Author/s:
Dharmarathne, Hetti Arachchige Sameera Gayan

Title:
Exploring the statistical aspects of expert elicited experiments

Date:
2020

Persistent Link:
https://hdl.handle.net/11343/238547

Exploring the Statistical Aspects of Expert Elicited Experiments

Hetti Arachchige Sameera Gayan Dharmarathne

The School of Mathematics and Statistics
The University of Melbourne

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

January 2020

Abstract

We explore the statistical aspects of some of the known methods of analysing experts' elicited data to identify potential improvements on the accuracy of their outcomes in this study. It can be identified that potential correlation structures induced in the probability predictions by the characteristics of experimental designs are ignored in computing experts' Brier scores. We show that the accuracy of the standard error estimates of experts' Brier scores can be improved by incorporating the within-question correlations of probability predictions in the second chapter of this thesis. Missing probability predictions of events can impact on assessing the prediction accuracy of experts using different sets of events (Merkle et al., 2016; Hanea et al., 2018). It is shown in the third chapter that multiple imputation method using a mixed-effects model with questions' effects as random effects can effectively estimate missing predictions to enhance the comparability of experts' Brier scores.

Testing experts' calibration on eliciting credible intervals of unknown quantities using hit rates; observed proportions of elicited intervals that contain realized values of given quantities (McBride, Fidler, and Burgman, 2012), has a property of obtaining lower values of power to correctly identify well-calibrated experts and more importantly, the power tends to decrease as the number of elicited intervals increases. The equivalence test of a single binomial proportion can be used to overcome these problems as shown in the fourth chapter. There is a possibility of allocating higher weights to some of the not well-calibrated experts by the way experts' calibration is assessed in the Cooke's classical model (Cooke, 1991) to derive experts' weights. We show that the multinomial equivalence test can be used to overcome this problem in the fifth chapter.

Experts' weights that derived from experiments to combine experts' elicited subjective probability distributions to obtain aggregated probability distributions of unknown quantities (O'Hagan, 2019) are random variables subject to uncertainty. We derive shrinkage experts' weights with reduced mean squared errors in the sixth chapter to enhance the precision of the resulting aggregated distributions of quantities.

Declaration

This is to certify that:

1. the thesis comprises only my original work towards the PhD;
2. due acknowledgement has been made in the text to all other material used; and
3. the thesis is fewer than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Hetti Arachchige Sameera Gayan Dharmarathne

Acknowledgements

It is a great pleasure to acknowledge my deepest thank and gratitude to my supervisor Associate Professor Andrew Robinson, my co-supervisor Dr. Anca Hanea and the chairperson of the Confirmation and Supervisory Panel Professor. David Balding for their great guidance and support to make this thesis a success. I also thank the University of Melbourne to provide me this opportunity to pursue a PhD with financial assistance.

I am really grateful to all the staff and friends at the School of Mathematics and Statistics and the Centre of Excellence for Biosecurity Risk Analysis (CEBRA) at the University of Melbourne who have been friendly and helpful throughout this period. Finally, I sincerely thank my wife and our family members for their encouragement and help to complete this thesis.

Contents

List of Figures

List of Tables

1. Introduction

Expert elicitation refers to employing formal procedures for obtaining and combining expert judgments, when existing data and models cannot provide required information for decision making in practice (Colson and Cooke, 2018). Furthermore, educated guesses from experts could play an important role in decision making of new and emerging contexts of which no data are available. The existing literature of applications in widespread areas (O'Hagan, 2019) speaks to itself about the importance of expert elicitation in practice. Therefore, we consider important to research on improving the accuracy of the outcomes of analysing experts' elicited data. Thus, we explore the statistical aspects of some of the known methods of analysing experts' elicited data to identify potential improvements on the accuracy of their outcomes in this study.

The Brier score is one of the commonly used scoring rules to assess the prediction accuracy of experts in forecasting probabilities of the occurrence of events (O'Hagan et al., 2006; McBride, 2013; Hanea et al., 2017). Experts' Brier scores are computed in situations of which the experts are predicting the probabilities of the occurrence of events of multiple questions. Therefore, the probability predictions form a two-level non-nested hierarchical data structure with questions at level 1, within experts at level 2. This hierarchical data structure can be illustrated as in figure 1.1 below. Furthermore, there can be instances where the predictions are made either with partners or groups leading to multi-level hierarchical data structures (Stockard, Peters, et al., 2007). Observations between layers in a hierarchical data structure can be considered independent but dependent within layers as they belong to the same subpopulation (Demidenko, 2013). Therefore, there can be correlation structures induced in the observations due to sharing the effects of same layers of an experimental design. It can be identified that typical computation of experts' Brier scores ignores

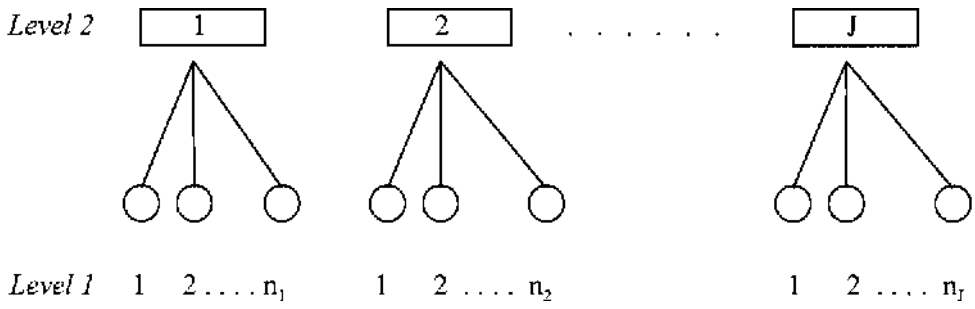


FIGURE 1.1: Illustration of a two-level hierarchical data structure due to Sullivan, Dukes, and Losina (1999)

the potential correlation structures that are induced in the probability predictions by the characteristics of experimental designs.

The accuracy of the standard error estimates of parameters can be affected by the ignored correlations between observations (Finch, Bolin, and Kelley, 2016). According to Jiang (2007), the random effects in mixed-effects models are used to model the variation due to correlated observations in hierarchical data structures. Therefore, we compute experts' Brier scores using a mixed-effects model with questions' effects as random effects to study the impact of incorporating the potential correlations between probability predictions due to the effects of common questions on the estimated standard errors of experts' Brier scores in the second chapter of this thesis. It is evident from the analysis that the accuracy of the standard error estimates of experts' Brier scores can be improved by incorporating the within-question correlations of probability predictions.

It is important to note that human judges (even experts) may prefer to assess only the subsets of events of which they feel comfortable to offer coherent predictions in practice (Predd et al., 2008). Missing probability predictions of events are typically ignored in computing experts' Brier scores to assess the prediction accuracy of experts in practice. If experts' Brier scores are computed using the probability predictions of different subsets of events, then the comparison of the prediction accuracy of experts using Brier scores can be challenging and perhaps may be less meaningful (Merkle et al., 2016; Hanea et al., 2018). Therefore, it is important to enhance the comparability of experts' Brier scores by adjusting for the missing probability predictions of events. Hence, we apply some missing value estimation methods to estimate missing probability predictions of events in computing experts' Brier scores in the third chapter of this

thesis. The results of a conducted simulation study show that multiple imputation method using a mixed-effects model with questions' effects as random effects can estimate missing probability predictions to compute experts' Brier scores with reduced errors compared to the typically computed experts' Brier scores that ignore missing predictions.

Experts' calibration on eliciting credible intervals of unknown quantities is tested by the direct comparison of experts' hit rates; observed proportions of experts' elicited intervals that contain realized values of given quantities (McBride, Fidler, and Burgman, 2012), with the level of intended coverage probability of elicited intervals (Speirs-Bridge et al., 2010). Here, the hit rates are considered as fixed quantities disregarding their random variation in elicitation contexts. We emphasize the importance of considering the random variation of hit rates in testing experts' calibration as computed experts' hit rates can randomly vary from their true levels of coverage in short term and therefore the long-term average hit rates will better reflect the true levels of coverage of experts' elicited intervals.

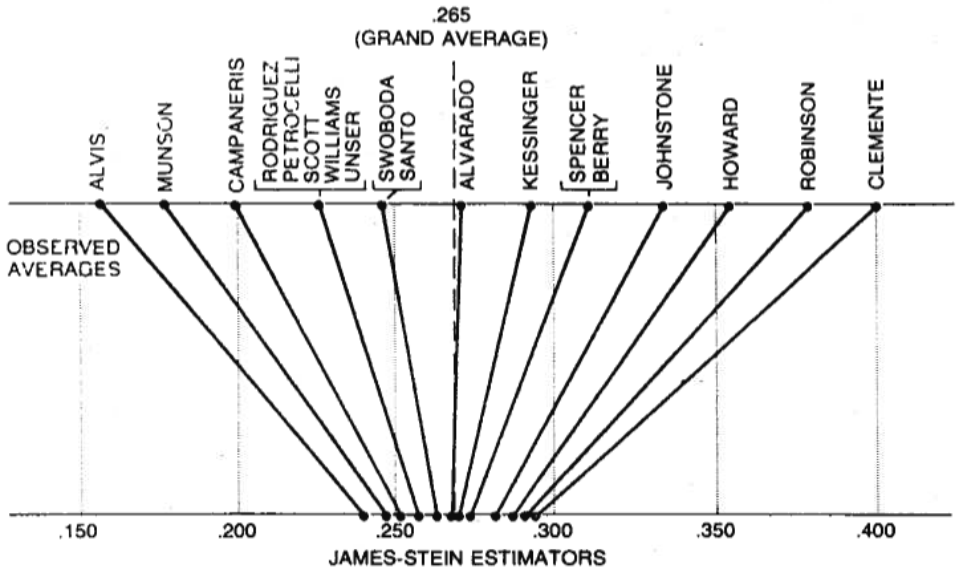
The results of a conducted simulation study at some selected standard levels of intended coverage probabilities show that the test has a property of obtaining lower values of power to correctly identify well-calibrated experts and more importantly, the power tends to decrease as the number of elicited intervals increases. We consider the hit rates as random variables and apply the equivalence test of a single binomial proportion to compare the population means of experts' hit rates with the level of intended coverage probability of elicited intervals to test experts' calibration statistically. The above identified problems of the direct comparison of hit rates can be solved using the equivalence test of a single binomial proportion for large number of elicited intervals as shown in the fourth chapter of this thesis.

Experts' weights are derived to combine experts' elicited subjective probability distributions to obtain aggregated probability distributions of unknown quantities (O'Hagan, 2019). Assessing the levels of experts' calibration on eliciting a specified number of percentiles from the probability distributions of quantities of some given seed questions (with known realized values of quantities to the researchers but not to the experts) using the calibration score component is a part of deriving experts' weights using

the Cooke's classical model (Cooke, 1991). The results of a conducted simulation study show that the calibration score component fails to detect not well-calibrated experts with adequately higher values of power even for reasonably large number of elicited quantities. Furthermore, average calibration scores are reasonably high when the calibration score component fails to detect not well-calibrated experts. Therefore, there is a possibility to allocate higher weights to some of the not well-calibrated experts if they have also obtained higher information scores. We show that the multinomial equivalence test can be used to identify not well-calibrated experts with higher accuracy and overcome the potential issue of allocating higher weights to not-well calibrated experts in the fifth chapter of this thesis.

It can be identified following the discussion on deriving experts' weights in Cooke's classical model (Cooke, 1991, chap. 12) that computed weights are considered as fixed quantities. We emphasizes the importance of considering the weights as random variables subject to uncertainty as they are derived from experiments. Therefore, it is important to address this underlying uncertainty of the derived experts' weights from experiments in computing aggregated distributions of quantities. James-Stein shrinkage estimation technique discussed in James and Stein (1961) can be used to estimate the mean of a multivariate normal distribution with reduced mean squared errors. This concept of reducing the mean squared errors of estimates can be illustrated using the well-known example of computing James-Stein estimators for the 18 baseball players by shrinking their individual batting averages toward the overall average of averages as in figure 1.2 below. Therefore, we explore the potential of applying the James-stein technique to obtain experts' weights with reduced mean squared errors in the sixth chapter of this thesis.

We apply an empirical Bayes development of the James-Stein shrinkage estimation technique discussed in Zhao (2010) that shrinks variables differently depending on their variances (larger the variance more shrinkage there should be) on Cooke's weights to derive shrinkage weights with reduced mean squared errors in this analysis. The results of the analysis show that overall Decision Maker (DM) calibration and information scores of some selected testing questions with known realized values are different between the normalized typical and shrinkage Cooke's weights. Considering the



JAMES-STEIN ESTIMATORS for the 18 baseball players were calculated by "shrinking" the individual batting averages toward the overall "average of the averages." In this case the grand average is .265 and each of the averages is shrunk about 80 percent of the distance to this value. Thus the theorem on which Stein's method is based asserts that the true batting abilities are more tightly clustered than the preliminary batting averages would seem to suggest they are.

FIGURE 1.2: Illustration of James-Stein shrinkage estimation due to Efron and Morris (1977)

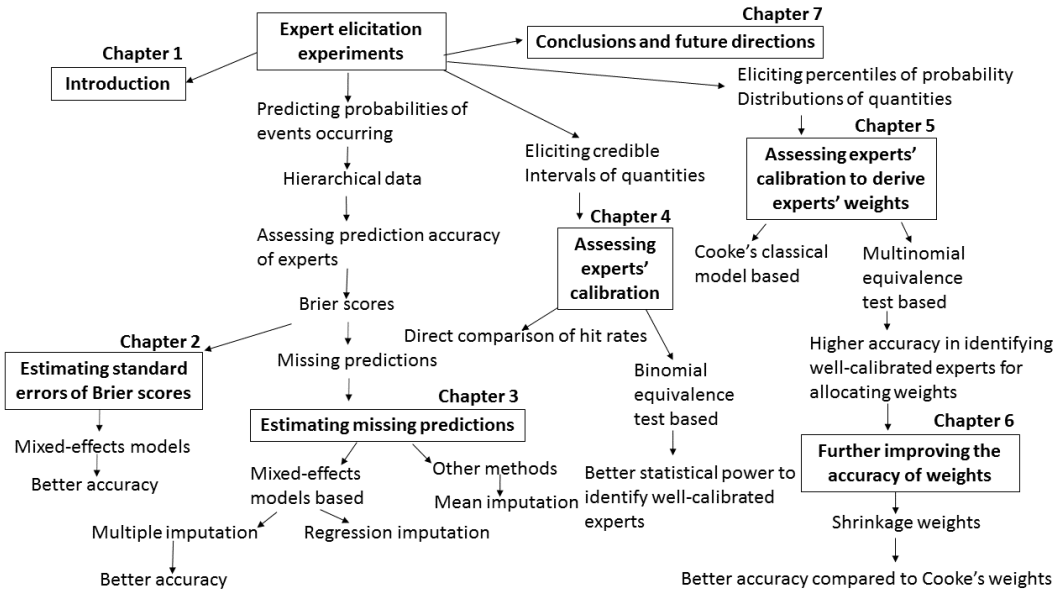


FIGURE 1.3: Mind map of the flow of the thesis

purpose of applying the shrinkage estimation technique to reduce the mean squared errors of estimators of the mean of a multivariate normal distribution due to James and Stein (1961) and the procedure of the employed empirical Bayes shrinkage approach to derive weights by reducing the allocated weights for experts with larger variances of weights in Zhao (2010), we suggest that the outcome of the shrinkage weights can expect to be more accurate than the typically computed weights. Overall, the flow of the thesis can be illustrated using the mind map given in figure 1.3 above.

2. Aligning the analysis and the design of expert elicitation experiments

Section 2.1: Introduction

This chapter focuses on methods for improving the estimation of standard errors of experts' Brier scores that are derived from expert elicited experiments to assess the prediction accuracy of experts. A brief summary on computing experts' Brier scores as follows. Experts' Brier scores are typically computed by obtaining average squared deviations between the predicted probabilities of the occurrence of events and their outcomes within experts. This aggregated analysis considering an expert as the unit of analysis ignores potential correlation structures that are induced in the probability predictions by the characteristics of experimental designs - for example, asking common questions from experts. If we consider experts' probability predictions on multiple questions in an experiment, it is reasonable to assume that the probability predictions on similar questions can be correlated due to common questions' effects.

The independent assumption of errors will be violated if we apply standard statistical methods ignoring potential correlations between observations. It will result in obtaining inaccurate standard error estimates of model parameters (Finch, Bolin, and Kelley, 2016). Mixed-effects models can incorporate potential design-based correlation structures of observations due to the groupings of data by associating common random effects to observations that share the same levels of grouping factors (Pinheiro and Bates, 2000). Therefore, we compute experts' Brier scores using a mixed-effects model that includes questions' effects as random effects to study the impact of incorporating the above discussed within-question correlations of probability predictions on the

standard error estimates of experts' Brier scores in this analysis. It is evident from the analysis that the accuracy of the standard error estimates of experts' Brier scores can be improved by incorporating the within-question correlations of probability predictions. We begin the analysis with the following review about the scoring rules that are relevant to the context of using Brier scores with some evidence suggesting the common use of Brier scores in practice to support our interest to work on improving the computation of Brier scores in this analysis.

Section 2.2: Scoring rules for probability predictions of events with multiple outcomes

A scoring rule can be considered as a function that measures the degree of association between expert's elicited probability distributions and the observed true values of some unknown quantities (O'Hagan et al., 2006). Therefore, a scoring rule can be used to compare the prediction accuracy of experts. According to O'Hagan et al. (2006), a scoring rule should motivate the experts to record their opinions well and train them to quantify their opinions accurately. Therefore, it is important that a scoring rule should encourage experts to be honest when providing their opinions. A scoring rule with this property is termed "proper". Here, we review the Brier, Logarithmic, and Spherical scores as they are the proper scoring rules relevant to our context of assessing the prediction accuracy of experts on probability predictions of the occurrence of events with multiple outcomes. Following details about these scoring rules are due to O'Hagan et al. (2006).

Suppose an event E takes exactly one of g outcomes, $O_1, O_2, ..., O_g$. Let p_i be the probability that the outcome O_i occurs. Hence, each p_i must be non-negative and $\sum_{i=1}^{g} p_i = 1$. Let $d_i = 1$ if O_i occurs and 0 otherwise. Thus, exactly one of $d_1, d_2, ..., d_g$ will be non-zero. The Brier score (BS), Logarithmic score (LS), and Spherical score (SS) can be defined as

$$\text{BS} = \sum_{i=1}^{g} (p_i - d_i)^2, \tag{2.1}$$

$$\text{LS} = \ln\Big(\sum_{i=1}^{g} p_i d_i\Big), \text{ and} \tag{2.2}$$

$$\text{SS} = \sum_{i=1}^{g} p_i d_i \Big/ \Big(\sum_{i=1}^{g} p_i^2\Big)^{1/2}. \tag{2.3}$$

The above scores can be defined as mean or average scores when probability predictions on multiple events are available (refer equation 2.4 for the average Brier score).

We are interested in identifying the commonly applied scoring rule among these three scoring rules. O'Hagan et al. (2006), McBride (2013), and Hanea et al. (2017) mentioned that the Brier score is commonly applied in practice. Furthermore, O'Hagan et al. (2006) reported a recommendation from meteorologists; who are the group of scientists that makes the most use of scoring rules, that Brier score is the most preferred scoring rule for judgements related to probability predictions. Thus, it is reasonable to assume that the Brier score is commonly applied than the Logarithmic and Spherical scores in practice. Therefore, we consider important to work on improving the computation of Brier score in this analysis.

Now consider the Brier score (Brier, 1950) that was originally defined as an average measure of accuracy over a set of m probability predictions of the occurrence of events with g outcomes.

$$BS = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{g}(p_{ji} - d_{ji})^2, \tag{2.4}$$

where p_{ji} is the probability prediction of the occurrence of the i^{th} outcome of the j^{th} event and d_{ji} takes 1 if the i^{th} outcome of the j^{th} event occurs and 0 otherwise. Following definition is also used when all the events are binary (Candille and Talagrand, 2005).

$$BS = \frac{1}{m}\sum_{j=1}^{m}(p_j - d_j)^2, \tag{2.5}$$

where p_j is the predicted probability of occurrence of the j^{th} event and d_j is equal to 1 or 0 depending on whether j^{th} event has occurred or not.

