However, there is still considerable variation among these smaller

groups. After reaching 10 experts, the calibration score begins to decline but remains above the 0.05 threshold.

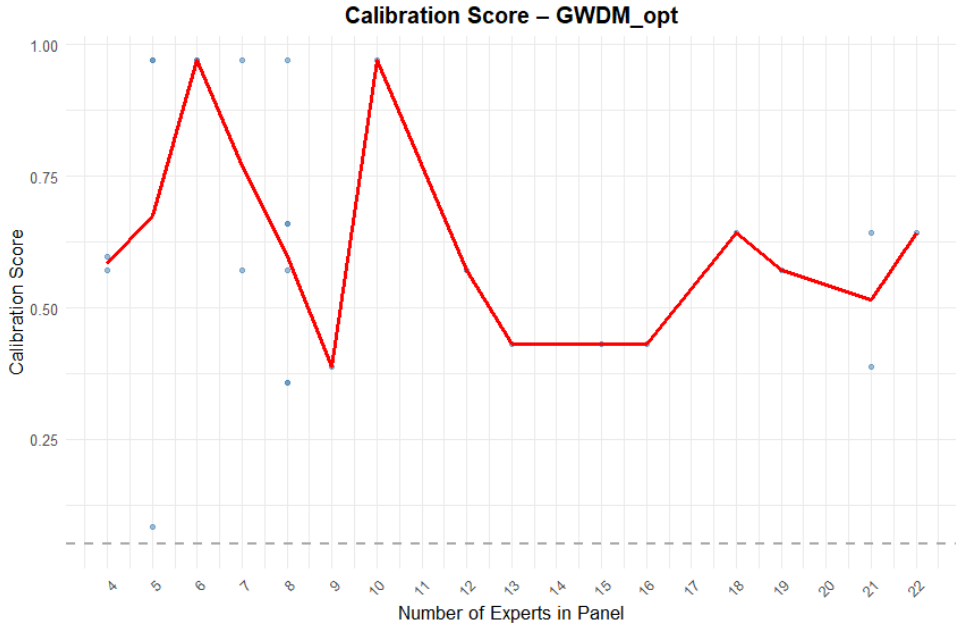


Figure 16: Calibration scores of the panels for various pathogens using GWDM_opt.

The GWDM_opt in Figure 16 demonstrates a higher calibration score in most panels. This figure exhibits a peak with 6 to 10 experts, after which the calibration score decreases beyond 10 experts. However, also in this DM the calibration score remains above the 0.05 threshold

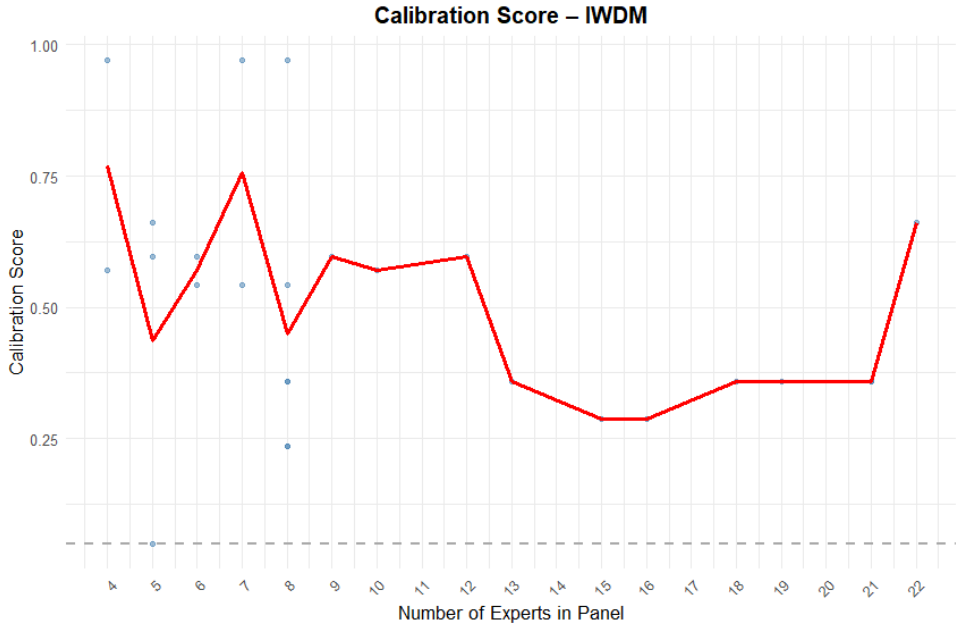


Figure 17: Calibration scores of the panels for various pathogens using IWDM.

The IWDM shown in Figure 17, has good calibration scores for small panels. However, there is considerable variation among these small panels. At times, the calibration score is very high, while at other times, it falls below the threshold of 0.05.