Section 2.3: Background of the research interest

Consider an experiment in which expert elicitation gives answers to multiple questions on predicting the probabilities of the occurrence of events. It follows from the discussion in section 2.2 that experts' Brier scores are commonly used to assess the prediction accuracy of experts in this context and they are typically computed by obtaining average squared deviations between the predicted probabilities of the occurrence of events and their outcomes within experts. This aggregated analysis considering an expert as the unit of analysis fails to capture important variation that remains within the data (Stockard, Peters, et al., 2007). If we consider how the standard error estimates are typically computed to assess the uncertainty of experts' Brier scores, it can be identified that potential correlation structure induced in the probability predictions by asking common questions is ignored in the analysis.

Suppose that the standard errors of experts' Brier scores are estimated in an experiment by asking 20 common questions from each of 10 experts. The traditional approach is to pretend that there are 200 questions, 20 for each expert, ignoring the potential correlation structure induced in the probability predictions by asking common questions in the experiment. This underlying within-question correlation of probability predictions can impact on the accuracy of the standard error estimates of experts' Brier scores as shown below. Therefore, we focus on improving the accuracy of the standard error estimates of experts' Brier scores by incorporating the correlations between probability predictions due to the effects of common questions in this analysis.

The above discussed issue of the impact of correlated observations on the accuracy of the standard error estimates has been discussed under the "design effect" of cluster sampling by Noortgate, Opdenakker, and Onghena (2005). The authors pointed out that the design effect can be large if the units belong to a same group are highly similar as indicated by a large intra-class correlation. Furthermore, they mentioned that standard error estimates can be substantially distorted even for a relatively small intra-class correlation. In relevance to this, Hox (2002) discussed a correction procedure for the standard error estimates through the computation of effective sample

sizes in two-stage cluster sampling as

$$\tilde{n}_{eff} = \tilde{n}/[1 + (\tilde{n}_{clus} - 1)\rho], \tag{2.6}$$

where $\tilde{n}_{eff}$ is the effective sample size, $\tilde{n}$ is the total sample size, $\tilde{n}_{clus}$ is the cluster size, and ρ indicates the intra-class correlation.

For example, let us assume that the within-question correlation (ρ) takes values 0, 0.5, and 1 in the above discussed context of estimating the standard errors of experts' Brier scores by asking 20 common questions from each of 10 experts in an experiment. Then, the effective sample sizes ($\tilde{n}_{eff}$) to be used for computing the standard error estimates of experts' Brier scores can be calculated as 200, 19, and 10, respectively for the assumed ρ values using the equation 2.6 above. Here, the total sample size ($\tilde{n}$) is 200 and the cluster size ($\tilde{n}_{clus}$) is 20. This demonstrates the potential impact of ignoring the underlying within-question correlation of probability predictions on the accuracy of the standard error estimates of experts' Brier scores through applying incorrect effective sample sizes.

The formula in equation 2.6 assumes equal number of cluster sizes, which is not always realistic. Hox (2002) explains that the effective sample size can be considerably reduced and the standard error estimates can be substantially increased even for a small intra-class correlation. Therefore, it is important to obtain corrected standard error estimates using effective sample sizes in cluster sampling. However, as Hox (2002) pointed out, applying standard error correction procedures can be difficult in general multilevel data with different design effects. Therefore, a more flexible approach is to use a suitable "multilevel model" in which different kinds of dependencies can be accounted through the model. We propose to do this here.

Multilevel models are also known as hierarchical models and they are fitted to data arising from multilevel or hierarchical data structures (Gelman and Hill, 2006). Multilevel or hierarchical data structures consist of multiple units of analysis that are ordered hierarchically, and they exist in general when some units of analysis can be considered as a subset of other units in a hierarchy (Steenbergen and Jones, 2002). The observations between levels in a hierarchical structure are considered independent

but dependent within levels as they belong to the same subpopulation (Demidenko, 2013). If we consider the above discussed experiment of which several experts are making probability predictions on multiple questions, it represents a hierarchical data structure with two levels for the experts and questions. Therefore, the probability predictions can be correlated due to the common grouping effects of experts and questions.

It was mentioned above that the typical procedure of computing experts' Brier scores ignores the potential correlation structure induced in the probability predictions due to the grouping effect of common questions. According to Pinheiro and Bates (2000, chap. 1), mixed-effects models can represent potential covariance structures (equivalently, correlation structures) induced due to the groupings of data by associating common random effects to observations that share the same levels of grouping factors. Random effects impose observations that share the same levels of grouping to have the same intercept and/or slope (Harrison et al., 2018). Therefore, we expect to compute experts' Brier scores using a mixed-effects model that includes questions' effects as random effects in this analysis. Hence, the estimated standard errors of experts' Brier scores from the fitted mixed-effects models can be expected to be more accurate than the typically computed standard error estimates ignoring potential correlated probability predictions due to the effects of common questions.

Section 2.4: Related literature

The definitions of Brier scores in equations 2.4 and 2.5 show that probability predictions of all questions have equal weights in computing experts' Brier scores. Evaluating the prediction accuracy of experts using equally weighted Brier scores may be suboptimal, unless all the experts have predicted all the questions (Merkle et al., 2016). In relevance to this, Hanea et al. (2018) mentioned that comparison of experts' Brier scores computed from different sets of questions can be challenging and comparisons will be more meaningful on a same set of questions. We consider this preferred complete set of predictions (all the experts have predicted all the questions) in this chapter to show that the accuracy of standard error estimates of experts' Brier scores

can still be improved even we use a complete set of predictions by incorporating the correlations between probability predictions due to the effects of common questions into the analysis. We also consider the fact that human judges (even experts) may not be able to provide coherent probability predictions for all the events of interest in forecasting contexts. Therefore, they may prefer to assess only the subset of events about which they feel comfortable to offer coherent predictions in practice (Predd et al., 2008). It causes to arise the need of computing experts' Brier scores and their standard errors from incomplete sets of predictions with missing values that will be discussed in the next chapter.

We consider the importance of computing experts' Brier scores not only as point estimates but also along with a suitable measure of uncertainty. It will provide the opportunity to make more realistic comparisons between experts' Brier scores considering their random variation than just comparing the point estimates alone. Note that confidence interval is one of the commonly used measures of uncertainty in practice (Speirs-Bridge et al., 2010). We found the use of confidence intervals in Wintle et al. (2012) and Hanea et al. (2017) to perform more realistic comparisons of groups' average Brier scores in predicting the probabilities of the occurrence of some global events on geopolitical, economic, and military sectors before and after group discussions. However, the accuracy of the estimated standard errors of Brier scores of their analyses can be improved by incorporating the potential correlation structures induced in the probability predictions by the common grouping effects of experts and questions.

We discuss a mixed-effects model based method to compute experts' Brier scores as fixed effects and to estimate their standard errors by incorporating potential correlations between probability predictions due to the effects of common questions using random effects. Our method can be extended to compute groups' Brier scores as fixed effects and to estimate their standard errors by incorporating potential correlations between probability predictions due to the common grouping effects of experts and questions using two random effects for the experts and questions.

Applications of mixed-effects models are not very common in judgement and decision making context. Stockard, Peters, et al. (2007) reanalysed data from two of the

previously published articles; modified Iowa gambling task (Peters and Slovic, 2000) and a prisoner's dilemma game (Mulford et al., 1998) to show that bringing in within-subject variation over performing aggregated analyses considering between-subject variation can enhance the analyses to examine the impact of the characteristics of participants on their decisions across the rounds of play in which they engage. Budescu and Johnson (2011) discussed a logistic mixed-effects model based approach to estimate probabilities of accurately throwing basketballs by participants under different experimental conditions to assess the calibration of probability judgements made by participants on their confidence of throwing basketballs accurately. If the probabilities are estimated using relative frequencies computed by aggregating data over observations, important characteristics of experiments and participants will be ignored and unstable probability estimates will be obtained. Budescu and Johnson (2011) used fixed effects to address different experimental conditions and random effects to address within-participant variation in their fitted models. We have not found applications of linear mixed-effects models to compute experts' Brier scores and their standard error estimates. Therefore, it can be considered that our suggested approach of computing experts' Brier scores and their standard error estimates using a mixed-effects model as an extension to the typical computation of experts' Brier scores.

In general, mixed-effects models are applied in situations of which data consist of grouped observations or clusters, where the observations within the same group are not independent. This within-group associations can be accounted by adding a group-specific random effect to a regression model. This type of applications can be found in variety of fields. We found some references in different fields as follows. Speelman, Heylen, and Geeraerts (2018) and Meteyard and Davies (2020) discussed applications in linguistics and psychological studies. Harrison et al. (2018) and Bolker et al. (2009) discussed applications of linear and generalized linear mixed-effects models in ecological studies. In addition to that, applications of mixed-effects models are also found in longitudinal and repeated measures analyses. Some applications in longitudinal and repeated measures analyses have been found in the field of medical studies in Yarkiner et al. (2013) and Andersen and Millen (2013), respectively.

Now we review the mixed-effects models and their properties relevant to the context of improving the accuracy of standard error estimates of fixed effect parameters using random effects. Following information about the mixed-effects models are due to Robinson and Hamann (2010). Mixed-effects models contain both fixed and random effects. The fixed effects are generally assumed to be purposively selected and the estimates of the levels represent nothing other than themselves. On the other hand, the random effects are assumed to be randomly sampled from a population of possible levels. In general, even though it is not always true, the random effects are suggested by the experimental design of a study. For example, the random effects may represent the experimental material such as blocks, plots, subplots and so on. Therefore, in order to reflect the experimental design, necessary random effects should be included in a given statistical model. It is also important to note that random effects are useful to organize the unexplained error variation and improve the diagnostic compatibility of a statistical model. The fixed effects are usually based on the hypotheses of interest of a study. They explain variation and the model has no practical meaning if they are not in the model. Therefore, a mixed-effects model should include all the fixed effects that are necessary for the hypotheses of interest and all the random effects that are necessary to reflect the experimental design of a study.

The mathematical model of the general linear mixed-effects model can be given in matrix form as follows (Littell et al., 1996).

$$Y = X\beta + Zu + e, \tag{2.7}$$

where

Y is the data vector,

β is the vector of parameters of the fixed effects,

X is the design matrix for the fixed effects,

u is the vector of coefficients corresponding to the random effects,

Z is the design matrix for the random effects,

e is the vector of random errors.

It is usually assumed that u and e are uncorrelated normally distributed random

variables with zero means and their covariance matrices G and R respectively. This information can be given as follows:

$$E\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{2.8}$$

$$Var\begin{bmatrix} u \\ e \end{bmatrix} = \sigma^2 \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}. \tag{2.9}$$

Let V denote the variance of Y. Then, V can be shown equal to $ZGZ^T + R$. This is the general specification of the mixed-effects model. The simple random effects is a special case of the general specification, where G is assumed to contain variance components in a diagonal structure (random effects are assumed to be uncorrelated) and $R = \sigma^2 I_{\tilde{n}}$, where $I_{\tilde{n}}$ denotes the $\tilde{n} \times \tilde{n}$ identity matrix of which the size depends on the total sample size $\tilde{n}$ (Littell et al., 1996).

If we consider the estimation of parameters of a mixed-effects model, it is a well-known fact that the main interests of fitting a mixed-effects model are to estimate the fixed effects parameters and the variance components (variation in a dependent variable that is associated with random-effects) of random effects. Robinson and Hamann (2010) mentioned that the maximum likelihood estimates of the covariance parameters of a mixed-effects model are usually biased. Demidenko (2004) suggested that (as described by Robinson and Hamann (2010)) the Restricted Maximum Likelihood (REML) estimates are expected to be less biased than the maximum likelihood estimates. Thus, Robinson and Hamann (2010) mentioned that the variance estimates based on REML are preferred for mixed-effects models. Venables and Ripley (2002) also suggested REML as the default estimation method for mixed-effects models. It is important to note that estimating random effects may be required at times for some prediction purposes. Robinson (1991) discussed in detail about the BLUP (Best Linear Unbiased Prediction) as a technique for estimating random effects.

The estimates of fixed effects ($\hat{\beta}$) and random effects ($\hat{u}$) of the general linear mixed-effects model in equation 2.7 can be given as follows:

$$\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} y \tag{2.10}$$

$$\hat{u} = \hat{G} Z^T \hat{V}^{-1} (y - X\hat{\beta}), \tag{2.11}$$

where $\hat{\beta}$ is the generalized least squares estimate of β and $\hat{u}$ is the BLUP of u (Littell et al., 1996).

If we consider the impact of ignoring the necessary random effects on the standard error estimates of fixed effects parameters, Demidenko (2013) discussed that if the necessary random effects are ignored and the fixed effects parameters are estimated using the Ordinary Least Squares (OLS) method, then the standard error estimates of fixed effects parameters will be inflated and they could be declared as statistically insignificant. Introduction of necessary random effects will reduce the standard error estimates of fixed effects parameters. Demidenko (2013) also mentioned that if a single random effect is ignored, then the error variance (σ^2) is roughly overestimated by the variance component of the corresponding random effect. Therefore, we consider important to incorporate variance component of questions in estimating standard error estimates of experts' Brier scores in this analysis.

Section 2.5: Methodology of computing Brier scores

The original definition in equation 2.4 implies that the Brier scores can be computed for events with multiple outcomes if the probabilities of the occurrence of each outcome are predicted. We restrict our attention to the events with two outcomes; either occurrence or the non-occurrence of an event, in this study. Therefore, we derive Brier scores that are similar to equation 2.5. Let p_{ij} be the probability of the j^{th} event occurrence predicted by the i^{th} expert and d_j represent the outcome of the j^{th} event; where 1 indicates the "occurrence" and 0 indicates the "non-occurrence". Further, define $Y_{ij} = (p_{ij} - d_j)^2$ as the squared deviation between p_{ij} and d_j. Observe that

Y_{ij}'s are grouped in a one-way classification because they are classified according to a single characteristic- the expert on which the predictions are made. According to Pinheiro and Bates (2000, chap. 1), data from a one-way classification like this can be analysed either with a fixed-effect model or a random-effects model. We are interested to estimate the Brier scores of those particular levels of experts that are used in the experiment but not to make inferences about the population from which the experts are drawn. Therefore, following the explanation in Pinheiro and Bates (2000, chap. 1), we fit the following linear fixed-effects model.

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, 2, 3, ..., q, \quad j = 1, 2, 3, ..., m, \tag{2.12}$$

where μ_i indicates the mean squared deviation between the predictions and the outcomes of the i^{th} expert and, the errors ε_{ij} are assumed to be independently distributed as $N(0, \sigma^2)$. Here, μ_i can also be viewed as the innate skill of the i^{th} expert and Y_{ij} as a noisy measure of this skill. We begin modeling by assuming the random errors as an independent, constant variance, normally distributed random variable with mean zero. However, the model can be modified if it does not seem appropriate.

The above model can generally be fitted in situations where the observations of a certain response variable of interest are obtained under different conditions in different treatment groups. The population means of responses are required to be estimated in each treatment group and they are estimated by the corresponding sample means within groups. It is not necessary to have an equal number of responses in each treatment group. The treatment groups in our context are the experts and the squared deviations between the predicted probabilities and the outcomes of events are the responses of interest. Obtaining the sample means of responses is equivalent to computing experts' Brier scores using the equation 2.5. Hence, the above linear model gives mean estimates that are identical to the typical experts' Brier scores irrespective of whether the experts have predicted all the events or not.

We mentioned in section 2.3 that typical computation of experts' Brier scores fails to capture the potential correlation structure induced in the probability predictions by asking common questions. Therefore, the independence assumption of ε_{ij} can

be violated if we ignore this potential correlation structure in the model 2.12. The standard errors of fixed effects parameters are estimated in linear models by assuming the independence of errors. Therefore, if the independence assumption of errors is violated, then the estimated standard errors of experts' Brier scores from the model 2.12 are not accurate.

The correlations between probability predictions due to the questions' effects can be included into the above model either as fixed effects or as random effects. If we include them as fixed effects, then the model needs to estimate the effects of each and every level of the questions as separate parameters in the model. These parameters will behave as nuisance parameters and avoid estimating μ_i's directly as experts' Brier scores in the model. If we include the questions' effects as random effects, then the experts' Brier scores can directly be estimated as μ_i's from the model with improved standard error estimates. The inclusion of questions' effects as random effects gives the following linear mixed-effects model.

$$Y_{ij} = \mu_i + \delta_j + \varepsilon_{ij}, \tag{2.13}$$

where δ_j represents the effect of the j^{th} level of the random effect for the questions. It is assumed that $\delta_j \sim N(0, \sigma^2_{que})$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$. Note that σ^2_{que} and σ^2 indicate the between question variation and the random error variation respectively. Similar to the case of linear fixed-effects model (equation 2.12) above, we begin with assuming both δ_j and the ε_{ij} as independent, constant variance, normally distributed random variables with mean zero and look for possible improvements if requires. The random effect for the questions will introduce the correlation structure induced in the responses; Y_{ij}, due to the effect of questions into the model. Therefore, it can be expected to improve the accuracy of the estimated standard errors of experts' Brier sores from this model.

The correlation between any two observations; Y_{ij}, within the same question; intra-class correlation of questions, can be defined from the standard definition of Pearson's

correlation coefficient as

$$\rho = \frac{\sigma^2_{que}}{\sigma^2_{que} + \sigma^2}, \tag{2.14}$$

and it can be interpreted as the proportion of unknown error variance explained by the common questions (Rodrıguez and Elo, 2003). Note that the above described models will be fitted on the data obtained from the "Intelligence Game" forecasting tournament.

2.5.1 The Intelligence Game

The Intelligence Advanced Research Projects Activity (IARPA) is an organization within the Office of the Director of National Intelligence in United States of America (USA). IARPA is responsible for leading research to overcome difficult challenges relevant to the United States Intelligence Community. It announced a program called the Aggregative Contingent Estimation (ACE) in year 2010 with the aim of "dramatically enhance the accuracy, precision and timelines of forecasts for a broad range of events types, through the development of advanced techniques that elicit, weight and combine the judgements of many intelligence analysts" (Wintle et al., 2012). The program was designed as a four year forecasting tournament to predict the probabilities of the occurrence of some global events on geopolitical, economic, and military sectors. Five collaborative research teams involved in a competition. This forecasting tournament was called the "Intelligence Game".

Following details about the intelligence game are due to Wintle et al. (2012). Each month IARPA released a list of questions asking to predict the probabilities of the occurrence of some global events relevant to the time period concerned. Participants of the teams submitted their probability predictions using the 3-step question format below in the first round and their second round predictions were made after a feedback and a discussion. The outcome of each event was classified as "occurred" and "not occurred" and the Brier scores were used to measure the prediction accuracy of the participants and the teams of the tournament.

The 3-step question format

For each question, the probabilities were elicited using the following structured 3-step question format:

1. The highest plausible probability of event occurrence:
   (Please answer with a percentage 0-100)
2. The lowest plausible probability of event occurrence:
   (Please answer with a percentage 0-100)
3. The best guess probability of event occurrence:
   (Please answer with a percentage 0-100) (Wintle et al., 2012)

The data

The Australian Centre of Excellence for Risk Analysis (ACERA) at the University of Melbourne has contributed to one of the teams was led by the members at the George Mason University in USA. Members of the two institutes formed a joint team called the Decomposition-Based Elicitation and Aggregation (DAGGRE). ACERA's role was to elicit predictions from the groups in USA and Australia using a structured Delphi-style iterative elicitation procedure. The data collected by ACERA in the first year of the tournament will be used for the analyses of this chapter.

Section 2.6: The analysis

We analyse the first-round best guess probabilities of event occurrence from the first-year data of the above tournament. Note that the second-round probability predictions were made by the participants after a feedback and a discussion within groups. Therefore, the probability predictions can be correlated due to the common groups' effects as well. We are interested to assess the impact of incorporating the potential correlations between probability predictions due to the effects of common questions on the estimated standard errors of experts' Brier scores in this analysis.

Therefore, we consider the first round predictions to avoid the potential correlations due to the effects of groups from the analysis.

We selected a complete subset of probability predictions by the participants for the analysis as mentioned in section 2.4. It includes probability predictions on 12 questions by all the selected 16 participants . There are 4 variables in the data as described below:

QuestionId - identification number of a question,
ParticipantId - identification number of a participant,
Bestguess - first round best guess probability of an event occurring, and
Outcome - outcome of the question.

We acknowledge the fact that Brier scores and their standard error estimates are more robust if we select more questions for the analysis. However, we could not manually select a complete subset of predictions of which a selected set of participants answering a large number of questions. A data filtering approach was used to select a complete subset of predictions for the analysis. The typical Brier scores and their standard error estimates for the selected participants are given in Appendix A.

It follows from the discussion in section 2.5 that typical and the linear fixed-effects model (equation 2.12) based Brier scores will be identical for the participants. If we consider the estimation of standard errors of Brier scores, potential correlations between probability predictions due to the effects of common questions have not been considered in both methods. Therefore, the independence assumption of random errors ε_{ij} in the linear fixed-effects model can be violated and inaccurate standard error estimates of Brier scores can be obtained. Furthermore, the constant variance assumption of random errors ε_{ij} in the model assumes that variances of responses are equal for the participants. Therefore, the individual variance estimates of random errors within participants will be pooled together to estimate the overall variance of random errors that will be used to compute the standard error estimates of participants' Brier scores. If we consider the context of a complete set of predictions we are working on, the standard errors will be estimated equal for all the participants as they all have answered an equal number of questions.

If we consider the typically computed standard error estimates of participants' Brier scores given in Appendix A, there seem to have considerable differences between them. Pinheiro and Bates (2000, chap. 5) discussed fitting linear fixed-effects models with both constant and non-constant within-group errors using the generalized least squares estimation method. Therefore, we are interested to fit the above discussed linear fixed-effects model together with the following extended linear fixed-effects model with non-constant variances of within-participant errors.

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \tag{2.15}$$

where the difference from the model in equation 2.12 is to assume that $\varepsilon_{ij} \sim N(0, \sigma_i^2)$ have non-constant variances within participants; $i = 1, 2, 3, ..., 16$. We refer the linear fixed-effects model (equation 2.12) with constant variance assumption of errors as *Linear_c* model and the linear fixed-effects model (equation 2.15) with non-constant variance assumption of errors as *Linear_nc* model in the analysis.

Note that the "gls (generalized least squares)" function of the "nlme (Linear and Nonlinear Mixed Effects Models)" package (Pinheiro et al., 2017) of the R software package was used to fit the above two linear models. Table B.1 in Appendix B shows that the standard error estimates of Brier scores from the *Linear_nc* model are identical to the corresponding typically computed estimates. Furthermore, the constant variance assumption of errors in *Linear_c* model leads to obtain equal standard error estimates of Brier scores due to the pooled variance estimation discussed above. The small p-value of the likelihood ratio test appears in table 2.1 confirms the assumption that the variability of the errors are different for participants, even though there is a mismatch between the AIC (Akaike information criterion) and BIC (Bayesian information criterion) information criteria values.

TABLE 2.1: Comparison of the adequacy of linear models

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
Linear_c	1	17.00	-15.70	39.67	24.85			
Linear_nc	2	32.00	-40.76	63.48	52.38	1 vs 2	55.06	0.00

It is a well-known fact that both AIC and BIC values apply penalty terms on the log-likelihood function to reduce the effect of overfitting models. The penalty is

stronger for BIC than AIC for any reasonable sample size as the penalty term of the BIC depends not only on the number of estimated parameters of the fitted model as for the case of AIC but also on the sample size of the analysis. Therefore, due to the applied stronger penalty, BIC tends to choose smaller models than AIC in general (Brewer, Butler, and Cooksley, 2016). Hence, sometimes AIC and BIC can disagree as they measure different things. Delattre, Lavielle, and Poursat (2014) also discussed a potential issue of using different expressions for BIC by different software packages due to the reason of not clearly define the effective sample size and the effective number of parameters in the context of fitting mixed-effects models. Fitting a mixed-effect model with questions' effects as random effects is also a part of this analysis as explained in equation 2.13 above. Therefore, we focus only the p-value of the likelihood ratio test for choosing models in this analysis.

Pinheiro and Bates (2000, chap. 5) also discussed the fact that choices between models should not be made just based on what the information criteria and likelihood tests suggest but it is important to consider the practical context of fitting models as well. Therefore, we consider the practical importance of identifying a model that produces Brier scores and their standard error estimates that are similar to the corresponding estimates from the typical computation of Brier scores and focus on possible improvements afterward. Figure 2.1 confirms this result.

It is important to note that potential correlations between probability predictions due to the effects of common questions has not been incorporated to the analysis yet in the model discussed above. Therefore, now consider fitting the linear mixed-effects model in equation 2.13 with questions' effects as random effects. The random effects for questions will accommodate potential correlations between probability predictions due to the common questions in the model. It follows from above that we assume non-constant variances for random errors within participants. Note that (Pinheiro and Bates, 2000, chap. 5) also discussed the possibility of extending linear mixed-effects models to allow heteroscedastic or non-constant variances for within-group errors of a given stratification variable. Furthermore, there is no restriction on the grouping factor to be a fixed-effect or a random-effect in the model. Therefore, we consider participants as the stratification variable of the model. It leads to fit the following adjusted linear

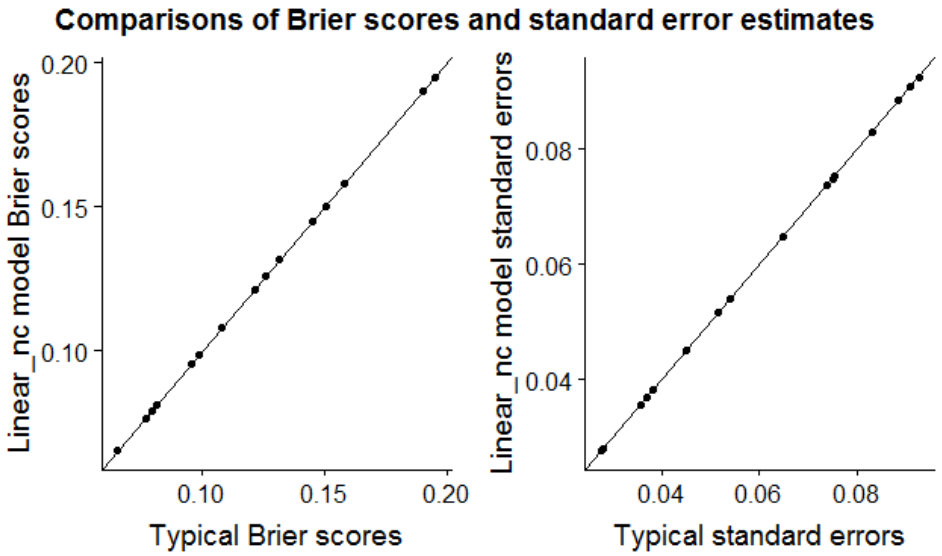


FIGURE 2.1: The scatter plots of Brier scores and standard error estimates with added x=y lines from typical and *Linear_nc* model based computations

mixed-effects model with non-constant variances of errors within participants.

$$Y_{ij} = \mu_i + \delta_j + \varepsilon_{ij}, \tag{2.16}$$

where the difference from the model 2.13 is to assume that $\varepsilon_{ij} \sim N(0, \sigma_i^2)$ have non-constant variances within participants; $i = 1, 2, 3, ..., 16$. We refer this model as *Mixed_que* as it includes questions' effects as random effects. The "nlme" package can also be used to fit this model.

It can be expected that participant's Brier scores similar to the typical computation will also be obtained from *Mixed_que* model for this complete set of predictions. However, we expect to improve the accuracy of the estimated standard errors of participant's Brier scores from *Mixed_que* model. The likelihood ratio test with a very small p-value suggests that *Mixed_que* model better reflects the data than the *Linear_nc* model as indicates in table 2.2.

TABLE 2.2: Comparison of the adequacy of linear and mixed-effects models

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
Linear_nc	1	32.00	-40.76	63.48	52.38			
Mixed_que	2	33.00	-122.09	-14.59	94.04	1 vs 2	83.33	0.00

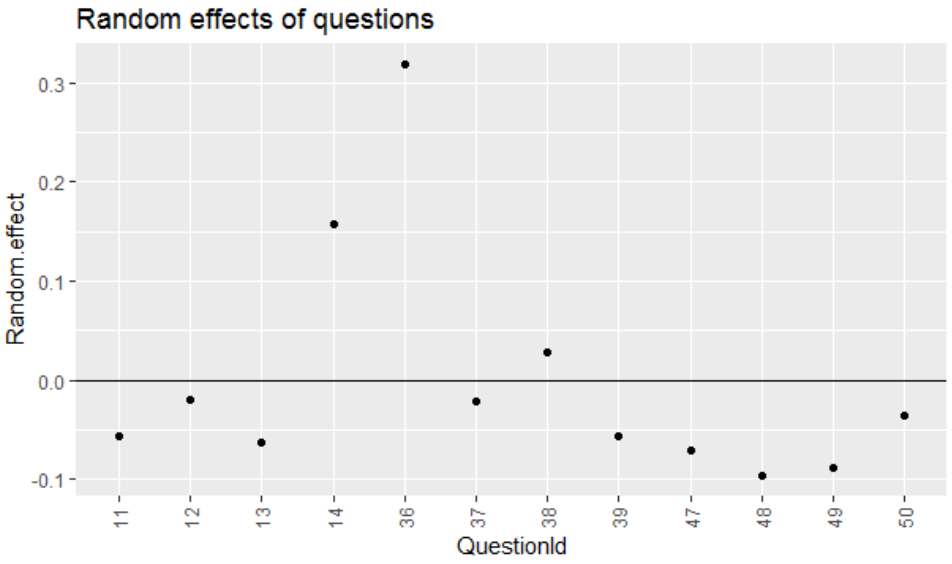


FIGURE 2.2: The computed random effects of questions from the *Mixed_que* model

The estimates of random effects of questions are given in figure 2.2. It indicates that there are differences between the effects of questions on the response of interest of the fitted *Mixed_que* model. This causes to have within-question correlations of responses that needs to be addressed to obtain accurate standard error estimates of experts' Brier scores. Therefore, it is reasonable to assume that the accuracy of the estimated standard errors of participants' Brier scores are improved from the *Mixed_que* model compared to the typical computation of Brier scores. Figure 2.3 indicates that participants' Brier scores are similar but the standard error estimates are different between the typical and mixed-effects model based computations as expected.

Also note that between questions variation and the unexplained random error variation were found to be approximately 0.0155 and 0.0144, respectively from the model. Therefore, the proportion of unknown error variance explained by the common questions was found approximately equal to 0.52 using the equation 2.14. The standard error estimates of participants' Brier scores obtained from *Mixed_que* model are given in Appendix C.

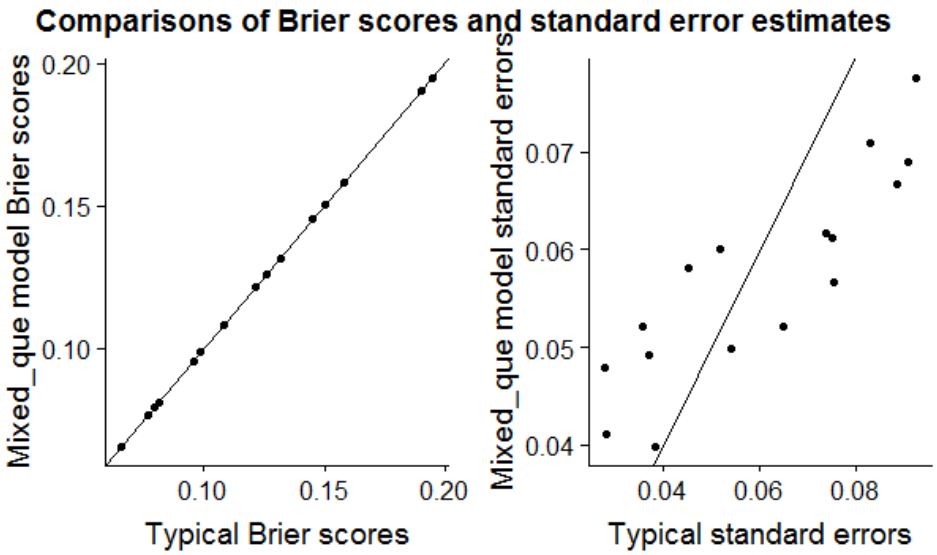


FIGURE 2.3: The scatter plots of Brier scores and standard error estimates with added x=y lines from typical and *Mixed_que* model based computations

Section 2.7: Discussion

It follows from the discussion in section 2.5 that the values of the response variable; squared deviations between the probability predictions of the occurrence of events and the outcomes of the corresponding events, of the fitted linear and mixed-effects models to compute Brier scores remain within the maximum range of [0, 1]. Therefore, we relaxed the normality assumption of errors of the fitted models. Furthermore, we also assumed non-constant variances of random errors within participants as suggested by the data. The normality and constant variance assumptions of errors are important when the estimated means are statistically tested for equality in the analysis of variance of models. We do not focus on comparing the estimated participants' Brier scores from the model. We are interested only on computing Brier scores and improve the accuracy of their standard error estimates by incorporating potential correlations between predictions due to the effects of common questions. Therefore, satisfying the model assumptions is not necessary for our computations. Once the accuracy of the estimated standard errors are improved, computed Brier scores from the model can be used to compute more precise confidence intervals for necessary comparisons between Brier scores.

One may think about a logistic mixed-effects model with $log\,\frac{p_{ij}}{1-p_{ij}}$ as the response variable to fit a model with satisfied model assumptions in this context. However, the deviation between the predictions and the outcomes would be less visible in this formulation and harder to interpret and relate back to the Brier score. Therefore, for the time being, we consider the suggested approach that produces similar estimates to Brier scores with improved standard error estimates is useful.

Mixed-effects models are developed to deal with multi-level or hierarchical data structures in a wide variety of experimental designs. Therefore, this mixed-effects model-based approach of computing Brier scores and their standard error estimates can be extended to multi-level data structures with more levels. Suppose we are interested to analyse the second-round best guess probabilities of event occurrence of the intelligence game data and add the grouping level of experts also as a random effects into the fitted *Mixed_que* model. Then, it is important to note that two random effects for the groups and questions will not be nested in the model. The "nlme" package of the R software package requires that the random effects are nested in the model (Bates et al., 2015). On the other hand, the "lme4 (Linear Mixed-Effects Models Using Eigen and S4" package (Bates et al., 2014) allows fitting non-nested random effects in mixed-effects models. However, it does not allow modelling non-constant within group errors in mixed-effects models (Bates et al., 2015). Thus, in circumstances where the within group error variances can be assumed constant, "lme4" package allows computing Brier scores with multiple random effects.

It is also important to note that the above discussed general approach of computing Brier scores and their standard errors using linear mixed-effects models can be extended to compute Logarithmic and Spherical scores. The response variables of the models to compute Logarithmic and Spherical scores can be defined accordingly. It should be noted that the linear model assumptions will not be satisfied according to the way the response variables are defined as in the case of computing Brier scores. However, the above discussed approach can be used for necessary comparisons of scores between experts with improved standard error estimates.

Section 2.8: Conclusion

Experts' Brier scores are derived from expert elicited experiments to assess the prediction accuracy of experts on predicting probabilities of the occurrence of events. The focus of this analysis is to improve the estimation of standard errors of experts' Brier scores by incorporating the potential correlation structure induced in the probability predictions by asking common questions from experts. This analysis using the data from the intelligence game (refer the section 2.5.1) shows that the standard error estimates of participants' Brier scores using the fitted linear mixed-effects model with questions' effects as random effects are different from the typically computed standard error estimates of participants' Brier scores ignoring correlations between probability predictions due to common questions' effects.

Following the statistical theory of obtaining inaccurate standard error estimates of parameters if potential correlations between observations are ignored (Finch, Bolin, and Kelley, 2016) and the ability of random effects of mixed-effects models to handle correlations between observations due to sharing the effects of the same levels of grouping factors (Pinheiro and Bates, 2000; Harrison et al., 2018) discussed , it can be concluded that the standard error estimates of participants' Brier scores from the fitted linear mixed-effects model with questions' effects as random effects are more accurate than the typically computed standard error estimates of participants' Brier scores. Therefore, we recommend to use mixed-effects models to compute Brier scores and their standard error estimates using necessary random effects to incorporate potential design-based correlated sources of predictions from hierarchical data.

3. Missing values, and ways of dealing with them

Section 3.1: Introduction

This chapter focuses on methods for estimating the missing probability predictions of events for computing experts' Brier scores that are used to assess the prediction accuracy of experts. It follows from the previous chapter due to Predd et al. (2008) that human judges (even experts) may prefer to assess only the subsets of events of which they feel comfortable to offer coherent predictions in practice. If experts' Brier scores are computed using the probability predictions of different subsets of events, then the comparison of the prediction accuracy of experts using Brier scores can be challenging and perhaps may be less meaningful (Merkle et al., 2016; Hanea et al., 2018). Hence, in order to enhance the comparability of experts' Brier scores, it is important to compute Brier scores by adjusting for the missing probability predictions of events by experts. Thus, we apply some missing value estimation methods to estimate missing predictions in computing experts' Brier scores in this analysis.

It was discussed in the previous chapter that experts' Brier scores are computed from hierarchical data of which the probability predictions can be correlated due to the effects of different levels of the hierarchy of experimental designs. Therefore, it can be suggested following the discussions in previous chapter that mixed-effects models will better reflect the data in these circumstances and will be useful to accurately estimate missing predictions. Hence, we investigate the potential of using mixed-effects models to satisfy the modelling requirement of some of the known missing value estimation methods in this analysis. The results of a conducted simulation study of introducing

missing predictions completely at random and not missing at random show that multiple imputation method using a mixed-effects model with questions' effects as random effects can estimate missing probability predictions to compute experts' Brier scores with reduced errors compared to the typically computed experts' Brier scores that ignore missing predictions. We begin the analysis with the following general review of observing missing values in data.

Section 3.2: Missing values in data

First consider the general background of observing missing values and the consequences of ignoring them in statistical analyses. According to Kaiser (2014), missing values are the values that were intended to be obtained during data collections but failed due to various reasons including respondents of studies have not answered all the questions, errors in data entry processes, incorrect measurements, experimental errors, censoring of data, and many others. If we consider the general consequences of ignoring missing values, they cause to obtain biased estimates of parameters and increase their standard error estimates. In addition to that, the statistical power of tests can be decreased and the generalizability of results can be weakened (Dong and Peng, 2013). Therefore, handling missing values is very important before carrying out statistical analyses.

3.2.1 Missing data mechanisms

Prior to discuss about different methods of dealing with missing values, we consider the following different types of missing data mechanisms under which missing data can occur in practice. According to Tabachnick and Fidell (2013), not only the percentages of missing values but identifying the missing data mechanisms and missing data patterns is also very important to assess the missing value problem. It is important to note that problems caused by missing values and potential solutions to them are different for different types of missing data mechanisms. Missing data mechanisms are usually classified as missing completely at random, missing at random, and not missing at random (Kaiser, 2014; Lakshminarayan, Harp, and Samad, 1999; Horton

and Kleinman, 2007). We consider some simple easy to understand definitions and examples of different missing data mechanisms due to Kaiser (2014) as follows.

Missing Completely at Random (MCAR)

The missing data mechanism is considered as MCAR if the probability of a record having a missing value for a variable does not depend on either the observed or the missing data. Suppose a laboratory sample is dropped and the resulting observation is missing without depending on either the observed or missing data. Therefore, this is an example for a MCAR mechanism. According to Scheffer (2002) and Bennett (2001), MCAR mechanism is very rarely observed in practice.

Missing at Random (MAR)

Now consider the MAR mechanism that occurs when the probability of a record having a missing value for a variable does not depend on the value of the missing data itself but could be dependent on the observed data. Consider a context of collecting income data with given information on property tax bands. It can be expected that people with higher incomes may not reveal the correct figures considering the amount of tax charges to be paid in practice. Therefore, a missing value in the income variable may not depend on the value of income itself but can be dependent on the value of the property tax band given. Hence, missing values in income data would be an example for a MAR mechanism. According to Scheffer (2002), it is more likely to meet the MAR condition if the data are collected on related variables.