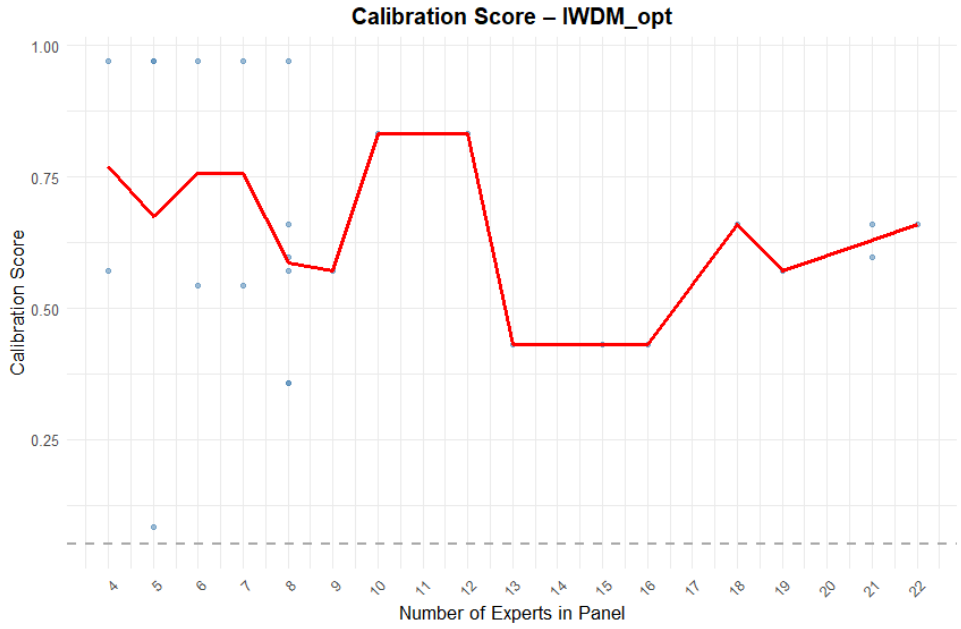


Figure 18: Calibration scores of the panels for various pathogens using IWDM_opt.

For the IWDM_opt in Figure 18 shows high calibration scores for nearly all panel sizes. Here, the calibration score drops after 12 experts, but still remains high.

When comparing the different DMs, the results are largely consistent with the earlier analysis. The EWDM once again performs the worst in most cases. For example, in the CRYPT panel, the calibration score for EWDM falls below the 0.05 threshold, while for all other DMs, the score remains above it.
The GWDM and IWDM continue to show similar behavior, with only small differences in calibration scores between them. These methods generally produce better results than EWDM, although their performance can still vary slightly depending on the specific panel. The optimized DMs again outperform all the other DMs. Across different panels and group sizes, they consistently achieve higher calibration scores. This confirms the earlier conclusion that the optimized methods are more reliable and robust across different settings.

When examining the relationship between calibration scores and panel sizes in this second dataset, several observations can be made. In all the different decision-makers (DMs), small panels consistently achieve high calibration scores. However, there is significant variation among these small panels, as noted in my initial analysis. While some small panels perform exceptionally well in terms of calibration, others score below the 0.05 threshold. Due to this variation in smaller panels, a clear performance peak does not always appear around the 6-8 experts. While panels of this size often perform well, they also sometimes perform significant lower. To better understand these differences in performance, the next section will take a closer look at the types of experts included in each panel.

3.5 Closer look to the experts in each panel

To clarify the variation between the calibration scores in smaller panels, we will examine the types of experts used. We do this by looking at the experts' assessments provided. Each expert provided estimates by giving the 5th, 50th, and 95th percentiles for 15 seed questions. Table 4 presents the proportions of correct (when realization $r \in [q_5, q_{95}]$), overestimated ($r < q_5$), and underestimated ($r > q_{95}$) responses for each expert. The difference between the number of overestimates and underestimates was used as a measure of expert bias. A positive difference indicates a tendency to overestimate, while a negative difference suggests a tendency to underestimate.

Expert ID	Perc_Good	Perc_Overestimate	Perc_Underestimate	Difference
1	40	33.3	26.7	1
2	60	13.3	26.7	-2
3	20	33.3	46.7	-2
4	73.3	6.7	20	-2
5	80	6.7	13.3	-1
6	80	6.7	13.3	-1
8	46.7	13.3	40	-4
9	33.3	20	46.7	-4
10	33.3	20	46.7	-4
11	60	20	20	0
12	66.7	26.7	6.7	3
13	66.7	20	13.3	1
14	33.3	33.3	33.3	0
15	53.3	20	26.7	-1
16	53.3	26.7	20	1
17	33.3	20	46.7	-4
18	60	33.3	6.7	4
19	80	13.3	6.7	1
20	33.3	40	26.7	2
22	40	20	40	-3
23	66.7	13.3	20	-1
24	40	26.7	33.3	-1
25	60	13.3	26.7	-2
26	60	20	20	0
27	60	13.3	26.7	-2
28	86.7	13.3	0	2
29	46.7	20	33.3	-2
30	40	40	20	3
31	53.3	6.7	40	-5
32	66.7	13.3	20	-1
33	20	20	60	-6
34	80	6.7	13.3	-1
35	66.7	6.7	26.7	-3
36	40	20	40	-3
37	53.3	13.3	33.3	-3
38	66.7	13.3	20	-1
40	46.7	13.3	40	-4
42	46.7	26.7	26.7	0
43	26.7	40	33.3	1
44	46.7	13.3	40	-4
45	93.3	6.7	0	1
46	60	20	20	0
47	93.3	6.7	0	1

Table 4: Percentages of good estimations, overestimations and underestimations per expert

Table 4, shows that most experts made a mix of good, overestimated, and underestimated predictions, but some showed clear behavioral patterns. For instance, Expert 33 had only 20% good answers and significantly more underestimates than overestimates. Referring back to Table 1, we see that expert 33 has a calibration score of $1.26e^{-09}$, which is the lowest calibration score among all 43 experts. This suggests a potential link between systematic underestimation or overestimation and poor calibration scores.