Not Missing at Random (NMAR)

The last type of mechanism to consider is the NMAR. It occurs when the probability of a record having a missing value for a variable could depend on the actual value of the variable itself. Consider a context of which a sensor is not able to detect temperatures below a certain threshold value. Thus, the observed missing values of temperature depend on the actual values themselves. Therefore, this is an example for a NMAR mechanism. According to Scheffer (2002), NMAR is non-ignorable, even though both

MCAR and MAR are ignorable for likelihood-based imputation methods. It means that the missing data mechanism has to be modelled as you deal with missing data in this context. Kaiser (2014) suggested to examine the source of the data to identify an appropriate model for the missing data mechanism. However, as Kaiser (2014) pointed out, it is very rare to meet circumstances where appropriate models can be identified to model missing data mechanisms in practice.

Section 3.3: Different ways of dealing with missing values

After the review of different missing data mechanisms above, now consider different methods of dealing with missing values can be applied under different data mechanisms. We review some of the known traditional and advanced methods in general and discuss their potential to apply for estimating missing predictions in computing experts' Brier scores. We begin with some of the traditional methods.

Case deletion methods

First consider the following result from a survey conducted by Peng et al. (2006) on quantitative studies published from 1998 to 2004 in 11 education and psychology journals. Survey reveals that among the studies that have shown evidence of missing values, 97% have used either listwise or pairwise deletions of missing values. These two deletion methods fall into the general category of case deletion. According to Dong and Peng (2013), both listwise and pairwise deletion methods are adhoc and produce biased and/or inefficient estimates in most situations.

Listwise deletion is all about deleting or ignoring cases with missing values for any of the variables and analyse the remaining data. This approach is also known as the complete case (or available case) analysis (Kang, 2013). Observe that listwise deletion method is used in typical computation of experts' Brier scores ignoring missing predictions. It leads to obtain experts' Brier scores that are computed from different sets of questions. Hence, the resulting comparisons of experts' Brier scores may be less meaningful in practice. We mentioned about two papers; (Merkle et al., 2016; Hanea et al., 2018), discussed about this drawback of comparing experts' Brier scores

that are computed from different sets of questions above. However, we have not found references on improving the comparability of experts' Brier scores by adjusting for the missing predictions. This is the focus of the analysis of this chapter.

If we consider the pairwise deletion method, it eliminates cases only when the data are missing in variables of which the pairwise analyses such as correlations, t-tests are carried out (Graham, 2009). Therefore, it is not necessary to delete all the cases with missing values elsewhere in the data set. There are no pairwise analyses between variables involved in computing experts' Brier scores. Therefore, we do not consider the pairwise deletion method in this analysis. According to Scheffer (2002), case deletion and mean imputation are the default methods that used in major statistical packages. Hence, we now review the mean imputation method as follows.

Mean imputation method

Following review on the mean imputation method is obtained from Kang (2013). Mean imputation substitutes the missing values of a given variable by the mean value of the observed values of that variable. Mean imputation method can lead to inconsistent bias if the missing values occur strictly not random and there are differences between the number of missing values of different variables. Furthermore, no new information is added but only the sample size is increased. Therefore, the standard error estimates of parameters can be underestimated. According to Bennett (2001), mean imputation method needs to satisfy the MCAR assumption to obtain unbiased results. Therefore, we will consider applying this method for imputing missing values of probability predictions in computing experts' Brier scores if the missing predictions can be assumed to satisfy the MCAR condition.

Regression imputation method

Now consider the regression imputation method that has been developed after the case deletion and mean imputation methods (Scheffer, 2002). This method fits a regression equation considering the observed values of a given variable with missing values as the response variable and all other relevant variables in the data set as predictor variables.

Then, the predicted values from the fitted regression equation are used to impute missing values (Bennett, 2001). According to Graham (2009), this method provides a sound basis for many of the modern missing value estimating methods. The results are unbiased if MCAR or MAR conditions hold (Bennett, 2001).

We mentioned about our interest to use mixed-effects models to incorporate potential correlations between probability predictions to better reflect the data that will be useful to accurately estimate missing values of probability predictions in computing experts' Brier scores above. It was also mentioned in the previous chapter that inclusion of necessary random effects to model correlations between probability predictions due to sharing the effects of the same levels of grouping factors of an experimental design will allow direct estimation of experts' Brier scores as fixed means from an appropriate mixed-effects model. If we consider the grouping factors as fixed-effects, then they will behave as nuisance parameters in the model avoiding direct estimation of experts' Brier scores form the model. Furthermore, the requirement to estimate the effects of each and every level of grouping factors as separate parameters in the model will restrict the number of levels of factors to be considered in the analysis. Therefore, in order to keep the flexibility to consider any available number of levels of grouping factors and to directly estimate experts' Brier scores as fixed means, we will use mixed-effects models appropriately in regression imputation in this analysis.

In addition to the above mentioned frequently used traditional methods, Bennett (2001) also suggested that the last value carried forward, hot-deck imputation, and cold-deck imputation are three of another possible methods. The last value carried forward method is used to estimate missing values of longitudinal data of which repeated measurements of a particular characteristic are obtained over time for each study participant (Laird and Ware, 1982). If the data are found to be missing at a particular point in time for some participants, then the last available values of participants will be carried forward to estimate the missing values (Bennett, 2001). Note that this method is not applicable for estimating missing predictions of different questions by experts.

If we consider the hot-deck and cold-deck imputation methods, they rely on replacing missing values with values taken from matching covariates that need to be identified

from the existing data for the hot-deck method and from the external or prior information for the cold-deck method. Identifying matching covariates may not be possible in forecasting scenarios of which the data are available only on predicting relevant probabilities of given sets of questions by experts. Thus, the hot-deck and cold-deck imputation methods seem not applicable to the context of our study. Therefore, we only consider mean imputation and regression imputation as the traditional methods of imputing missing values in this analysis.

The above discussed traditional methods only perform single imputations of missing values. Harel and Zhou (2007) discussed the importance of considering the uncertainty of an imputation process by repeating the process multiple times. Therefore, it is also of interest to apply suitable advanced missing value estimation methods that employ iterative procedures to take into account the underlying uncertainty of the imputation process. Dong and Peng (2013) and Graham (2009) suggested that the multiple imputation, full information maximum likelihood, and EM (Expectation-Maximization) algorithm are three of such commonly used advanced missing value estimation methods. In addition to that, Markov-chain imputation method is also used in practice (Bennett, 2001).

The Markov-chain imputation method is usually applied to longitudinal (or repeated measures) data (Bennett, 2001). Therefore, this method can also be ignored from our context of computing experts' Brier scores as similar to the case of ignoring the traditional method of last value carried forward described above for applying with longitudinal data. The full information maximum likelihood is a model-based missing data estimation method commonly applied in structural equation modelling (Dong and Peng, 2013). The structural equation modelling technique is widely used in behavioural sciences. It provides a general and a convenient framework to conduct statistical analyses in several multivariate procedures including factor analysis, multivariate regression analysis, discriminant analysis, canonical analysis, and so on (Hox and Bechger, 1998). Therefore, this method is also not applicable to the context of computing experts' Brier scores.

Following the above discussion, we can restrict our attention to the multiple imputation and EM algorithm methods that employ iterative procedures to estimate missing

values. The EM algorithm proceeds as an iterative approach that consists of two steps in each iteration. They are the Expectation step (E-step) and Maximization step (M-step) (Bennett, 2001). In the E-step, best guesses for the missing values are obtained by evaluating the specified model on the observed data. In the M-step, the missing data are substituted by the expected or the fitted values obtained in the E-step and updated parameter estimates of the model are obtained by maximizing the likelihood function assuming no data are missing. The updated parameter estimates will be substituted back into the E-step and a new M-step will be performed. This procedure will be continued until the change of the parameter estimates will be negligible between iterations where the convergence is reached (Bennett, 2001).

It was mentioned above that we are interested to apply mixed-effects models to compute experts' Brier scores as fixed mean estimates with missing predictions. However, we have not found the applications of EM algorithm for estimating the parameters of fixed effects of linear mixed-effects models with missing values. Therefore, we restrict our attention only to the multiple imputation method as described below.

Multiple imputation method

Imputing one value for each missing value in single imputation has an obvious disadvantage that arising from the fact that the imputed single value cannot represent any uncertainty about which value to input. Hence, treating imputed values just as observed values here generally systematically underestimate the uncertainty even under assuming the precise reasons for nonresponse are known. Furthermore, single imputation cannot represent any additional uncertainty when the reasons for nonresponse are unknown. In contrast, the multiple imputation method for nonresponse that was proposed in Rubin (1978) and Rubin (2004b) replaces each missing value by two or more plausible values and the values can be chosen to represent uncertainty about which values to impute under assuming the reasons for nonresponse are both known and unknown (Rubin, 1988). The theoretical underpinnings and several examples relevant to this context can be found in Rubin (2004a).

Multiple imputation replaces each missing value by a vector composed of $M \geq 2$ possible values as mentioned above. These imputed missing values are ordered to

create separate completed data sets that can be separately analysed using standard complete-data methods. The uncertainty of which values to impute is of two types: sampling variability assuming the reasons for nonresponse are known and unknown. Under each hypothesized model for nonresponse, two or more imputations are created to reflect sampling variability under the model. Here, imputations under more than one model for nonresponse reflect uncertainty about the reasons for nonresponse and the inferences under different models can be contrasted to reveal sensitivity of answers to posited reasons for nonresponse. Furthermore, repeated analyses based on multiple imputations within one model can be combined to form valid inferences under that model (Rubin, 2004a). There is an extensive literature on multiple imputation. Herzog and Rubin (1983), Rubin (1986), Rubin and Schenker (1987), Treiman, Bielby, and Cheng (1988), and Rubin (2004b) are some of them mentioned in (Rubin, 1988). We focus on Little and Rubin (2002) that provides further explanation on the steps of multiple imputation following the description above.

There are three main stages in the multiple imputation process. They are the imputation stage, analysis stage, and combining stage of results. The imputation stage involves replacing each missing value by a vector $M \geq 2$ of imputed values and order them to create M complete data sets. First complete data set is obtained by replacing each missing value by the first component of the vector of imputations M and the second complete data set is obtained by replacing the corresponding second component and so on. Here, M sets of imputations are assumed to be repeated random draws from the predictive distribution of missing values under a particular model for nonresponse. In the analysis stage, M complete data sets are analysed separately using the intended statistical analysis technique. In the combining stage of results, the separate results obtained from M complete data sets are combined to form one overall result of the analysis.

Combining or the pooling stage produces a single set of estimates of the parameters of interest and their standard errors. The pooled standard error estimate of a parameter incorporates the between imputation uncertainty into the uncertainty inherent in the estimation method (the within imputation uncertainty) applied. Therefore, the pooled standard error estimate is larger than the corresponding estimate derived from a single

imputation that does not consider the between imputation uncertainty. Hence, the bias in the standard error estimate of a parameter in a single imputation method (e.g., mean imputation and regression imputation discussed above) can be reduced using the multiple imputation method (Dong and Peng, 2013). Following details about the estimation of parameters in the multiple imputation method are due to Little and Rubin (2002):

Let $\hat{\theta}_i; i = 1, 2, 3, ..., M$ and $W_i; i = 1, 2, 3, ..., M$ be the estimates for θ and their variances calculated from M repeated imputations from a given model. Then, the combined estimator for θ is

$$\overline{\theta}_M = \frac{1}{M} \sum_{i=1}^{M} \hat{\theta}_i. \tag{3.1}$$

The variability associated with this estimate has two components: the average within-imputation component,

$$\overline{W}_M = \frac{1}{M} \sum_{i=1}^{M} W_i \tag{3.2}$$

and the between-imputation component,

$$B_M = \frac{1}{M-1} \sum_{i=1}^{M} (\hat{\theta}_i - \overline{\theta}_M)^2. \tag{3.3}$$

Hence, the total variability associated with $\overline{\theta}_M$ is

$$T_M = \overline{W}_M + \Big(1 + \frac{1}{M}\Big) B_M, \tag{3.4}$$

where $\Big(1 + \frac{1}{M}\Big)$ is an adjustment for finite M.

A good imputation model with a random component is required to create M complete data sets with imputed missing values for a given analysis. Identifying a good imputation model follows the need of understanding the structure of the data and the variables that need to predict missing values accurately. We described above that a mixed-effects model with necessary random effects to incorporate potential

correlations between probability predictions will better reflect the data in the context of computing experts' Brier scores. Therefore, an appropriate mixed-effects model will be useful to accurately estimate missing values of probability predictions. Once the missing values are estimated, we need to use the same model to estimate experts' Brier scores as fixed mean estimates from complete data sets. Theoretically, it is possible to use the same model for both imputation and analysis models (Dong and Peng, 2013). Therefore, we will use a suitable mixed-effects model for both models in the analysis.

It is also important to note that multiple imputation method is valid if the missing data mechanism is either MCAR or MAR (but not valid under NMAR) and the percentage of missing values is not too high (Scheffer, 2002). Harel and Zhou (2007) mentioned that "the distinction between missing at random and not missing at random is based on a non-testable assumption". In relevance to this, Dong and Peng (2013) pointed out that statistical tests cannot be expected to provide definitive evidence of satisfying either MCAR or MAR conditions in practice. Therefore, we consider assessing the modes of missing data mechanisms by looking at the context of the data rather than employing statistical tests in this study.

Section 3.4: Methodology of the study

We will perform a simulation study to explore the potential of the mean imputation, regression imputation, and multiple imputation methods to estimate the missing probability predictions of events for computing experts' Brier scores. A complete subset of probability predictions made by experts will be considered as in the previous chapter and some selected percentages of missing values will be randomly introduced into the predictions. Computed experts' Brier scores with estimated missing predictions using the above mentioned missing value estimation methods will be compared with the typical Brier scores that ignore missing predictions. A suitable measure of distance from the original Brier scores that are computed without introducing missing predictions will be used for necessary comparisons between Brier scores above.

The exact missing data mechanism of observing missing predictions is practically unknown in the context of computing experts' Brier scores. Therefore, we assume

two scenarios of which missing data occur completely at random and missing data occur not at random but depending on the actual values of themselves conditional on the levels of difficulty of questions in the analysis. We also consider two different ways of introducing the selected percentages of missing values into the probability predictions made by experts in the analysis. In the first case, we introduce the selected percentages of missing values directly into the overall set of predictions made by experts to represent a context of which different experts have different number of missing predictions. Secondly, we introduce the selected percentages of missing values equally into the predictions made by each expert individually.

3.4.1 The data

The same intelligence game data was used as in the previous chapter and a complete subset of predictions without missing values was selected from the third year data of the forecasting competition. Third year second round data was used considering the possibility of incorporating the potential correlations between probability predictions not only due to the effects of common questions but also due to the common grouping effects of participants into the analysis. The selected subset includes data from 6 participants answering 31 questions each. Therefore, the data contain a total of 186 probability predictions. Following variables are included in the data.

ParticipantId - identification number of a participant,
QuestionId - identification number of a question,
GroupId - identification number of a group,
Bestguess - second round best guess probability prediction of an event occurring, and
Outcome - outcome of the question.

Missing values will be randomly introduced into the “Bestguess” variable of probability predictions as described above. Once the missing values are estimated, (Bestguess - Outcome)^2 will be computed for computing experts’ Brier scores in the analysis.

Section 3.5: The analysis

It was mentioned above that the purpose of using the third year second round data of the intelligence game is to consider the possibility of incorporating potential correlations between probability predictions due to the effects of common questions and the groups of participants into the analysis. If we consider the participants' Brier scores and their standard error estimates of the selected complete set of predictions given in table 3.1, it can be observed that standard error estimates are not varied considerably between participants as in the analysis of the previous chapter. Therefore, we can relax the need of using non-constant within group variances of predictions for participants and look into fitting a mixed-effects model with non-nested random effects for both questions and groups as described in the discussion section of the previous chapter.

TABLE 3.1: Brier scores and their standard error estimates

ParticipantId	Brier score	Standard error estimate
51	0.14732	0.04906
73	0.17228	0.04276
76	0.16387	0.03052
133	0.14272	0.03398
251	0.14703	0.03401
252	0.16968	0.03131

Considering the total number of 186 predictions for introducing random missing values directly into the overall set of predictions made by experts and 31 predictions for introducing random missing values for each individual expert, we consider 10% as a reasonably small percentage of missing values and 25% as a reasonably large percentage of missing values. Therefore, we reduce the scope of this analysis to 10% and 25% missing values. The analysis will be repeated 1000 times at each percentage of missing values and participants' average Brier scores with missing values will be computed to take into account the random variation of Brier scores due to the introduction of a given percentage of missing values randomly. Root mean squared errors (RMSE) of computed Brier scores with missing values will also be estimated over 1000 repeats since the original Brier scores without missing values are known. Even though, the Brier score is a unit less measure, RMSE is chosen to make sure that the measure of

variation of the estimated Brier scores from their true values is in the same scale of measure of the Brier scores.

Now recall from above that there are 6 participants in the selected data for the analysis. Let $\tilde{O}_i$ indicate the Brier score of the i^{th} participant; $i = 1, 2, 3, ..., 6$, from the original data without introducing missing values and $\tilde{I}_i$ indicate the corresponding Brier score estimate with or without estimating the randomly introduced missing values. Therefore, mean error of computing Brier scores over 1000 repeats; $\frac{\sum_{i=1}^{6}(\tilde{O}_i - \tilde{I}_i)^2}{1000}$, will be used for necessary comparisons between Brier scores with and without estimating missing values in the analysis. Here, we assume that distances from the original Brier scores are equally important irrespective of the sizes of Brier scores. It is also important to note that computing the squared distances between Brier scores ensures that sum of the distances will not be affected by the cancellation between positive and negative distances together.

3.5.1 Introducing missing values completely at random

First consider analysing the case of introducing missing predictions completely at random. Therefore, MCAR condition is directly satisfied and all the suggested missing value estimation methods of the analysis can be applied to estimate missing predictions. We begin with introducing 10% missing values directly into the overall set of predictions made by experts. This represents a context of which different experts have different number of missing predictions that can be considered more realistic in practice. Here, we do not consider the number of missing values for each participant separately in each repetition of the analysis. We let the number of missing values to be varied between repetitions for each participant and compute their average Brier scores with missing values to take into account the random variation of Brier scores due to the random introduction of 10% missing values into the overall set of predictions made by experts.

Table 3.2 indicates that on average there are only slight deviations between the original Brier scores and the Brier scores that are computed by ignoring 10% overall missing values of predictions introduced completely at random. It can also be seen that the

TABLE 3.2: Brier scores and estimated root mean squared errors (RMSE) of Brier scores

ParticipantId	Original Brier score	Estimated Brier score	RMSE estimate
51	0.14732	0.14762	0.01703
73	0.17228	0.17304	0.01450
76	0.16387	0.16367	0.01000
133	0.14272	0.14304	0.01140
251	0.14703	0.14693	0.01183
252	0.16968	0.16963	0.01095

estimated root mean squared errors of the computed Brier scores with missing values are reasonably small in this instance.

Now consider introducing 25% overall missing values of predictions completely at random. Figure 3.1 indicates that the estimated root mean squared errors of the computed Brier scores are increased by increasing the percentage of missing values from 10% to 25%. We also considered the case of introducing 10% and 25% percentages of missing values equally into the predictions made by each participant individually. Figures D.1 and D.2 in Appendix D show that on average most of the estimated root mean squared errors of Brier scores are higher for the individually introduced 10% and 25% missing values except some deviations due to potential random variation in the case of 25% missing values. Overall, it can be seen that the computed Brier scores tend to deviate more from the corresponding true values with increased percentages of missing values. Therefore, we consider important to estimate missing values of probability predictions for reliable comparisons between Brier scores in practice.

We now move into estimating the missing values of probability predictions for computing participants' Brier scores. First consider the mean imputation method. Suppose we consider grouping of probability predictions by participants and use the average of the predicted probabilities of questions by participants to estimate their missing probability predictions of questions. This provides the opportunity to study the potential of using the estimated participants' effects in estimating their missing probability predictions of questions. However, it is important to incorporate not only the corresponding participant's effect but also the corresponding question's effect in estimating the missing probability prediction of a question. Hence, it leads to consider the following mixed effect model including fixed mean parameters to incorporate

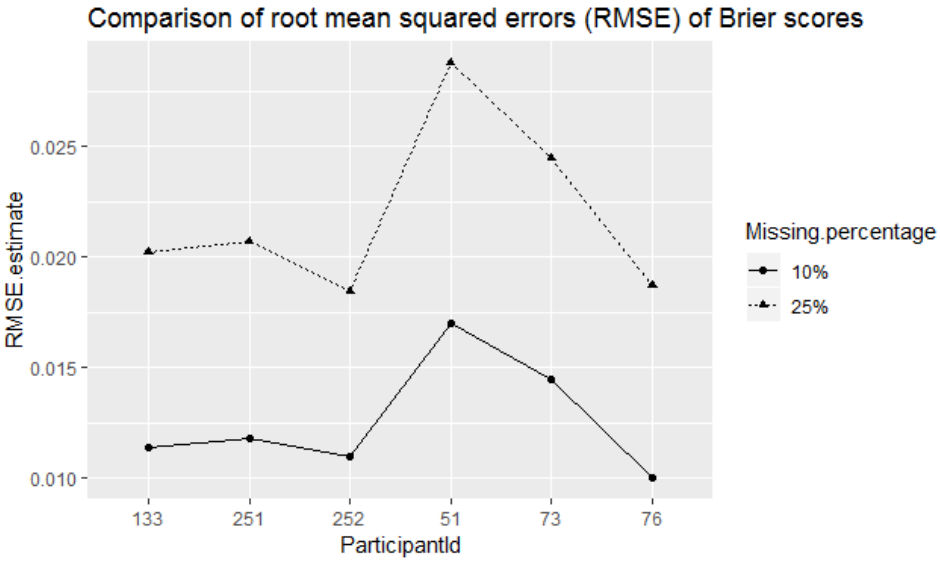


FIGURE 3.1: Comparison of root mean squared errors of computed participants' Brier scores with overall percentages of missing values introduced completely at random

participants' effects and random effects to incorporate questions' effects to estimate the missing best guess probability predictions of questions.

$$Bestguess_{ij} = \mu_i + \delta_j + \varepsilon_{ij}, \quad i = 1, 2, 3, ..., 6, \quad j = 1, 2, 3, ..., 31. \tag{3.5}$$

Note that $Bestguess_{ij}$ indicates the best guess probability prediction of the j^{th} question by the i^{th} participant, μ_i indicates the mean level probability predictions by the i^{th} participant, and δ_j indicates the effect of the j^{th} level of the questions. Furthermore, ε_{ij} represents the random error of probability predictions due to unknown sources of variation.

We also considered the possibility of using random effects to incorporate potential correlations between probability predictions due to the common grouping effects of participants into the analysis as discussed above. However, fitted mixed-effects models with random effects for groups and questions produced singular fits. Insufficient number of groups and inadequate number of participants within groups of the selected data may have caused to obtain models with singular fits. Therefore, regression imputation was carried out using the imputation model 3.5 with random effects only for questions. Note that regression imputation generates a random value each from the respective distributions defined by the combinations of relevant parameter effects

(fixed and/or random) of the suggested imputation model for imputing each of the missing values. Therefore, the underlying uncertainty of estimating parameters of the analysis model can be high due to performing only a single imputation for each of the missing values. Hence, multiple imputation was carried out to reduce the uncertainty of estimating parameters of the analysis model by obtaining the average of estimated parameters over 100 imputed data sets.

Now consider the figure 3.2 indicating the estimated mean errors of computing Brier scores with 95% confidence intervals for the mean imputed, regression imputed, multiple imputed and typically computed Brier scores ignoring missing values under 10% overall missing values of predictions introduced completely at random. Here, we assumed that estimated errors of computing Brier scores follow normal distributions with unknown means and variances under different methods of estimating missing values. Thus, we used the following well known interval estimate of population mean with unknown variance to quantify the uncertainty of the point estimates of mean errors in the analysis.

Suppose $\bar{x}$ denote the sample mean of the estimated errors of computing Brier scores under a given method of estimating missing values over 1000 repetitions of the analysis. Thus, we used the following well known interval estimate of population mean with unknown variance to quantify the uncertainty of the point estimates of mean errors in the analysis. Let us denote the $100(1-\alpha/2)$ percentile of the Student's t-distribution with 999 degrees of freedom as $t_{\alpha/2}$. For random samples of sufficiently large size, and with standard deviation s, the end points of the interval estimate at $(1-\alpha)$ confidence level is given as follows: $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{1000}}$.

According to figure 3.2, mean errors of computing Brier scores with multiple imputed missing values is lowest and statistically different to mean errors of computing Brier scores from other methods. It can also be seen that confidence intervals of other three methods overlaps each other implying that mean errors of computing Brier scores are not statistically different under 0.05 level of significance. Regression imputation method with a single imputation seems to have slightly higher sample mean error compared to the typical computation of Brier scores ignoring missing values. We acknowledge the fact there is an underlying uncertainty of generating random values

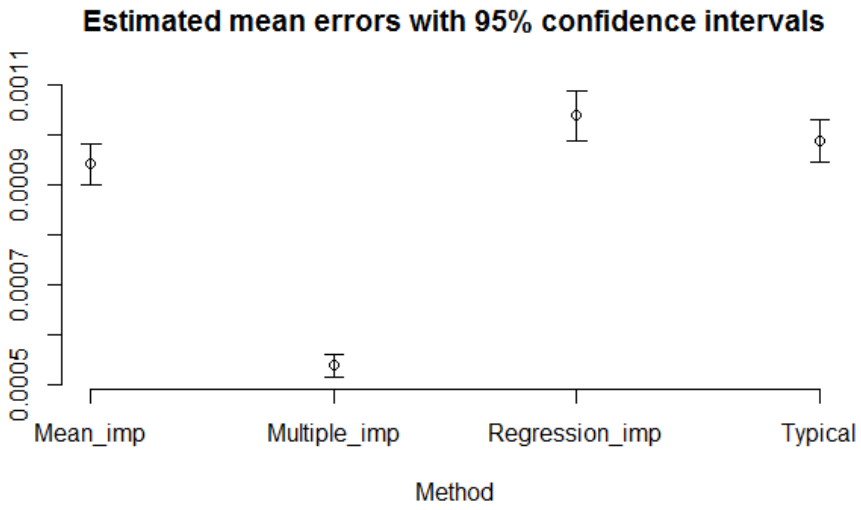


FIGURE 3.2: Estimated mean errors with 95% confidence intervals for computing Brier scores with 10% overall missing values introduced completely at random

for imputing missing values using the suggested mixed-effects model in regression imputation method. That is the reason for using multiply generated values to take this uncertainty into account in multiple imputation method. Also note that mean imputation method of using participants' effect to estimate missing values can reduce the sample mean error into some extent compared to that of the typically computed Brier scores ignoring missing values. However, it failed to achieve a significant difference. Figure 3.3 has almost similar interpretation of results for the case of 25% overall missing values of predictions introduced completely at random except for increased sample mean error for mean imputation method compared to typically computed Brier scores ignoring missing values.

Figures E.1 and E.2 in Appendix E show the results for computing Brier scores with 10% and 25% individually introduced missing values completely at random. We do not focus on the individual differences between mean errors of the mean imputation, regression imputation, and typical computation of missing values. We consider the fact that mean error of multiple imputed missing values is lowest and statistically different to mean errors from other methods. According to the results of the previous chapter, incorporating potential correlations between probability predictions due to the effects of common questions will improve the accuracy of estimated standard errors of participants' Brier scores in the analysis. Here, we use a mixed-effects model

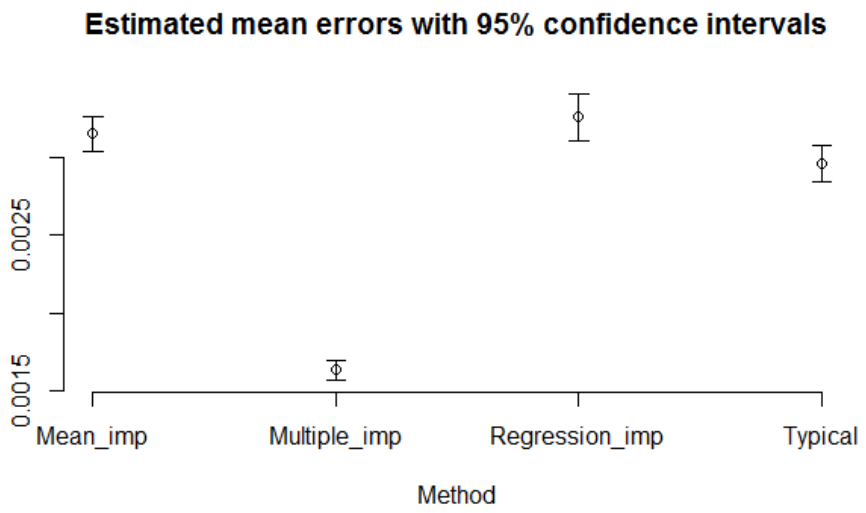


FIGURE 3.3: Estimated mean errors with 95% confidence intervals for computing Brier scores with 25% overall missing values introduced completely at random

with random effects for questions for both imputation and analysis models of the multiple imputation method. It was also discussed in section 3.3 that standard error estimates of parameters are also accounted for the between imputation uncertainty of the multiple imputation method. Therefore, it is reasonable to assume that standard error estimates of participants' Brier scores are accurate from the multiple imputation method. Thus, we report estimated participants' Brier scores together with their standard error estimates from the multiple imputation method in Appendix F for both overall and individually introduced missing values in this section.

3.5.2 Introducing missing values not at random

Now consider the case of introducing missing predictions not at random but depending on the actual values of themselves conditional on the levels of difficulty of questions in the analysis. Here, we randomly introduce missing values with probabilities proportional to the questions' Brier scores, representing the scenario that more difficult questions could be more likely to be missing. Therefore, the missing data mechanism would satisfy the not missing at random (NMAR) condition that is based on a non-testable assumption as discussed in section 3.3. Use of mixed-effects models in both regression and multiple imputation methods with random effects for questions

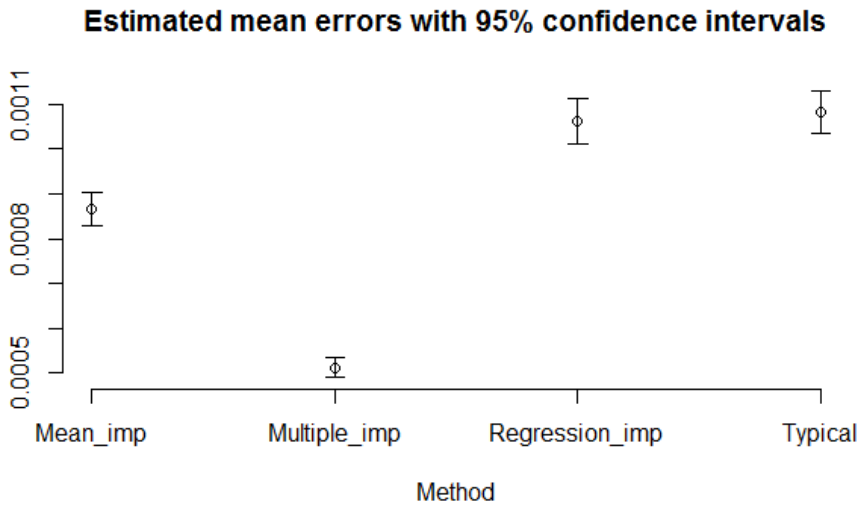


FIGURE 3.4: Estimated mean errors with 95% confidence intervals for computing Brier scores with 10% overall missing values introduced not at random

will incorporate the potential correlations between probability predictions caused by underlying differences between the levels of difficulty of questions into the imputation of missing values. Therefore, it will reduce the impact of ignoring the potential NMAR condition in the analysis. However, ignoring potential NMAR condition can have an impact on the results of mean imputation method in the analysis.

Similar analysis has been carried out as in the previous section for missing values introduced completely at random. Figure 3.4 shows that the mean errors of multiple imputed missing values are lowest and statistically different to the mean errors from other methods for computing Brier scores with 10% overall missing values. According to the figure 3.5, this result holds for the case of computing Brier scores with 25% overall missing values as well. Figures G.1 and G.2 in Appendix G also indicate similar results for the case of introducing 10% and 25% individual missing values, respectively. Therefore, we report the estimated participants' Brier scores together with their standard error estimates from the multiple imputation method in Appendix H as similar to the case of missing values introduced completely at random in the above section.

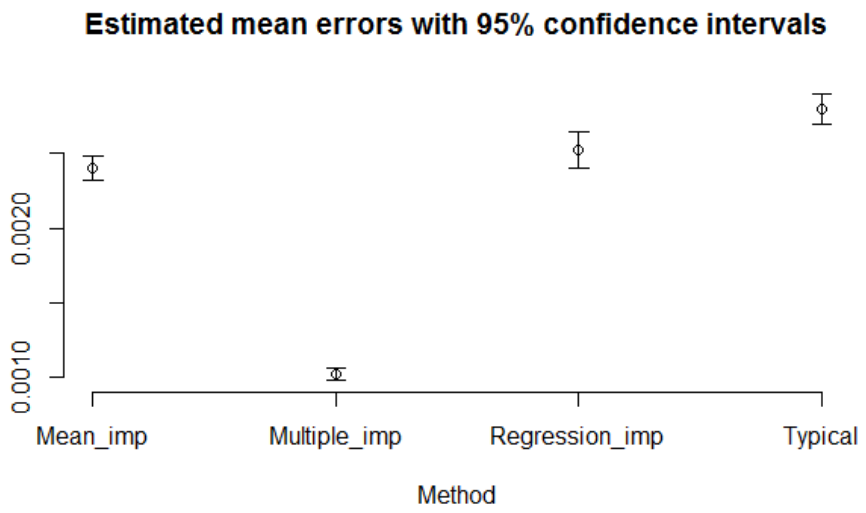


FIGURE 3.5: Estimated mean errors with 95% confidence intervals for computing Brier scores with 25% overall missing values introduced not at random

Section 3.6: Discussion

It is important to note that this simulation study was performed based on a specific data set. Therefore, there is a possibility that some specific characteristics of the data set may have caused multiple imputation to work well. We do not focus on performing a simulation study to theoretically prove the ability of the multiple imputation method with a given mixed-effects model to estimate missing probability predictions in computing experts' Brier scores over the other considered methods in some specific conditions. If so, it is required to perform a simulation study of which true effects of experts and the questions have to be chosen appropriately and simulate responses based on them. It will be followed by introducing missing values and estimating them using the suggested methods in the analysis. Instead, we consider this exercise just repeat existing simulation exercises that demonstrate the importance of imputation (Janssen et al., 2010; Ambler, Omar, and Royston, 2007; Demirtas, 2004) in general conditions, and the purpose of this chapter is to show how important it is to handle missing probability predictions correctly in the context of computing experts' Brier scores.

In general, regression imputation method with a single imputation was found to produce higher sample mean errors of computing Brier scores compared to the typically computed Brier scores ignoring missing values in the analysis. There is an underlying uncertainty of generating random values for imputing missing values using the suggested mixed-effects model in regression imputation method. It would have been the reason to observe higher errors of computing Brier scores compared to ignoring missing values in the typical computation of Brier scores. Averaging out this uncertainty using multiple generated random values for missing predictions in the multiple imputation method managed to produce lower errors of computing Brier scores compared to the typically computed Brier scores.

Applying a mixed-effects model with questions' effects as random effects will incorporate the potential correlations between probability predictions caused by the underlying differences between the levels of difficulty of questions into the imputation process of missing predictions. Therefore, whether the missing predictions are introduced completely at random or probability proportionately to questions' Brier scores (not at random), the corresponding questions' effects of missing questions will be incorporated into the process of estimating missing predictions by the fitted mixed-effects model. It is evident by not observing different patterns of comparisons between the mean errors of computing Brier scores under different methods of estimating missing values for both completely at random and not at random missing data mechanisms. More importantly, actual missing data mechanisms of observing missing predictions are unknown in practice. Therefore, what we can practically think about achieving is to estimate missing probability predictions with lowest possible errors for computing experts' Brier scores compared to to the typically computed Brier scores ignoring missing predictions. It is achieved for both missing data mechanisms within the considered percentages of missing values in the analysis.

Section 3.7: Conclusion

The focus of this analysis is to show that estimating missing values of probability predictions of questions can be useful to enhance the comparability of experts' Brier

scores to assess the prediction accuracy of experts. We considered the fact that experts' Brier scores that are computed without estimating missing values of probability predictions of questions are not adjusted for the effects of missing questions for some experts in a given analysis. Therefore, the comparisons between experts' Brier scores may not be accurate enough in such situations. We employed some selected missing value estimation methods that are suitable to the context of computing Brier scores and found that applying the multiple imputation method for estimating missing values using a mixed-effects model with questions' effects as random effects can be used to improve the computation of experts' Brier scores by adjusting for the missing probability predictions of questions within the considered scope of the analysis.

We acknowledge the fact that a base level may not be available for comparing the accuracy of missing values imputed Brier scores and the typically computed Brier scores ignoring missing values in a given analysis. However, the suggested multiple imputation method of estimating missing predictions uses all the available predictions to estimate missing predictions and compute a new set of experts' Brier scores with estimated missing predictions. This is more effective than throwing away some collected data to compute Brier scores using same sets of predictions for all the experts when there are missing predictions by experts. Therefore, we recommend to apply multiple imputation method using mixed-effects models with necessary random effects to bringing in any design-based correlated sources of predictions to compute Brier scores with estimated missing predictions. If there are considerable differences between the ranks of Brier scores with and without estimating missing predictions, then it is advisable to have an idea about the possibility of obtaining difference performances form experts in practice than what their original ranks suggest.

4. Testing experts' calibration I

Section 4.1: Introduction

This chapter focuses on assessing the properties of testing experts' calibration on eliciting credible intervals of unknown quantities. According to Speirs-Bridge et al. (2010), experts' calibration on eliciting credible intervals of unknown quantities is tested by the direct comparison of experts' hit rates; observed proportions of experts' elicited intervals that contain realized values of given quantities (McBride, Fidler, and Burgman, 2012), with the level of intended coverage probability of elicited intervals. Here, the hit rates are considered as fixed quantities disregarding their random variation in elicitation contexts. We emphasize the importance of considering the random variation of hit rates in testing experts' calibration as computed experts' hit rates can randomly vary from their true levels of coverage in short term and therefore the long-term average hit rates will better reflect the true levels of coverage of experts' elicited intervals. Thus, we consider the hit rates as random variables and apply the equivalence test of a single binomial proportion to compare the population means of experts' hit rates with the level of intended coverage probability of elicited intervals to test experts' calibration statistically.

It is evident from a conducted simulation study at some selected standard levels of intended coverage probabilities that the direct comparison of hit rates has a property of obtaining lower values of power to correctly identify well-calibrated experts when the true levels of coverage of experts' elicited intervals are taken equal to the corresponding levels of intended coverage probabilities of elicited intervals. Furthermore, there is a major problem of decreasing the power above with the increase of number of elicited

intervals. We show that the equivalence test of a single binomial proportion can be used to overcome these problems for large number of elicited intervals.

Section 4.2: Background of eliciting credible intervals

Expert opinion on quantities of interest is often elicited in the form of subjective credible intervals because they contain information about uncertainty that is not provided by point estimates (Speirs-Bridge et al., 2010). Thus, best guess estimates of quantities can be elicited together with credible intervals to reflect experts' uncertainty about the quantities. There are three types of elicited intervals that are created depending on the elicitation question format (Speirs-Bridge et al., 2010). The first type is the range question format where a lower limit and an upper limit are elicited to create an interval of a certain level of intended coverage probability. The second type is the 3-point question format where a lower limit, an upper limit, and a best guess are elicited for each quantity. Note that two elicited limits create an interval as above and the best guess is considered useful to reduce common overconfidence in eliciting intervals (Plous, 1993). This follows the knowledge sampling theory due to Klayman, Juslin, and Winman (2006) that if we sample our knowledge base more times to create an estimate, then the estimate is less likely to be overconfident. The third type is the 4-step format where the same procedure of the 3-point format is followed except the level of intended coverage probability is not assigned by the experimenter. In the fourth step, experts rate their anticipated level of coverage probability of the elicited interval. This format can be considered as an improved method to reduce overconfidence (Speirs-Bridge et al., 2010). We do not focus on explaining how the 3-point and 4-step formats reduce overconfidence in eliciting intervals here (refer Soll and Klayman (2004) and Teigen and Jørgensen (2005) for useful explanations on this subject matter).