Expert ID	Difference	Calibration score
8	-4	1.05e-04
9	-4	1.48e-06
10	-4	1.48e-06
12	3	0.0222
17	-4	1.48e-06
18	4	0.0017
22	-3	3.12e-05
30	3	2.25e-05
31	-5	0.0007
33	-6	1.26e-09
35	-3	0.0467
36	-3	2.25e-05
37	-3	0.0008
40	-4	2.96e-05
44	-4	0.0002

Table 5: Underestimators and overestimators with corresponding calibration score

As shown in Table 5, the experts who can be classified as underestimators or overestimators were listed, meaning their difference is less than or equal to -3 or greater than or equal to 3. For each of these experts we observe that their individual calibration score is below the threshold 0.05.
This consistent pattern suggest a link between extreme prediction bias and poor statistical accuracy. In other words, experts who systematically underestimate or overestimate tend to receive significantly lower calibration scores.

3.5.1 Mathematical explanation

As mentioned above, all experts identified as either underestimators or overestimators (based on their difference being ≤ -3 or ≥ 3) have calibration scores below the threshold 0.05 in my dataset. This pattern also makes sense mathematically when we look back at how the calibration score is computed, as explained in Section 1.2.

To show this, consider expert 33, who is classified as an underestimator. This expert consistently provides lower estimates compared to the actual outcomes. In Table 6, expert 33 gave for each of the 15 seed questions the 5th, 50th, and 95th percentiles. The actual realizations were also included.

Expert ID	Question	5th percentile	50th percentile	95th percentile	realization
33	1	810000	845000	860000	864653
33	2	325	350	365	370
33	3	65000	85000	90000	57000
33	4	1	5	8	9
33	5	19000	21000	22000	21224
33	6	750	775	825	1165
33	7	10	15	20	4
33	8	0,180	0,240	0,260	0,182
33	9	0,45	0,52	0,55	0,618
33	10	0,005	0,012	0,016	0,05
33	11	0,270	0,320	0,350	0,15
33	12	0,06	0,08	0,11	0,1
33	13	0,005	0,007	0,008	0,012
33	14	0,030	0,040	0,060	0,2727
33	15	890	910	940	1500

Table 6: percentiles and realization of expert 33

As mentioned in section 1.2, the 3 percentiles divide the range of possible outcomes into four probability intervals. Then we check in which interval the realization of each question falls, and we count how many times each interval is hit. Dividing these counts by the number of seed questions (in our case 15) gives us the following observed proportions: $S = (\frac{3}{15}, \frac{1}{15}, \frac{2}{15}, \frac{9}{15})$.
The expected proportions, under perfect calibration, are: $p = (0.05, 0.45, 0.45, 0.05)$. Here we can already see that more than 50% of the realizations fall above the 95th percentile (in the 4th range), while we would only expect 5% in that range. This leads to a large Kullback-Leibler divergence:

$$l(S,p) = \sum_{k=1}^{4} S_k \ln\left(\frac{S_k}{p_k}\right) = \frac{3}{15}\ln(\frac{\frac{3}{15}}{\frac{1}{20}}) + \frac{1}{15}\ln(\frac{\frac{1}{15}}{\frac{9}{20}}) + \frac{2}{15}\ln(\frac{\frac{2}{15}}{\frac{9}{20}}) + \frac{9}{15}\ln(\frac{\frac{9}{15}}{\frac{1}{20}}) \approx 1.479 \quad (6)$$

Therefore we also get a higher value of the test statistic:

$$T = 2M \cdot l(S,p) = 2 \cdot 15 \cdot 1.479 = 44.37 \quad (7)$$

And this results in a lower p-value, which is the expert's calibration score:

$$Cal(e) = 1 - F(44.37) \approx 1.26 \cdot 10^{-9} \quad (8)$$

As shown in this example, the observed proportions S for experts identified as either overestimators or underestimators tend to fall more frequently into the outer intervals S_1 or S_4. Since these intervals are only expected to contain 5% of the realizations each under perfect calibration, such deviations lead to a large difference between the observed and expected proportions. This results in a high test statistic and, therefore in a low calibration score.

So far, we have shown that all experts classified as underestimators and overestimators have low calibration scores. However, the reverse is not always true. A low calibration score does not necessarily imply that an expert is an underestimator or overestimator. Consider for example, expert 15, who is not classified as an underestimator or overestimator. In Table 7, we see the 5th, 50th, and 95th percentile of all the 15 seed questions of expert 15. The realizations are also included.

Expert ID	Question	5th percentile	50th percentile	95th percentile	realization
15	1	550000	750000	950000	864653
15	2	310	350	370	370
15	3	57000	60000	70000	57000
15	4	2	10	25	9
15	5	15000	20000	22000	21224
15	6	850	900	1100	1165
15	7	10	15	25	4
15	8	0,35	0,42	0,60	0,182
15	9	0,57	0,62	0,68	0,618
15	10	0,01	0,02	0,02	0,05
15	11	0,31	0,35	0,38	0,15
15	12	0,05	0,10	0,15	0,1
15	13	0,00	0,01	0,01	0,012
15	14	0,07	0,16	0,25	0,2727
15	15	1000	1400	2000	1500

Table 7: percentiles and realization of expert 15

This gives us the following observed proportion: $S = (\frac{3}{15}, \frac{3}{15}, \frac{4}{15}, \frac{5}{15})$. The expected proportions under perfect calibration are: $p = (0.05, 0.45, 0.45, 0.05)$. Using the observed proportion and the expected proportion, we can compute the Kullback-Leibler divergence:

$$l(S,p) = \sum_{k=1}^{4} S_k \ln\left(\frac{S_k}{p_k}\right) = \frac{3}{15}\ln(\frac{\frac{3}{15}}{\frac{1}{20}}) + \frac{3}{15}\ln(\frac{\frac{3}{15}}{\frac{9}{20}}) + \frac{4}{15}\ln(\frac{\frac{4}{15}}{\frac{9}{20}}) + \frac{5}{15}\ln(\frac{\frac{5}{15}}{\frac{1}{20}}) \approx 0.6079 \quad (9)$$

Then the test statistic becomes:

$$T = 2M \cdot l(S,p) = 2 \cdot 15 \cdot 0.6079 = 18.237 \quad (10)$$

Using this test statistic, the p-value can be computed and this p-value is the calibration score of the expert

$$Cal(e) = 1 - F(18.237) \approx 0.0004 \quad (11)$$

As you can see, the manually computed calibration score for expert 15 does not completely match the R outputs of 0.0005, which was shown in Table 1. This small difference can be attributed to rounding and the higher numerical precision used by R when evaluating the chi-squared cumulative distribution.
Importantly, both results lead to the same interpretation, as they fall below the threshold of 0.05.

So the example above illustrates that experts who are not classified as underestimators or overestimators can still receive a low calibration score. Although this expert's estimates are not strongly biased in one direction, the realizations often fall outside the central intervals of their predicted ranges. As a result, the observed proportions deviate significantly from the expected probabilities, leading to a high test statistic and thus a low calibration score.

3.5.2 Closer look at pathogen panels

In Table 3, we observed that one panel, consisting of five experts, had for the different decision makers a significantly lower calibration score compared to all the other panels. This panel is known as the BRUCL panel. In this section, the experts involved in the BRUCL panel were closely examined to determine whether the low calibration score is due to underestimation, overestimation, or other influencing factors.

Below, you will find a table listing the experts in this panel, along with the proportions of good estimations, overestimations, and underestimations. The table also includes the Difference, which represents the number of overestimated questions minus the number of underestimated questions.

Expert ID	Perc_Good	Perc_Overestimate	Perc_Underestimate	Difference
8	46.7	13.3	40	-4
9	33.3	20	46.7	-4
15	53.3	20	26.7	-1
33	20	20	60	-6
38	66.7	13.3	20	-1

Table 8: Percentage of good estimations, overestimations, and underestimations of the BRUCL panel

All the experts in this panel have a negative value for their differences, indicating that each expert has more underestimations than overestimations. As defined earlier, we classify an expert as an underestimator if their difference is less than or equal to -3. In this panel, three out of the five experts (8,9 and 33) meet this criterion. Therefore, the poor calibration of the BRUCL panel may be attributed to the presence of these underestimators. As mentioned in section 3.5.1 all the underestimators have a low individual calibration score. In addition, expert 15 also performs poorly, with a calibration score of 0.0005, which is below the 0.05 threshold. Only expert 38 achieves a calibration score of 0.0842, which is the only score in this panel above the threshold. This example illustrates that the presence of underestimators in a panel can contribute to low calibration scores of the different decision makers however, it's not the only possible cause, since even experts who are not classified as underestimators may still perform poorly.

3.5.3 Effect of underestimators on Decision Maker Performance

To further explore the relationship between panel composition and the calibration score of the Decision Makers, I investigate whether an increasing percentage of underestimators in a panel corresponds to a decrease in the calibration score of the DMs. Since my data only has 3 overestimators, I decided to only focus on the effect of underestimators. To do this, I generated scatter plots for each type of decision maker, where each blue point represents a panel and shows the calibration score plotted against the percentage of underestimators in that panel. A linear regession line (the red line) is included to help identify trends. Figures 19 to 22 show the results for the five decision makers.