If we consider how the experts' calibration is tested on interval judgements, the current approach is to directly compare the computed proportions of intervals that contain true realized values of quantities with the level of intended coverage probability of elicited intervals for experts. The computed proportions of intervals that contain

true values can be referred as "hit rates" (McBride, Fidler, and Burgman, 2012). According to Speirs-Bridge et al. (2010), experts' calibration can be tested using the hit rates as follows. The authors referred this approach of testing calibration as " hit-and-miss calibration". Suppose 10 intervals at 80% intended coverage probability have been elicited from an expert. Then, the expert can be considered perfectly calibrated if 8 out of 10 intervals contain true values. The expert should be considered 20% overconfident if only 6 out of 10 intervals contain true values. Following the same argument of " hit-and-miss calibration", experts can also be considered to be underconfident. Suppose 9 out of 10 intervals that have been elicited from an expert contain true values in the above described context. Then, the expert should be considered 10% underconfident.

The range question format and the 3-point format of eliciting intervals discussed above produce intervals at a given fixed level of coverage probability. Therefore, experts' calibration can be tested by comparing the computed experts' hit rates directly with the intended fixed level of coverage probability. The 4-step format produces intervals at varied levels of coverage probabilities by experts. Therefore, testing experts' calibration has to be performed after transforming the elicited intervals into a standard level of coverage probability to allow comparisons between experts at a common level of coverage probability. According to Hemming et al. (2018), intervals are typically transformed into 80% or 90% standard levels of coverage probabilities.

One method of transformation is to fit a suitable distribution to the elicited points (lower limit, upper limit, and the best guess) of a quantity of interest by an expert. It is followed by deriving the credible interval with the level of intended coverage probability from the fitted distribution. This process will be carried out for all the intervals that are necessary to be transformed for all the experts. It should be noted that identifying a suitable distribution to be fitted to a quantity using the limited quantiles that have been elicited from an expert can be practically difficult (O'Hagan et al., 2006; Speirs-Bridge et al., 2010; McBride, 2013). The linear extrapolation is also used to transform intervals into a standard level of intended coverage probability. In this method, the elicited best guess (B), lower bound (L) and upper bound (U) of an interval, and the level of coverage probability assigned by an expert (C) are used

to derive a credible interval (L_s, U_s) of a given standard level of intended coverage probability (S) as $L_s = B - ((B - L) \times (S/C))$ and $U_s = B + ((U - B) \times (S/C))$, respectively. If the adjusted intervals fall outside of reasonable bounds (such as [0, 1] for probabilities), then distributions will be truncated at their extremes (Hemming et al., 2018). Here, we do not focus on how accurately these transformations can be made in practice. We focus on the end result that the intervals are transformed into a standard level of coverage probability to test experts' calibration using experts' hit rates at a common level of coverage probability.

It is important to note that experts' hit rates are assumed as fixed quantities in testing experts' calibration on interval judgements. In general, an expert can be considered well-calibrated if "over the long run, for all predictions assigned a given probability, the proportion that are true equals the probability assigned" (Lichtenstein and Fischhoff, 1980). Therefore, experts' calibration on eliciting intervals should be tested based on how close the computed experts' hit rates to the level of intended coverage probability of elicited intervals in the long run. It underlines the fact that computed experts' hit rates can randomly vary from their true levels of coverage in short term and therefore long-term average hit rates will better reflect their true levels of coverage. Thus, we consider experts' hit rates as random variables and compare the population means of experts' hit rates with the level of intended coverage probability of elicited intervals to test experts' calibration on interval judgements statistically.

Section 4.3: Statistical testing of experts' calibration

It follows from the above discussion that considering the random variation of experts' hit rates is important in testing experts' calibration on eliciting credible intervals of unknown quantities. Here, we consider a theoretical population of hit rates that results from eliciting a certain fixed number of intervals repeatedly at a certain fixed level of coverage probability under similar experimental conditions by a particular expert. Hence, one of the computed hit rates from a given set of questions can be considered as a single realization from that theoretical population of the considered expert. The statistical inference of interest here is to test the equality of the population mean of

hit rates and the level of intended coverage probability of elicited intervals to assess expert's calibration.

If we consider the potential statistical tests to be used in this context, the obvious first choice would be to consider the one proportion equality test. If we apply the equality test, the null hypothesis will indicate that the population mean hit rate is equal to the level of intended coverage probability of elicited intervals and the alternative hypothesis will indicate that they are not equal. In this setting, the null hypothesis should not be rejected to conclude that the considered expert is well-calibrated. However, it does not necessarily mean that the expert is well-calibrated. It only implies that the data do not provide sufficient evidence to reject the null hypothesis that the expert is well-calibrated. Thus, we consider employing a different form of a hypothesis test that seems more applicable to the context.

Now consider the equivalence test of a single binomial proportion. In general, there are two null hypotheses and a single alternative hypothesis in the equivalence test of a single binomial proportion (refer the section 4.3.1 for details). The alternative hypothesis indicates the equivalence of the two population proportions that are compared and the two null hypotheses indicate the non-equivalence of them. It is required to reject both the null hypotheses to declare equivalence. Otherwise, there is not enough evidence to declare the equivalence from the test. If we apply this test in our context, the alternative hypothesis will indicate that the expert is well-calibrated (population mean of hit rates is equivalent to the level of intended coverage probability of elicited intervals) and the two null hypotheses will indicate other way around. The considered expert can be declared well-calibrated if both the null hypotheses are rejected. Otherwise, it should be concluded that there is not enough evidence to reject the claim that the expert is not well-calibrated.

If equality and the equivalence tests are compared, the null hypothesis of the equality test should not be rejected and both the null hypotheses of the equivalence test should be rejected to declare a given expert as well-calibrated. When the null hypothesis of a test is not rejected, the type II error (probability of not rejecting the false null hypothesis) of the test can occur and the size of the error is generally unknown. However, when the null hypothesis of a test is rejected, the maximum possible type I

error (probability of rejecting the true null hypothesis) that could have been occurred is known (level of significance of the test). Generally, a small value (e.g. 0.05) is set for the type I error when developing a test. Therefore, it is statistically more reliable to obtain a decision by rejecting a null hypothesis or a set of null hypotheses.

If we consider the context of concluding an expert is well-calibrated, it is advisable to assume an expert is not well-calibrated unless the data have sufficient evidence to disprove the assumption. Here, we assume that the error of concluding a not well-calibrated expert as well-calibrated is more serious than the error of concluding a well-calibrated expert as not well-calibrated. Definition of the equivalence test also ensures that the true levels of coverage of elicited intervals of the identified experts as well-calibrated remain within a known acceptable margin of deviation from the level of intended coverage probability of elicited intervals as described below. Therefore, we use these facts to justify our intention to apply the equivalence test of a single binomial proportion over the one proportion equality test to test experts' calibration in this analysis.

4.3.1 Exact version of the equivalence test of a single binomial proportion

We consider applying the equivalence test of a single binomial proportion to test experts' calibration on eliciting intervals under varied number of elicited intervals. Therefore, it is important to apply the exact version of the test in place of the large sample approximated test to keep the flexibility of considering both small and large number of elicited intervals. Following details about the binomial distribution and the exact version of the equivalence test of a single binomial proportion are obtained from Wellek (2010, chap. 4). The binomial distribution can be considered as a sum $T(\tilde{X}) = \sum_{i=1}^{m} \tilde{X}_i$ of m mutually independent Bernoulli random variables; $\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_m$, with two outcomes, say "success" and "failure" with probability $\tilde{p}$ and $1 - \tilde{p}$, respectively. The probability mass function of the distribution of T is:

$$f(T = t; m, \tilde{p}) = \binom{m}{t} \tilde{p}^t (1 - \tilde{p})^{m-t}, t \in \{0, 1, 2, ..., m\}. \tag{4.1}$$

Suppose that an unknown population proportion of success $\tilde{P}$ is required to be statistically tested for equivalence with a reference value $\tilde{P}_r$. The statistics of interest in defining the equivalence test of a single binomial proportion is the number of successes T out of a certain m number of Bernoulli trials. Let us assume that the equivalence can be claimed if $\tilde{P}$ remains between $\tilde{P}_1 = \tilde{P}_r - \epsilon_1$ and $\tilde{P}_2 = \tilde{P}_r + \epsilon_2$, where an acceptable margin of deviation around the reference value is allowed in the test. Wellek (2010, chap. 4) explained that $\tilde{P}_1$ and $\tilde{P}_2$ can be symmetrized around $\tilde{P}_r$ by taking ϵ_1 and ϵ_2 are equal to some specified value ϵ. This approach can generally be followed unless specific values can be chosen for ϵ_1 and ϵ_2 depending on the context of interest.

It was mentioned above that an equivalence test includes two null hypotheses and a single alternative hypothesis of which equivalence can be claimed by rejecting both the null hypotheses. Furthermore, allowing an acceptable margin of deviation ϵ around the reference value $\tilde{P}_r$ leads to claim the equivalence if $\tilde{P}$ remains between $\tilde{P}_1 = \tilde{P}_r - \epsilon$ and $\tilde{P}_2 = \tilde{P}_r + \epsilon$. Therefore, the alternative hypothesis should be defined as $\tilde{P}_1 < \tilde{P} < \tilde{P}_2$ and the two null hypotheses should be defined as $0 < \tilde{P} \leq \tilde{P}_1$ and $\tilde{P}_2 \leq \tilde{P} < 1$. Hence, the equivalence test of a single binomial proportion can be stated as follows.

$$\begin{aligned} \mathrm{H}_0 &: 0 < \tilde{P} \leq \tilde{P}_1 \text{ or } \tilde{P}_2 \leq \tilde{P} < 1 \\ H_1 &: \tilde{P}_1 < \tilde{P} < \tilde{P}_2, \text{ where } (0 < \tilde{P}_1 < \tilde{P}_2 < 1). \end{aligned} \tag{4.2}$$

There exists a uniformly most powerful (UMP) level α test that can be defined by the following set of rules:

1. Rejection of H_0 for $C^1_\alpha(m; \tilde{P}_1, \tilde{P}_2) < T < \mathrm{C}^2_\alpha(m; \tilde{P}_1, \tilde{P}_2)$
2. Rejection with probability $\gamma^1_\alpha(m; \tilde{P}_1, \tilde{P}_2)$ for $T = C^1_\alpha(m; \tilde{P}_1, \tilde{P}_2)$

3. Rejection with probability $\gamma_\alpha^2(m;\tilde{P}_1,\tilde{P}_2)$ for $T = C_\alpha^2(m;\tilde{P}_1,\tilde{P}_2)$

4. Acceptance for $T < C_\alpha^1(m;\tilde{P}_1,\tilde{P}_2)$ or $T > C_\alpha^2(m;\tilde{P}_1,\tilde{P}_2)$

We use the abbreviations C_1, C_2, γ_1 and γ_2 for $C_\alpha^1(m;\tilde{P}_1,\tilde{P}_2), C_\alpha^2(m;\tilde{P}_1,\tilde{P}_2)$, $\gamma_\alpha^1(m;\tilde{P}_1,\tilde{P}_2)$ and $\gamma_\alpha^2(m;\tilde{P}_1,\tilde{P}_2)$, respectively. Following equation can be used to derive the constants; C_1, C_2, γ_1 and γ_2 using the probability mass function $f(T = t; m, \tilde{P})$ of the binomial random variable T (equation 4.1).

$$\sum_{t=C_1+1}^{C_2-1} f(T=t;m,\tilde{P}_1) + \sum_{\nu=1}^{2} \gamma_\nu\, f(T=C_\nu;m,\tilde{P}_1) = \alpha = \sum_{t=C_1+1}^{C_2-1} f(T=t;m,\tilde{P}_2) + \sum_{\nu=1}^{2} \gamma_\nu\, f(T=C_\nu;m,\tilde{P}_2); \quad 0 \le C_1 \le C_2 \le m, \quad 0 \le \gamma_1, \gamma_2 < 1. \tag{4.3}$$

Following steps to derive a solution; C_1, C_2, γ_1 and γ_2 are due to Wellek (2010, chap. 4).

1. Select an initial value C_1^0 for the lower bound of the rejection region knowing that it is greater or equal to the correct value C_1.

2. Keeping C_1^0 fixed, find the largest integer $C_2^0 > C_1^0$ such that the probability of observing T to take on its value in the closed interval $[C_1^0 + 1, C_2^0 - 1]$ does not exceed α, neither for $\tilde{P} = \tilde{P}_1$ nor for $\tilde{P} = \tilde{P}_2$.

3. Consider equation 4.3 as a system of linear equations in the two unknowns γ_1 and γ_2 and compute its solution γ_1^0, γ_2^0.

4. Test whether the condition $0 \le \gamma_1^0, \gamma_2^0 < 1$ is satisfied. If so, a solution of the full system in equation 4.3 is found and can be given as $(C_1, C_2, \gamma_1, \gamma_2) = (C_1^0, C_2^0, \gamma_1^0, \gamma_2^0)$. If not, reduce C_1^0 by 1 and repeat steps (2) and (3).

The above set of rules to define the level α test indicate that the equivalence test should be defined as a randomized test with γ_1 and γ_2 rejection probabilities at $T = C_1$ and $T = C_2$. We implement the equivalence test to test the calibration of a given expert at a certain level of intended coverage probability as follows:

1. Consider the level of intended coverage probability as the reference value $\tilde{P}_r$.

2. Define $\tilde{P}_1 = \tilde{P}_r - \epsilon$ and $\tilde{P}_2 = \tilde{P}_r + \epsilon$ based on the selected value of ϵ.

3. Compute the limits of the rejection region C_1 and C_2 at a given number of elicited intervals.

4. Observe the number of intervals containing true values (m_{int}) as a random variable for the expert.

5. If $m_{int} < C_1$ OR $m_{int} > C_2$, then do not reject the null hypotheses.

6. If $m_{int} > C_1$ AND $m_{int} < C_2$, then reject the null hypotheses and conclude the equivalence indicting that the expert is well-calibrated at the intended coverage probability $\tilde{P}_r$.

7. If $m_{int} = C_1$, then generate a uniform random number u. If $u < \gamma_1$, then reject the null hypotheses and conclude the equivalence. Otherwise do not reject the null hypotheses.

8. If $m_{int} = C_2$, then generate a uniform random number u. If $u < \gamma_2$, then reject the null hypotheses and conclude the equivalence. Otherwise do not reject the null hypotheses.

Section 4.4: An important statistical problem

Suppose we apply the direct comparison of hit rates to test experts' calibration on eliciting 80% and 90% credible intervals of quantities. Here, we assume that the true levels of coverage of experts' elicited intervals of quantities are equal to the corresponding levels of intended coverage probabilities of elicited credible intervals. Therefore, we focus on investigating the properties of the power of the test to correctly identify well-calibrated experts on eliciting 80% and 90% credible intervals. Figure 4.1 plots the values of power of the test above at some selected number of elicited intervals (m); 10, 20, 30, 40, 50, 80, 100, 150, 200, and 250. Note that binomial probabilities are used to compute the values of power at different combinations of true levels of coverage and number of elicited intervals. For example, the binomial probability of

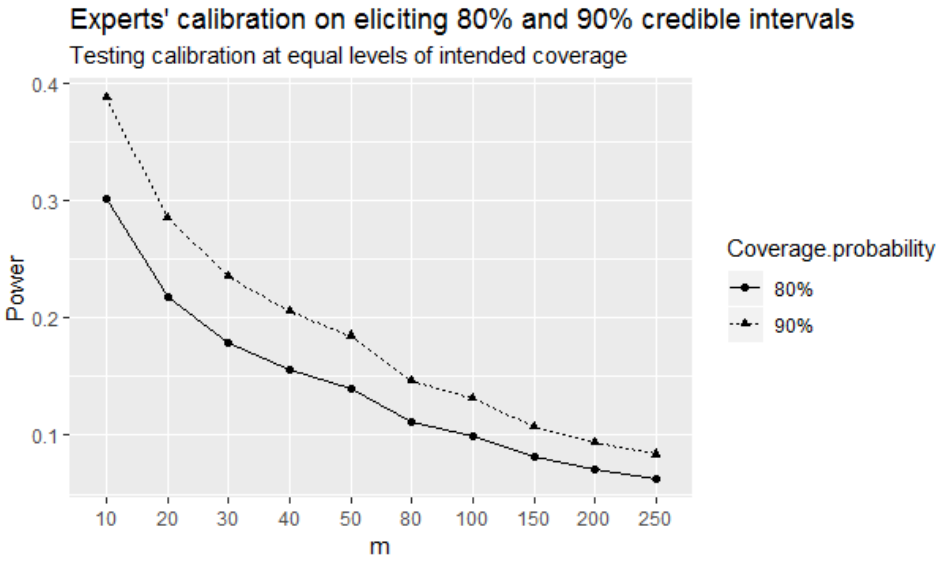


FIGURE 4.1: The power of the direct comparison of hit rates to correctly identify well-calibrated experts

obtaining 8 successes out of 10 trials at 0.8 probability of success gives the power of correctly identifying 80% well-calibrated experts on eliciting 10 intervals at 80% true level of coverage.

Here, we analyse a theoretical context of which the true levels of coverage of experts' elicited intervals are assumed equal to the corresponding levels of intended coverage probabilities of elicited credible intervals. Therefore, it is reasonable to expect to obtain higher values of power to correctly identify well-calibrated experts at each level of coverage probability. However, the observed values of power are considerably low at each level of coverage probability. More importantly, the values of power tend to decrease as the number of elicited intervals increases. This is a contradictory result from a statistical point of view as the power of the test to correctly identify well-calibrated experts can be expected to increase with increased number of elicited intervals.

The above result of obtaining lower values of power to correctly identifying well-calibrated experts over increased number of elicited intervals can be interpreted as follows. Consider the case of computing the power of the test to correctly identifying 80% well-calibrated experts at 80% true level of coverage of elicited intervals. Note that the binomial probability of obtaining 8 successes out of 10 trials at 0.8 probability of success is higher than the corresponding binomial probability of obtaining 16 successes

out of 20 trials at 0.8 probability of success. If the number of elements of a sample space increases, then the probability of observing a single element decreases. The same principle applies here. Therefore, the power of the test tends to decrease as the number of elicited intervals increases.

It is also important to note that the values of power to correctly identify well-calibrated experts are higher on eliciting 90% credible intervals compared to eliciting 80% credible intervals. This result holds in general as the computation of binomial probabilities ensures that the probability of obtaining 0.9 proportion of successes under 0.9 success probability is higher than the probability of obtaining 0.8 proportion of successes under 0.8 success probability for a given number of trials.

Section 4.5: Methodology of the study

It follows from the discussion above that the direct comparison of hit rates has a property of obtaining lower values of power to correctly identify 80% and 90% well-calibrated experts when the true levels of coverage of experts' elicited intervals are taken equal to the corresponding 80% and 90% levels of intended coverage probabilities of elicited credible intervals. Furthermore, there is a major problem of decreasing the power over the increase of number of elicited intervals. Therefore, we perform a simulation study here to investigate whether addressing the random variation of hit rates using the equivalence test of a single binomial proportion can be used to overcome the above identified problems of the direct comparison of hit rates. For convenience, we refer the equivalence test of a single binomial proportion just as the equivalence test and the direct comparison of hit rates as the direct test hereafter.

When simulating data, the outcome of each elicited interval; whether the realized true value of the quantity of interest is included or not included in the interval, will be considered as a Bernoulli outcome with a fixed probability of success that is equal to the assumed true level of coverage of experts' elicited intervals. Therefore, the total number of elicited intervals that include true values of quantities can be considered as a binomial outcome with the same fixed probability of success and the number of trials equals to the total number of elicited intervals of quantities. Hence, random

observations from binomial distributions will be used accordingly to generate total number of intervals that include true values of quantities for experts in the analysis. Here, we focus on comparing the power of the direct and equivalence tests to correctly identify well-calibrated experts on eliciting credible intervals of quantities with 80% and 90% coverage probabilities. Therefore, the data will be generated at 80% and 90% true levels of coverage of experts' elicited intervals for the analysis at the number of elicited intervals (m) considered in the section 4.4 above.

The above discussed random generation of total number of elicited intervals containing true values will be repeated 25 times at each combination of total number of elicited intervals and the assumed true levels of coverage of experts' elicited intervals mentioned above. Here, the number of repeats can conceptually be considered as the number of experts in the analysis. It will be followed by computing the proportion of experts who are correctly identified as well-calibrated on eliciting 80% and 90% credible intervals respectively from the direct and equivalence tests. Furthermore, the above process will be repeated 100000 times to obtain the average proportions of experts who are correctly identified as well-calibrated at each combination mentioned above.