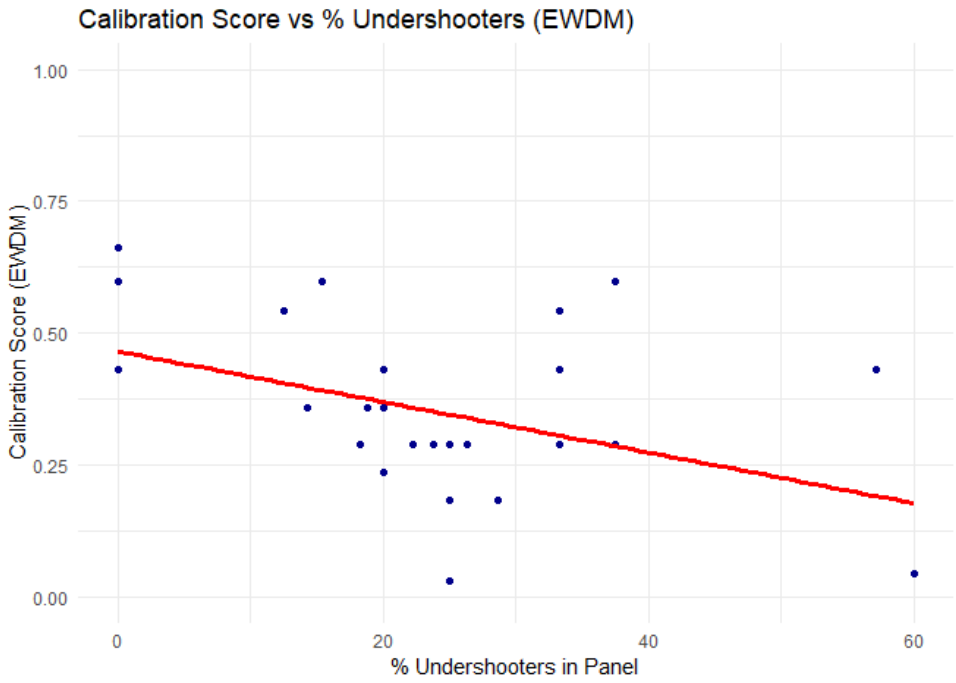


Figure 19: Calibration scores vs proportion of underestimators for EWDM

In Figure 19, the EWDM shows a clear negative trend. Panels with a higher percentage of underestimators tend to have lower calibration scores. Since this method gives equal weight to all experts in the panel, the presence of poorly calibrated individuals (such as underestimators) has a direct negative effect on the final distribution. The regression line suggests that calibration drops significantly as the proportion of underestimators increases.

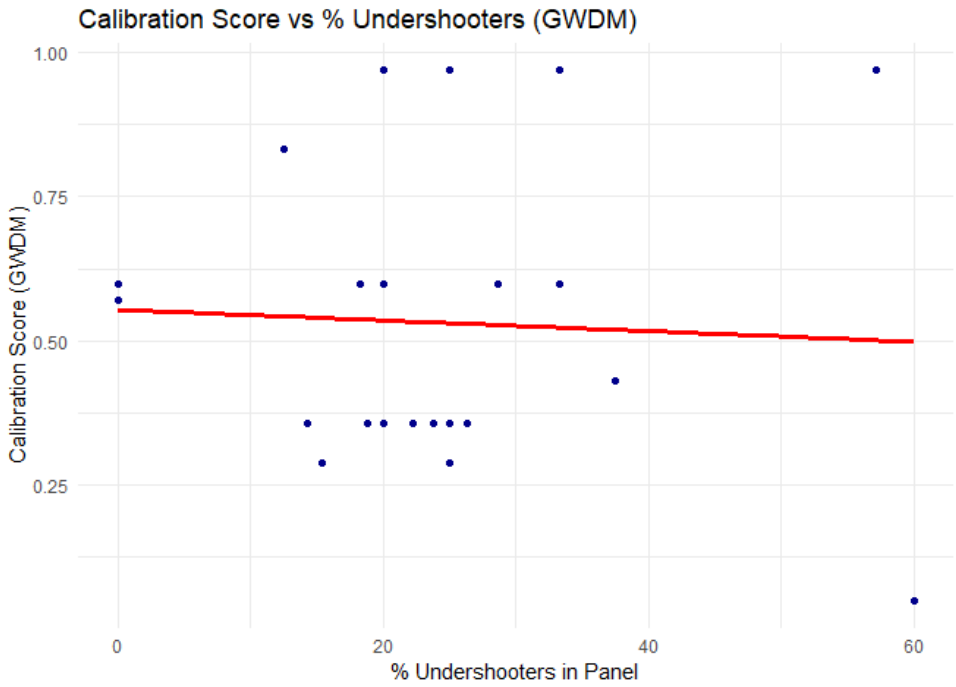


Figure 20: Calibration scores vs proportion of underestimators for GWDM

In Figure 20 and Figure 21, the GWDM and GWDM_opt show a less pronounced decline. The GWDM assigns weights to each expert based on their overall calibration and information scores, while GWDM_opt further optimizes these weights to maximize performance. Since these methods give lower weights to poorly performing experts, like underestimators, their influence on the final distribution becomes smaller. This explains why the calibration score does not drop as much.

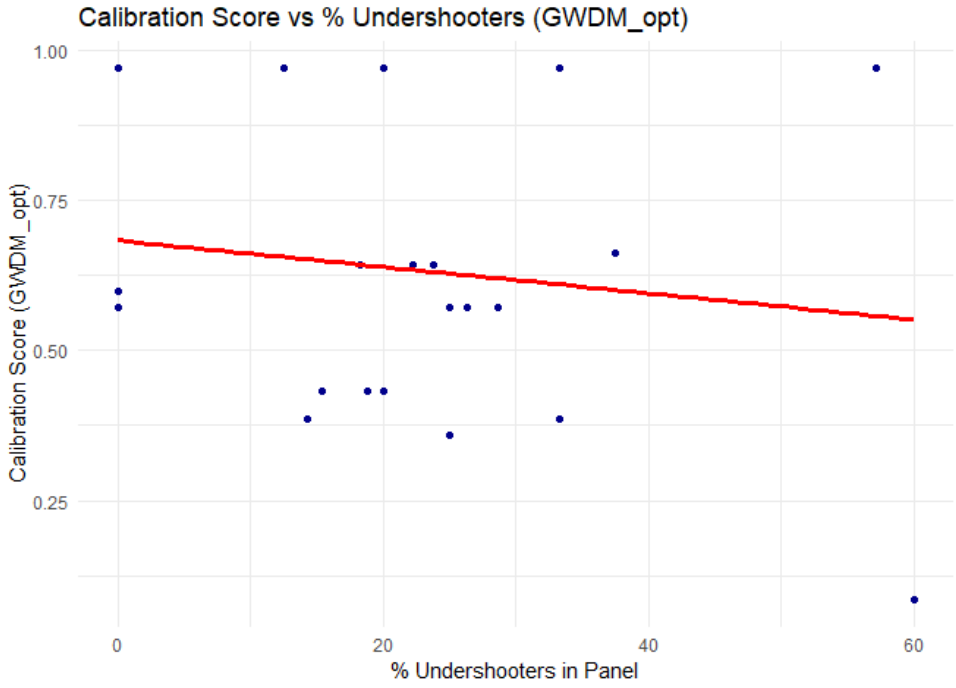


Figure 21: Calibration scores vs proportion of underestimators for GWDM_opt

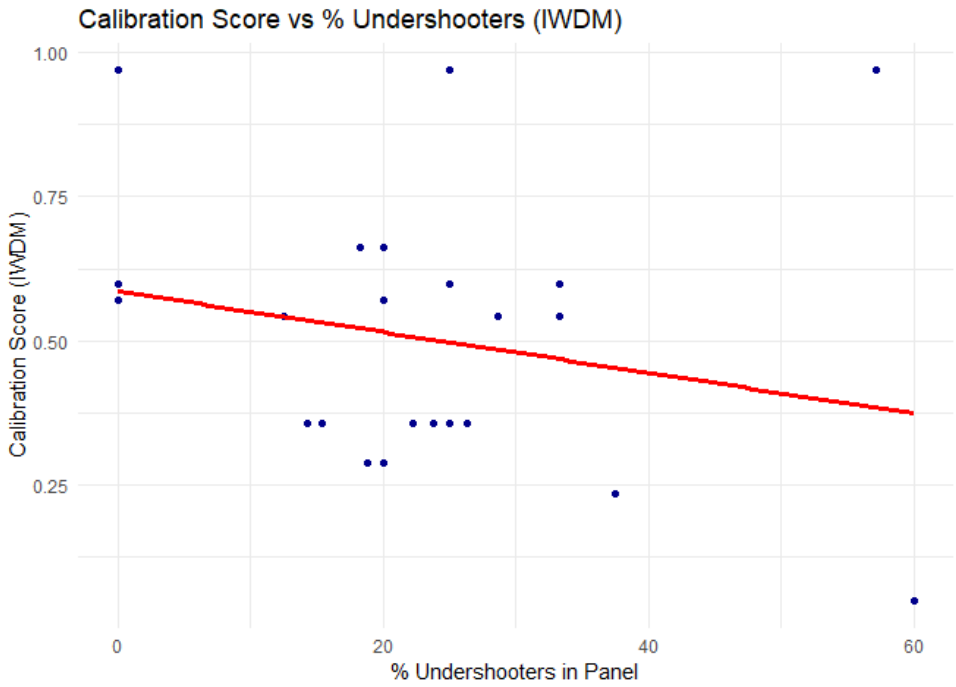


Figure 22: Calibration scores vs proportion of underestimators for IWDM

In Figure 22 and Figure 23, the IWDM and IWDM_opt have a steeper regression line than those for GWDM and GWDM_opt however, this is mainly due to the fact that these methods often achieve higher calibration scores when there are few or no underestimators in the panel. The reason that the decline look sharper is because they start from a better position. Instead of assigning one weight per expert, these methods give weights per question. This allows the decision maker to focus more on strong estimates and ignore weak ones for each specific