We conceptually considered 25 experts for the computation above. However, it would be more useful to carry out the analysis with probabilities of correctly identifying well-calibrated experts without considering the number of experts as it is not straight forward to assume a reasonable number of experts for the analysis. It can be assumed that the average proportions obtained from the direct and equivalence tests above represent the probabilities of correctly identifying well-calibrated experts as they are computed from a large number of repeats. Therefore, the analysis can be carried out using the computed probabilities without considering the number of experts to reduce the scope of the study.

The significance level of the test α, and the acceptable margin of deviation ϵ around the reference value $\tilde{P}_r$ of the equivalence test will be taken as 0.05 in the analysis. Here, we assume that 0.05 is a commonly used acceptable level of significance for statistical analyses and 0.05 deviation of true level of coverage of elicited intervals from the considered levels of intended coverage probabilities of credible intervals is reasonable enough to consider a given expert as well-calibrated.

Section 4.6: Testing experts' calibration on eliciting 90% credible intervals

We first analyse the power of the direct and equivalence tests to correctly identify well-calibrated experts on eliciting 90% credible intervals by taking true level of coverage of experts' elicited intervals equals to 90%. Following critical regions (C_1, C_2) of the equivalence test given in table 4.1 to test experts' calibration on eliciting 90% credible intervals are obtained using the R program given in Appendix I due to Wellek (2010).

TABLE 4.1: Critical regions of the equivalence test

m	C1	C2
10	9	10
20	18	19
30	27	28
40	36	37
50	45	46
80	72	73
100	90	92
150	134	138
200	178	185
250	222	232

Figure 4.2 plots the power of the direct and equivalence tests to correctly identify 90% well-calibrated experts at 90% true level of coverage of experts' elicited intervals. It should be noted that the values of power of the direct test are not only decreasing when the number of elicited intervals are increasing but also they are considerably low at large number of elicited intervals. The values of power of the equivalence test tend to increase as intuitively expected with the increase of number of elicited intervals and they are higher than the corresponding values of the direct test for 80 or more intervals.

4.6.1 Assessing the properties of the tests at different true levels of coverage of intervals

The purpose of the analysis of this section is to assess the properties of the direct and equivalence tests in testing experts' calibration on eliciting 90% credible intervals when the true levels of coverage of elicited intervals are different to 90%. Here, we

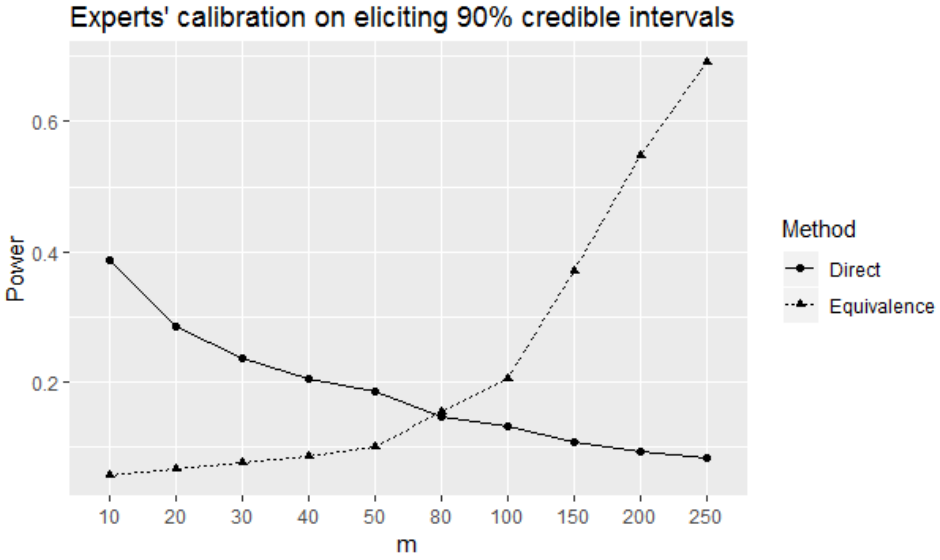


FIGURE 4.2: The power of the direct and equivalence tests to correctly identify 90% well-calibrated experts at 90% true level of coverage of elicited intervals

assess the outcomes of the direct and equivalence tests in testing experts' calibration on eliciting 90% credible intervals at true levels of coverage of elicited intervals that are less than 90% under the assumption that experts are usually overconfident in expert elicitation context. We restrict our attention to 100 or more intervals where the equivalence test was found to have comparatively higher values of power than the direct test to correctly identify 90% well-calibrated experts. Figure 4.3 plots the probabilities of identifying the experts as 90% well-calibrated at true levels of coverage of elicited intervals remain between 85% and 89% from the direct and equivalence tests.

It was discussed above that comparatively higher values of power to correctly identify 90% well-calibrated experts at 90% true level of coverage of elicited intervals can be obtained using the equivalence test compared to the direct test for 100 or more intervals. According to the figure 4.3, it is achieved at the expense of receiving comparatively higher probabilities of incorrectly identifying the experts as 90% well-calibrated at true levels of coverage remain between 85% and 89% from the test. Furthermore, these probabilities increase with the increase of number of elicited intervals. The alternative hypothesis of the equivalence test $\tilde{P}_1 < \tilde{P} < \tilde{P}_2$ to declare equivalence (to declare an expert as well-calibrated) at 90% coverage probability with the acceptable margin of deviation $\epsilon = 0.05$ around the reference value $\tilde{P} = 0.9$ contains values

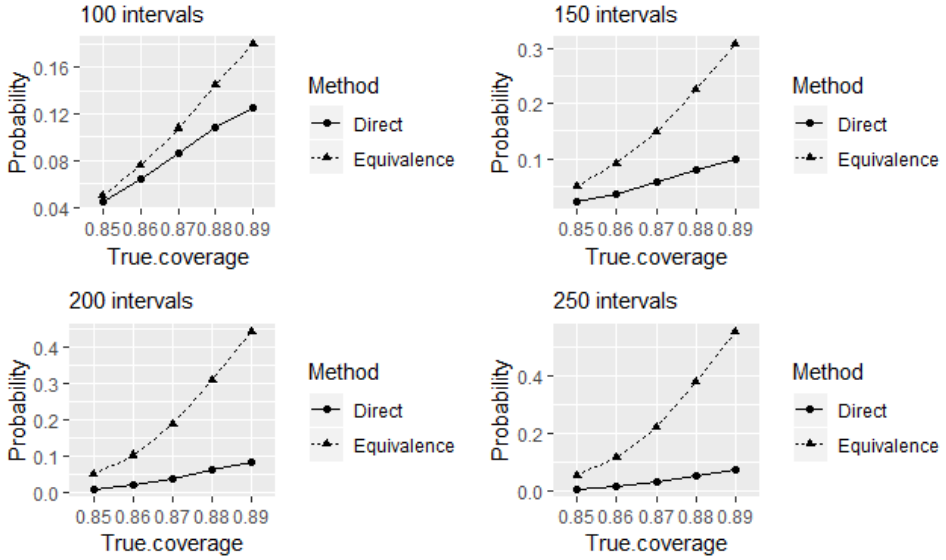


FIGURE 4.3: The probabilities of the direct and equivalence tests to identify the experts as 90% well-calibrated when true levels of coverage of elicited intervals are less than 90%

between $\tilde{P}_1 = 0.85$ to $\tilde{P}_2 = 0.95$ (refer the section 4.3.1 for details). Therefore, from the equivalence test point of view, the probabilities of identifying experts as 90% well-calibrated at different true levels of coverage of elicited intervals within this range can be considered as the values of power (the probabilities of rejecting the null hypotheses when they are false) of the test at corresponding values in the rejection region of the test. Therefore, receiving comparatively higher probabilities of incorrectly identifying the experts as 90% well-calibrated at true levels of coverage remain between 85% and 89% here happens due to a characteristic of the equivalence test.

Now focus on the type I error probabilities (the probabilities of incorrectly accepting the alternative hypothesis when one of the null hypotheses is true) of the equivalence test. Observe that type I error probabilities of incorrectly identifying the experts as 90% well-calibrated are approximately equal to the nominal value 0.05 at 85% true level of coverage of intervals which is on the border of the rejection region of the test. Direct comparison of observed hit rates of experts' elicited intervals with the level of intended coverage probability of credible intervals without considering the random variation of hit rates cannot be considered as testing experts' calibration statistically. However, considering the direct test as a special form of the equivalence test of which the alternative hypothesis includes only a single value of $\tilde{P} = 0.9$ can ease the interpretation of results of the direct test in figure 4.3 above.

If we consider the direct test as a special case of the equivalence test, the probabilities given in figure 4.3 of the direct test can be considered as type I error probabilities of incorrectly identifying the experts as 90% well-calibrated at the corresponding true levels of coverage of elicited intervals. It is a good property of the direct test to have lower type I error probabilities of incorrectly identifying the experts as 90% well-calibrated even at true levels of coverage of elicited intervals that are slightly less than 90%. It can be observed that the corresponding probabilities have reached the maximum of approximately 0.12 for the range of true levels of coverage and the number of elicited intervals considered. More importantly, these probabilities tend to decrease as expected when the number of elicited intervals increases. However, this property of decreasing probabilities over the increase of number of elicited intervals is applicable to the probabilities of correctly identifying 90% well-calibrated experts at the 90% true level of coverage of elicited intervals as well. We can overcome this problem using the equivalence test if we can accept observing higher probabilities of incorrectly identifying the experts as 90% well-calibrated at slightly lower true levels of coverage of elicited intervals as observed above.

The equivalence test needs more elicited intervals to obtain reasonably higher value of power to correctly identify well-calibrated experts even though it can overcome the above problem of the direct test. This is a drawback of applying the equivalence test. However, the application of the direct test cannot be accepted with the existence of the above problem. If we neglect that fact and apply the direct test, then it is required to reduce the number of elicited intervals to obtain higher probabilities of correctly identifying the experts as well-calibrated as mentioned above. It cannot statistically be accepted to encourage to perform an analysis with reduced number of observations to obtain better performance of the test, unless the additional observations do not add any valid information. It is not the case here. The properties of the test suggest to reduce the number of observations for the analysis.

It follows from above that the equivalence test needs more elicited intervals to obtain reasonably higher values of power to correctly identify well-calibrated experts. Therefore, one might have an interest to apply the direct test for small number of intervals. Let us further review the outcomes of the direct test at small number of elicited

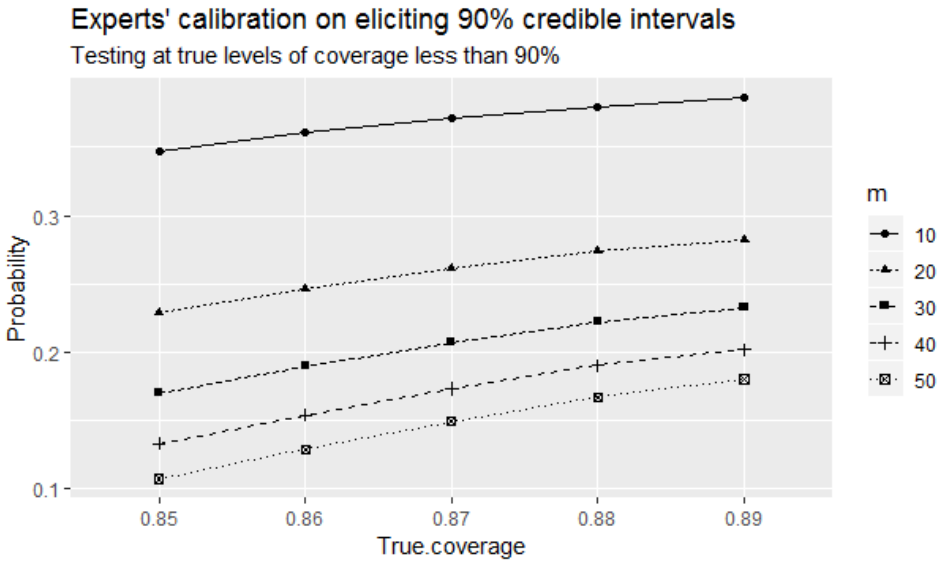


FIGURE 4.4: The probabilities of the direct test to identify the experts as 90% well-calibrated when true levels of coverage of elicited intervals are less than 90% for small number of elicited intervals

intervals. It is important to note from the figure 4.2 that the corresponding values of power to correctly identify well-calibrated experts are not large enough at small number of elicited intervals. In addition to that, figure 4.4 indicates another problem of the direct test to increase the type I error probabilities of incorrectly identifying the experts as 90% well-calibrated at the considered true levels of coverage of elicited intervals that are less than 90% with decreased number of elicited intervals. Observe that these probabilities are higher than the corresponding type I error probabilities at large number of elicited intervals shown in figure 4.3 above.

Following the discussion above, we need to think about whether it is sensible to use small number of elicited intervals to test experts' calibration with not large enough probabilities of correctly identifying well-calibrated experts and also with comparatively large type I error probabilities of incorrectly identifying experts as well-calibrated. Figure 4.5 with additionally included number of elicited intervals reveals another potential problem of the direct test of giving zero power when the required hit rate that needs to be equal to the level of intended coverage probability of elicited intervals cannot be observed from the number of intervals elicited. For example, an exact 90% hit rate cannot be observed from 25, 75, 125, and 175 elicited intervals to declare experts as well-calibrated at 90% level of coverage probability. It was the reason for using 80 elicited intervals instead of more appealing 75 intervals

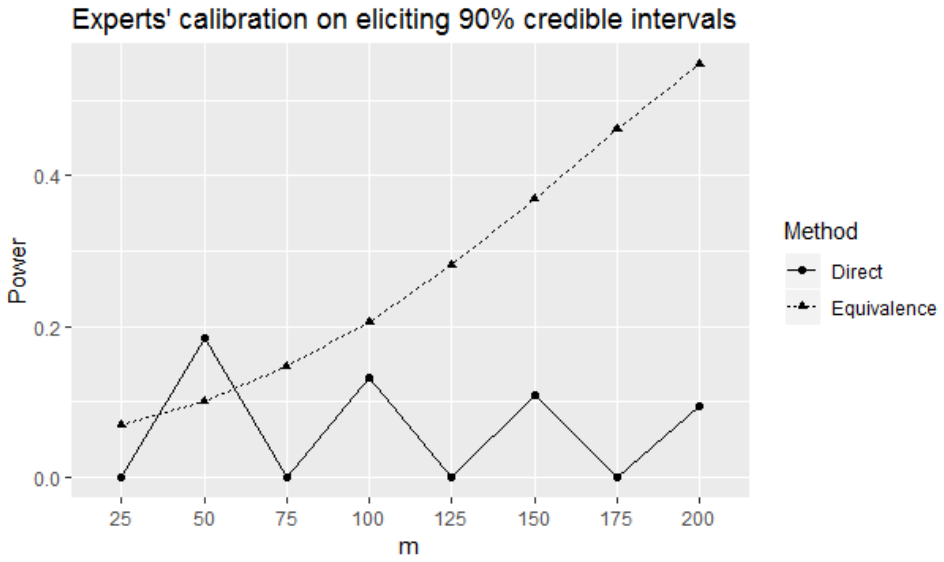


FIGURE 4.5: The power of the direct and randomized equivalence tests to correctly identify 90% well-calibrated experts at 90% true level of coverage of elicited intervals with additionally included number of elicited intervals

as the middle value between 50 and 100 intervals in this analysis. Therefore, it can be suggested considering these issues to apply the equivalence test with large number of elicited intervals to obtain higher probabilities of correctly identifying 90% well-calibrated experts by tolerating the higher probabilities of incorrectly identifying the experts as 90% well-calibrated at true levels of coverage of elicited intervals that are slightly less than 90%.

Section 4.7: Testing experts' calibration on eliciting 80% credible intervals

The above mentioned patterns of results are also common in testing for experts' calibration on eliciting 80% credible intervals of quantities. The observed values of power to correctly identify well-calibrated experts on eliciting 80% credible intervals were lower than the corresponding values on eliciting 90% credible intervals for the considered range of number of elicited intervals for both direct and equivalence tests. Therefore, we increased the number of elicited intervals further to obtain reasonably higher values of power from the tests. Figure 4.6 plots the values of power of the direct and equivalence tests for the extended range of elicited intervals. It also indicates

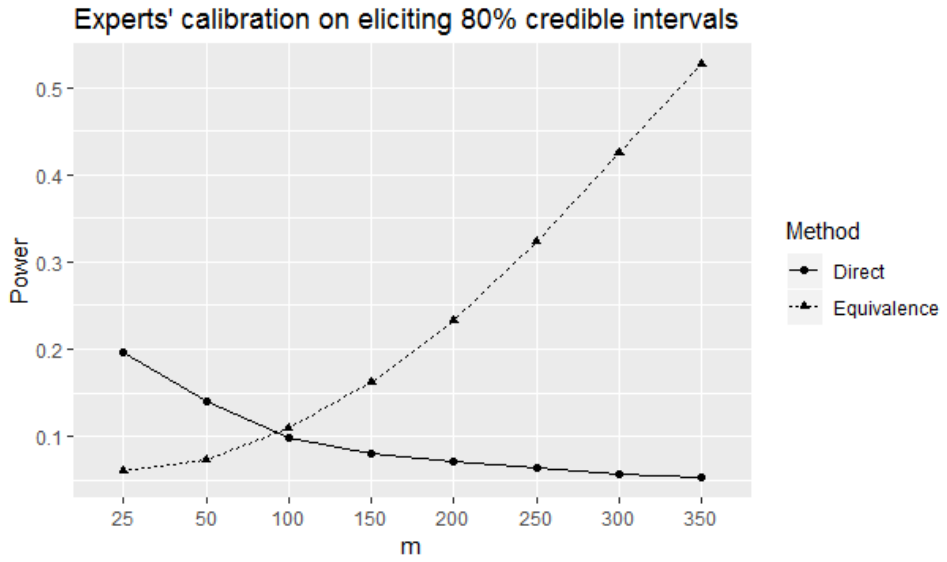


FIGURE 4.6: The power of the direct and equivalence tests to correctly identify 80% well-calibrated experts at 80% true level of coverage of elicited intervals

that the equivalence test works better than the direct test to correctly identify well-calibrated experts at large number of elicited intervals. The critical regions of the equivalence test for testing experts' calibration on eliciting 80% credible intervals are given in the table 4.2 below.

TABLE 4.2: Critical regions of the equivalence test

m	C1	C2
25	20	21
50	40	41
100	80	81
150	119	121
200	159	162
250	198	203
300	237	245
350	276	286

The plots indicating the probabilities of the direct and equivalence tests to identify the experts as 80% well-calibrated when true levels of coverage of elicited intervals that are less than 80% for large number of elicited intervals and the probabilities of the direct test to identify the experts as 80% well-calibrated when true levels of coverage of elicited intervals that are less than 80% for small number of elicited intervals are given in figures J.1 and J.2 of Appendix J, respectively. The patterns of behavior of the plots can be interpreted similarly as in the case of eliciting 90% credible intervals

above.

Section 4.8: Testing experts' calibration on eliciting credible intervals of different coverage probabilities

It was discussed above that the values of power to correctly identify well-calibrated experts on eliciting 80% credible intervals are lower than the corresponding values on eliciting 90% credible intervals from the direct and equivalence tests. Therefore, it is of interest to study whether the power of the direct and equivalence tests to correctly identify well-calibrated experts depends on the levels of intended coverage probabilities of elicited credible intervals. Figure 4.7 plots the values of power of the direct test at some selected levels of intended coverage probabilities. Note that the direct test produced zero power at certain number of elicited intervals at some selected levels of coverage probabilities as explained in the section 4.6.1 above. We ignored those zero values to clearly identify the overall pattern of the plot. Overall, it can be seen that the power of the direct test tends to increase with increased levels of coverage probabilities. It suggests to look into possibilities of optimizing the power by increasing the level of intended coverage probability of elicited credible intervals. However, the direct test has the problem of decreasing the power over the increase of number of elicited intervals for all the levels of coverage probabilities considered.