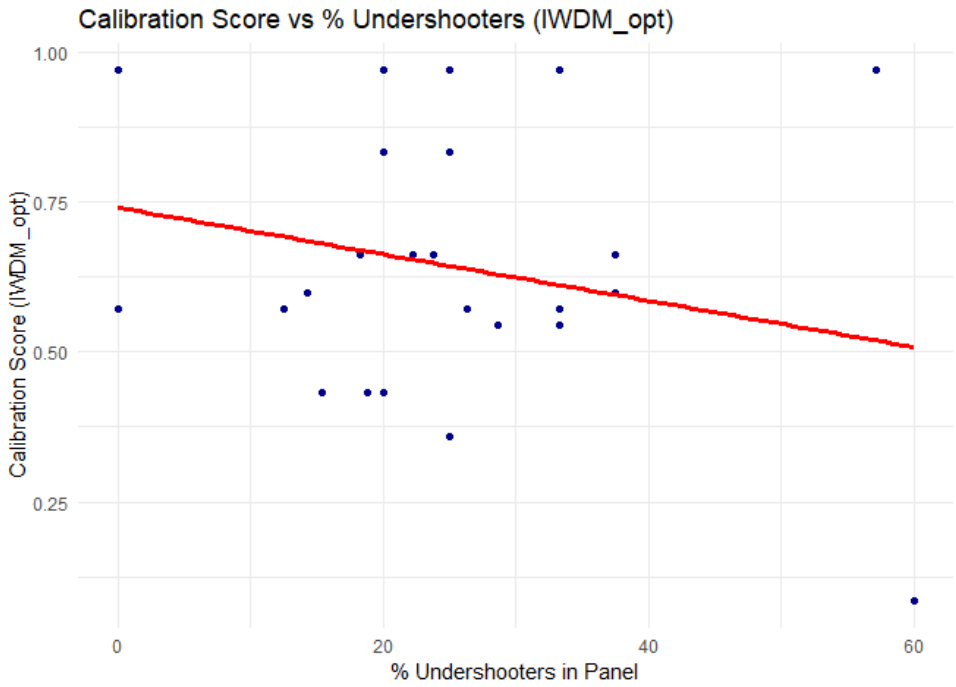


Figure 23: Calibration scores vs proportion of underestimators for IWDM_opt

When looking back at the BRUCL panel in Table 3, the poorest performance is obtained when using the EWDM, which assigns equal weight to all experts. This outcome is expected since four out of five experts in the panel have calibration score below the 0.05 threshold.
For GWDM and IWDM, we see a slight improvement. These methods assign weights to better performing experts. However, the calibration scores for both remain below 0.05, indicating that the presence of multiple poorly calibrated experts still limits overall decision maker performance.
Only with the optimized decision makers (GWDM_opt and IWDM_opt) does the calibration score rise above the threshold and is equal to the calibration score of expert 38, who is the only member of the BRUCL panel with a calibration score above 0.05. The calibration score of expert 38 is equal to 0.0842 which is still not very high. Therefore for all the DMs the BRUCL panel is scoring lower then all other panels. This is partly due to the presence of multiple underestimators, however also the composition of the panel plays an important role.

Expert ID	Perc_Good	Perc_Overestimate	Perc_Underestimate	Difference
1	40	33.3	26.7	1
3	20	33.3	46.7	-2
6	80	6.7	13.3	-1
8	46.7	13.3	40	-4
19	80	13.3	6.7	1
23	66.7	13.3	20	-1
42	46.7	26.7	26.7	0
44	46.7	13.3	40	-4

Table 9: Percentage of good estimations, overestimations, and underestimations of the CRYPT panel

Table 9 presents the CRYPT panel, which consists of eight experts. As shown, this panel includes two underestimators, both of whom have individual calibration scores below 0.05. However, when we look at Table 1, we see that experts 1,3,23, and 42 also have calibration scores below 0.05, even though they are not classified as underestimators. Only experts 6 and 19 have calibration scores above the threshold, both scoring 0.5710.
In the earlier figures, we saw that this panel receives a calibration score below 0.05 under the EWDM method. This outcome is expected, as EWDM assigns equal weights to all experts causing the poor performance of most experts in the panel to dominate the final results. This panel provides a clear illustration of how alternative decision makers can significantly improve performance. Once better-performing experts get more weight, the panel's calibration score improves. We also see in this panel that a low calibration score can also be caused by experts with low individual calibration score who are not classified as underestimators. Again, the panel composition is crucial for the final performance of the different Decision Makers.

Now, we will analyze the LEPTO panel, which is shown in Table 10. This panel consists of seven experts, of whom four are classified as underestimators. Interestingly, despite the high number of underestimators, the panel achieves a very high calibration score. All underestimators in this panel have low individual calibration scores, and expert 43, who is not an underestimator, also scores extremely low in calibration. On the other hand, expert 6 and expert 38 stand out with good calibration scores. As shown in Table 3, the EWDM performs the worst among all decision makers in terms of calibration score. However, despite being the least effective of the five methods, EWDM still has an good calibration score above 0.05. Since this method assigns equal weights to all experts it is notable that the calibration score of the EWDM is good. This suggests that in this panel, the presence of a few well-performing experts is sufficient to compensate for those with poor individual scores when they are equally distributed. That said, the benefits of performance-based weighting become immediately clear when we look at the other decision makers. Once weights are assigned in favor of the better-performing experts, the calibration scores of the DMs improve significantly.

Expert ID	Perc_Good	Perc_Overestimate	Perc_Underestimate	Difference
6	80	6.7	13.3	-1
9	33.3	20	46.7	-4
36	40	20	40	-3
38	66.7	13.3	20	-1
40	46.7	13.3	40	-4
43	26.7	40	33.3	1
44	46.7	13.3	40	-4

Table 10: Percentage of good estimations, overestimations, and underestimations of the LEPTO panel

These panel analyses demonstrate that the presence of underestimators can indeed contribute to lower calibration scores for the decision makers. This effect is most evident in the EWDM, where all experts are given equal weight. Therefore the presence of underestimators who all have low calibration scores can reduce the overall score.
In contrast, the other decision maker methods are less sensitive to the number of underestimators, since better-performing experts are assigned more weight. As a result, their

impact on the aggregates result is more limited.
However, it's important to note that the extent to which underestimators affect the final calibration score depends heavily on the composition of the panel. For instance, in the BRUCL panel, the effect of underestimators is pronouced because only one expert has a calibration score above the 0.05 threshold. This makes it difficult for any of the decision makers to reach a satisfactory overall performance.
In contrast, the LEPTO panel shows that even with several underestimators, a high calibration score is still possible if the panel includes strong experts. In such case, they compensate the weaker performance of the other experts especially when performance-based weighting methods are applied.
Finally, the CRYPT panel illustrates the poor calibration scores can also result from experts who are not classified as underestimators but still perform poorly in terms of calibration score. This underlines that it is not only the number of undershooters that matters, but the overall distribution of expert quality in the panel.
Together, these cases highlight that both the number and the quality of the experts play a key role in shaping the performance of decision makers. Especially in the presence of weak experts, methods that assign weights based on performance are better equipped to maintain high calibration scores.

4 Conclusion

The main questions of this thesis are: How many experts are needed in a panel for reliable performance? And when are there too many experts? To answer these questions, the Classical Model (CM) is used to evaluate each expert based on two scores: the calibration score and the information score. This thesis mainly focused on how the calibration score changes in different panel sizes. Besides this, this thesis also focused on how individual expert behavior, such as a tendency to systematically underestimate, affects both individual calibration scores and calibration scores of the 5 different Decision Makers (DMs).

Based on the analysis of how the calibration score of the different DMs changes for different panel sizes, several conclusions can be made. It was observed that smaller panels tend to show greater variability in their calibration scores. This is likely due to the fact that small groups are more sensitive to the particular experts included. As panel size increases, the calibration scores become more stable. However, the highest calibration scores are not achieved in the large panels. Instead, the optimal performance appears to occur with panels of approximately 6 to 8 experts. Beyond this range, the calibration score tend to decline again. This suggests that adding more experts after 6 to 8 experts does not further improve the quality of the panel.

That said, it's important to note that even larger panels still maintain calibration scores above the threshold of 0.05. Therefore, if we only consider the calibration score, one could argue that there is no such thing as "too many experts". However, in practice, factors such as limited expert availability and the cost of involving additional experts might play a role. Based on the data used in this thesis, a panel of 6 to 8 experts appears to be the best amount of experts in a panel.

The performance of the different Decision Makers were also compared to each other. EWDM, which assigns equal weights to all experts, consistently performs the worst comparing to the different DMs. GWDM and IWDM followed a similar trend, although IWDM performs slightly better. The optimized versions GWDM_opt and IWDM_opt clearly outperformed all the other DMs and also here the IWDM_opt showed a slight advantage over GWDM_opt. These results show the benefits of using performance-based strategies and also show that it's even better to also imply optimization.

In addition, this thesis investigated the role of underestimators and overestimators in explaining certain calibration outcomes. It was evident that both underestimators and overestimators consistently reveice low individual calibration scores.
When examining how the proportion of underestimators in a panel affects the calibration score of the decision makers, a clear trend was observed. The impact is strongest in the case of EWDM, where all experts contribute equally to the final judgment. For the other decision makers, where better performing experts are weighted more heavily, the effect of underestimators is much less pronounced.

Overall, the results show that the presence of underestimators can reduce decision maker performance, however this is strongly dependent on the composition of the panel. In some cases, as seen in the BRUCL panel, the underestimators can significantly lower the calibration score, especially when the panel lacks strong performers. In contrast, other

panels with also a high proportion of underestimators, such as the LEPTO panel, can still achieve high scores when strong experts are present.

5 Discussion

This thesis aimed to identify an optimal panel size for expert judgment using the Classical Model (CM), and to understand how individual expert behavior, particularly systematic underestimation, affects both individual and group performance in terms of the calibration score. Based on the analysis of two datasets from the RIVM, the results suggest that panels consisting of 6 to 8 experts gives the best calibration scores. Beyond this size, while calibration scores remain acceptable, there is no clear performance gain from adding more experts, and in some cases, a slight decline is observed. This implies that according to this dataset, adding more experts after a certain point may not improve the overall performance.

However, these conclusion should be interpreted with caution. They are based on a single elicitation involving 43 experts and 15 seed questions. Therefore, the results may not generalize across all expert judgment scenarios. The following factors could influence the overall outcomes.

First, the specific set of questions used in the elicitation strongly shapes the results. If a different set of 15 seed questions were used the performance of experts and the resulting panel scores could look very different. Some experts may perform well in one topic area but poorly in another. Similarly, expert performance may vary simply due to the conditions on the day of elicitation, such as fatigue or misinterpretation of instructions.

Second, the number of seed questions plays also a role in the reliability of the calibration score. Our elicitation used 15 seed questions and I think this is a reasonable amount of questions. However, with more questions it could stabilize the scores and offer a more nuanced view of expert performance.

In summary, while this thesis provided valuable insight into how panel sizes and composition affect the performance of decision makers, the results applies only in the context of this specific experimental setup. So with 43 experts, 15 seed questions, and with the kind of seed questions. Further research with varied datasets is needed to determine whether the observed trend hold more generally.

Bibliography

References

[1] Roger M. Cooke. "Validating Expert Judgment with the Classical Model". In: *Experts and Consensus in Social Science*. Ed. by Carlo Martini and Marcel Boumans. Springer, 2013, pp. 1–20.

[2] Roger M. Cooke and Louis L.H.J. Goossens. "TU Delft expert judgment data base". In: *Reliability Engineering and System Safety* 93.5 (2008), pp. 657–674. DOI: 10.1016/j.ress.2007.03.005.

[3] A.M. Hanea and G.F. Nane. "An in-depth perspective on the Classical Model". In: *Elicitation: The Science and Art of Structuring Judgment*. Cham: Springer, 2022.

[4] Arie H. Havelaar et al. "Attribution of Foodborne Pathogens Using Structured Expert Elicitation". In: *Foodborne Pathogens and Disease* 5.5 (2008), pp. 649–660. DOI: 10.1089/fpd.2008.0115.

[5] OpenAI. *ChatGPT used for grammar and spelling corrections in this thesis (June 2025 version)*. https://chat.openai.com. 2025.