Now consider the figure 4.8 that contains the corresponding values of power above using the equivalence test. It also indicates the same pattern of dependency of power on the levels of intended coverage probabilities of elicited credible intervals. Also note that the values of power tend to increase as expected over the increase of number of elicited intervals. Therefore, it can be suggested to increase the power of the equivalence test to correctly identify well-calibrated experts by increasing the intended level of coverage probability and the number of elicited intervals within the considered range of this analysis.

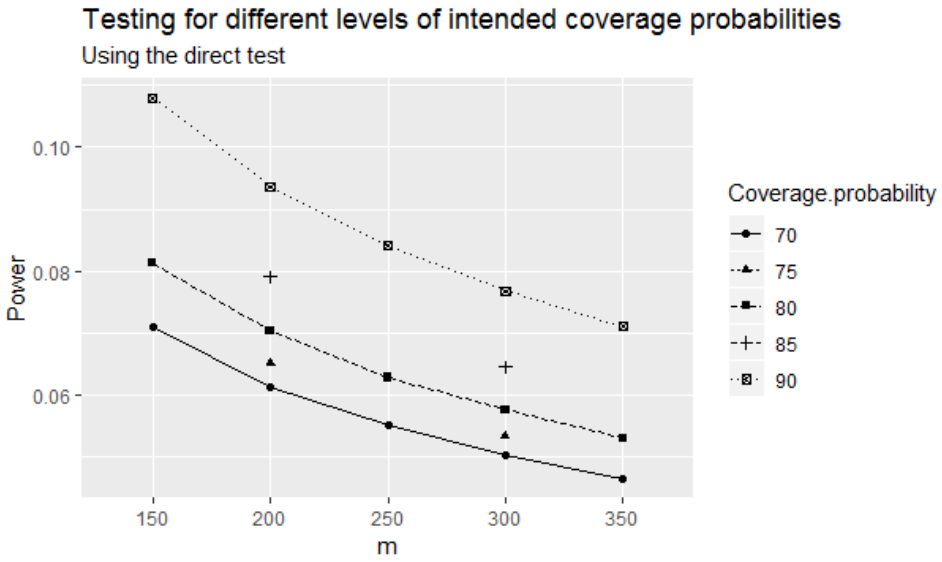


FIGURE 4.7: The power of the direct test to correctly identify well-calibrated experts at different levels of intended coverage probabilities

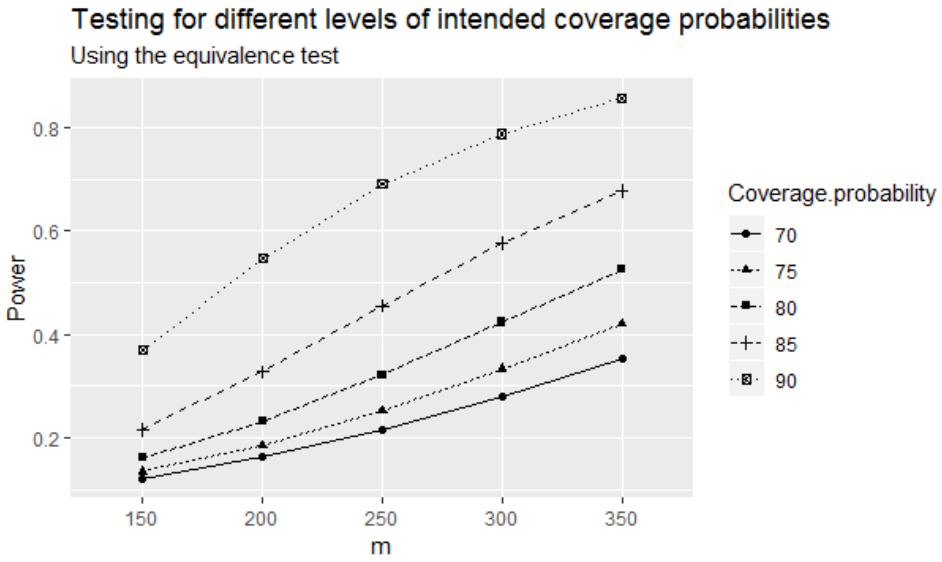


FIGURE 4.8: The power of the equivalence test to correctly identify well-calibrated experts at different levels of intended coverage probabilities

Section 4.9: Applying the non-randomized equivalence test

If we refer developing the equivalence test using the number of intervals containing true values (m_{int}) in the section 4.3.1, it can be seen that the rejection or acceptance of null hypotheses at $m_{int} = C_1$ and $m_{int} = C_2$ of the equivalence test should be decided based on comparing γ_1 and γ_2 with two uniform random numbers. This may be considered as a drawback of applying the exact test of size α in this context. Now consider applying the non-randomized equivalence test with the critical region of $C_1 < m_{int} < C_2$, even though it is a less than size α test. It can be identified from table 4.1 that the critical regions for testing experts' calibration on eliciting 90% credible intervals do not contain values greater than C_1 and less than C_2 for the number of elicited intervals less than or equal 80. Therefore, the non-randomized equivalence test with the critical region of $C_1 < m_{int} < C_2$ can only be applied for 100 or more intervals.

Figure 4.2 shows that the equivalence test can only be considered more effective than the direct test for 100 or more intervals. The values of power between tests are not considerably different at 80 intervals. Therefore, it seems meaningful to apply and observe the implications of the non-randomized equivalence test for 100 or more intervals. The values of power of the non-randomized equivalence test will be lower than the equivalence test due to the reduction of critical regions. Figure 4.9 plots the values of power of the direct and non-randomized equivalence tests. It shows that they are almost equal at 100 intervals and start to increase after that for the non-randomized equivalence test with the required property of increasing the power over the increase of number of elicited intervals. Also note that the computed values of power are less than the corresponding values of power from the equivalence test as expected.

We also consider the probabilities of the direct and non-randomized equivalence tests to identify the experts as 90% well-calibrated when true levels of coverage of elicited intervals that are less than 90% for large number of elicited intervals in figure 4.10. Observe that type I error probabilities of incorrectly identifying the experts as 90%

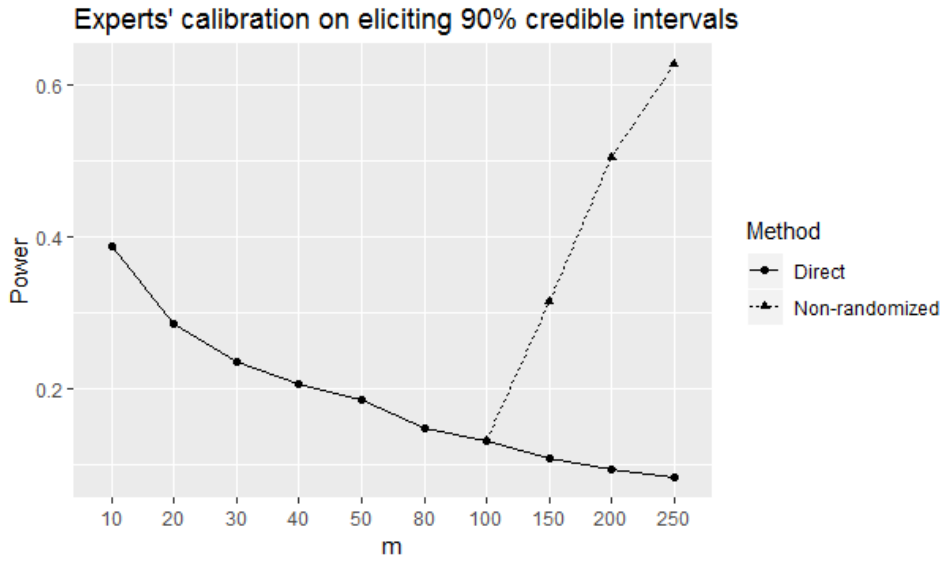


FIGURE 4.9: The power of the direct and non-randomized equivalence tests to correctly identify 90% well-calibrated experts at 90% true level of coverage of elicited intervals

well-calibrated at 85% true level of coverage of elicited intervals of the non-randomized test are less than 0.05 for the considered number of elicited intervals. It implies that the significance level of the non-randomized test is less than the nominal value 0.05. Therefore, the test is conservative with reduced power of rejecting the null hypotheses when they are false. It is indicated by observing lower values of probabilities compared to the equivalence test to identify the experts as 90% well-calibrated for the considered range of true levels of coverage of elicited intervals from 85% to 89%.

One might consider this as a better property of the non-randomized test to reduce the probabilities of incorrectly identifying the experts as 90% well calibrated when the true levels of coverage are less than 90%. However, this reduction of probabilities also affects the power of the test to correctly identify the experts as 90% well-calibrated at 90% true level of coverage which is also in the rejection region of the test. Therefore, it may require to increase the number of elicited intervals further to obtain sufficient level of power to correctly identify 90% well-calibrated experts. It will cause to increase the above mentioned probabilities of incorrectly identifying the experts as 90% well calibrated when true levels of coverage are less than 90% again. Hence, we consider to stay with the results of the equivalence test.

We do not report the case of testing experts' calibration on eliciting 80% credible

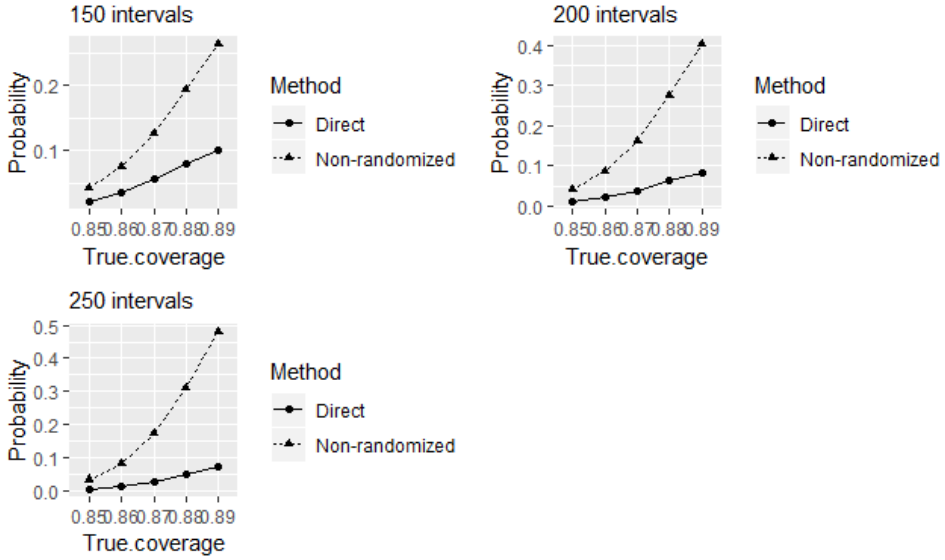


FIGURE 4.10: The probabilities of the direct and non-randomized equivalence tests to identify the experts as 90% well-calibrated when true levels of coverage of elicited intervals are less than 90%

intervals using the non-randomized equivalence test here as the observed results can be similarly interpreted as the 90% case above. Testing experts' calibration at different levels of intended coverage probabilities of elicited credible intervals in figure 4.11 indicates that higher values of power to correctly identify well-calibrated experts can be obtained by increasing the level of intended coverage probability for the non-randomized test as well.

Section 4.10: Discussion

We acknowledge the fact that it is difficult to conduct experiments in which expert elicitation gives interval judgements on a large number of quantities in practice. However, results of the conducted simulation study show that it is required to elicit a large number of intervals to obtain reasonably large value of power to correctly identify well-calibrated experts from the equivalence test of a single binomial proportion. The observed values of power of the equivalence test are comparatively higher than the direct comparison of hit rates for large number of elicited intervals considered in the analysis. However, the values of power are not very high within the considered range of elicited intervals. Therefore, it may be required to further increase the number of elicited intervals if higher power is necessary to identify well-calibrated experts.

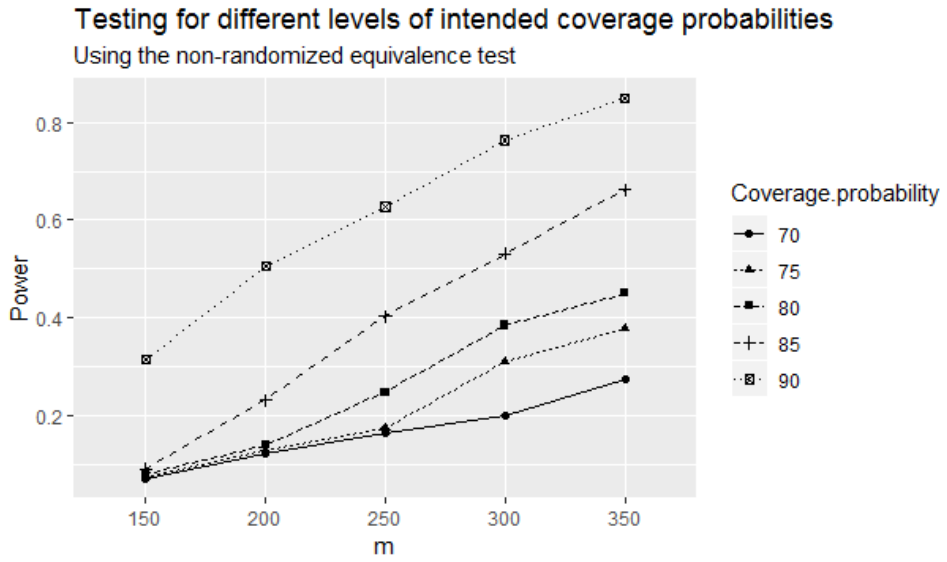


FIGURE 4.11: The power of the non-randomized equivalence test to correctly identify well-calibrated experts at different levels of intended coverage probabilities

We selected the value of ϵ (the acceptable margin of deviation around the level of intended coverage probability of elicited credible intervals) of the equivalence test equals to 0.05 in this analysis. It was assumed that deviation of such amount of true level of coverage of elicited intervals from the level of intended coverage probability is reasonable enough to consider the experts as well-calibrated. It can be considered reasonable not to think about increasing the margin further in practice. However, one might think about reducing the margin depending on the context. The reduction of the margin will reduce the critical regions of the test and reduce the power of the test to correctly identify well-calibrated experts accordingly.

Section 4.11: Conclusion

The focus of this section of the analysis is to assess the properties of testing experts' calibration on eliciting credible intervals for unknown quantities using the direct comparison of experts' hit rates; observed proportions of experts' elicited intervals that contain realized values of given quantities (McBride, Fidler, and Burgman, 2012), with the level of intended coverage probability of elicited credible intervals. The results of the conducted simulation study at 90% and 80% standard levels of intended coverage probabilities show that the direct comparison of hit rates has a property of

obtaining lower values of power to correctly identify well-calibrated experts and more importantly, the power tends to decrease as the number of elicited intervals increases. This is a contradictory result from a statistical point of view as the power of the test to correctly identify well-calibrated experts can be expected to increase over the increase of number of elicited intervals. Furthermore, the power of the test at small number of intervals are also not large enough with the additional problem of observing relatively higher values of probabilities; compared to the observed values of power to correctly identify well-calibrated experts, to incorrectly identify the experts as well-calibrated for the considered true levels of coverage of elicited intervals that remain within 0.05 margin less than the levels of intended coverage probabilities considered.

We explored the potential of applying the equivalence test of a single binomial proportion to overcome the above discussed issues of the direct comparison of hit rates. The values of power of the equivalence test tend to increase as intuitively expected over the increase of number of elicited intervals and they are comparatively higher than the corresponding values of the direct test for large number of elicited intervals. Therefore, this avoids the need of performing the analysis with small number of elicited intervals with lower values of power of the direct comparison of hit rates. However, it is achieved at the expense of receiving higher probabilities of incorrectly identifying the experts as well-calibrated compared to the direct comparison of hit rates at true levels of coverage of elicited intervals that remain within the 0.05 margin mentioned above due to a characteristic of the equivalence test.

It can be considered reasonable to declare experts as well-calibrated if their true levels of coverage of elicited intervals remain within a certain acceptable margin of deviation from the levels of intended coverage probabilities of elicited credible intervals. Therefore, we recommend to apply the equivalence test of a single binomial proportion with large number of elicited intervals to test experts' calibration in this context. It was also observed that the power of the direct and equivalence tests to correctly identify well-calibrated experts tends to increase with increased levels of intended coverage probabilities. Therefore, we suggest to look into possibilities of optimizing the power to correctly identify well-calibrated experts by increasing the levels of intended coverage probabilities of elicited intervals of quantities, appropriately.

5. Testing experts' calibration II

Section 5.1: Introduction

The preceding chapter described testing expert's calibration on eliciting credible intervals of unknown quantities. We considered the case of eliciting credible intervals in general without imposing any additional constraints on what the intervals represent, for example there was not a restriction to elicit medians (or the 50% percentiles) of probability distributions of quantities and to produce centralized credible intervals around the elicited medians of quantities. Therefore, the experts have the flexibility to produce lower and upper limits of credible intervals with given level of intended coverage probability from any position of their subjective probability distributions of quantities. Now we focus on testing experts' calibration in a more structured context of eliciting a specified number of percentiles from the probability distributions of unknown quantities.

Experts' weights are derived to combine experts' elicited subjective probability distributions to obtain aggregated probability distributions of unknown quantities (O'Hagan, 2019). Assessing the levels of experts' calibration on eliciting a specified number of percentiles from the probability distributions of quantities of some given seed questions (with known realized values of quantities to the researchers but not to the experts) using the calibration score component is a part of deriving experts' weights using the Cooke's classical model (Cooke, 1991). We consider some commonly specified number of percentiles to be elicited from probability distributions and assess the levels of experts' calibration using the calibration score component of the Cooke's classical model (Cooke, 1991). Here, we refer the application of the calibration score component to assess the levels of experts' calibration as the calibration test.

The results of a conducted simulation study at some commonly specified number of percentiles to be elicited show that the calibration test fails to detect not well-calibrated experts with adequately higher values of power even for reasonably large number of elicited quantities. Furthermore, average calibration scores are reasonably high when the calibration test fails to detect not well-calibrated experts. It indicates a possibility of allocating higher weights to some of the not well-calibrated experts from the Cooke's classical model if they have also obtained higher information scores. Therefore, there is a need of providing a safe guard for allocating higher weights to not well-calibrated experts from the process. It is evident from the analysis that the multinomial equivalence test can be used to identify not well-calibrated experts with higher accuracy and overcome the potential issue of allocating higher weights to not-well calibrated experts.

Section 5.2: Assessing experts' calibration using the Cooke's classical model

Now we review the background of the calibration component of the Cooke's classical model to assess the levels of experts' calibration on eliciting percentiles of probability distributions of quantities to derive experts' weights due to Cooke (1991, chap. 12) as follows. Consider a scenario in which expert elicitation gives $(n-1)$ fixed number of increasing percentiles from the probability distributions of m quantities; $X(1), X(2), ..., X(m)$. Let $X(i)_r$ denote the r^{th} percentile ($r \in (0, 100)$) of the probability distribution of the quantity $X(i); i = 1, 2, ..., m$.

Now define

$$P_r = Prob(X(i) \leq X(i)_r); \qquad r = 1, 2, ..., (n-1).$$

$$p_r = P_r - P_{r-1}; \qquad r = 1, 2, ..., n.$$

Here, we use the notation $Prob$ to denote the probability. It follows from the definitions above that, $p_1 = P_1$, $p_n = 1 - P_{n-1}$, and the probability that a realized value $x(i)$ of $X(i)$ falls between the $(r-1)^{th}$ and r^{th} quantiles is p_r, where $p_r \geq 0; r = 1, 2, ..., n$.

The relative information is one of the key concepts used in deriving Cooke's classical model. Let

$$S^n = \Big(p = (p_1, p_2, ..., p_n) \in \mathbb{R}^n | p_r \geq 0, r = 1, 2, ..., n; \sum_{i=1}^{n} p_r = 1\Big), \tag{5.1}$$

where S^n is the set of positive probability vectors over n cells. Now for $s, p \in S^n$, the relative information of s with respect to p; $I(s, p)$, can be defined as

$$I(s, p) = \sum_{r=1}^{n} s_r \ln\Big(\frac{s_r}{p_r}\Big). \tag{5.2}$$

The theoretical probabilities p_r of the cells $r = 1, 2, ..., n$ can be compared with the corresponding empirical cell probabilities s_r can be estimated using the realized values of m quantities as follows.

$$s_r = \frac{\#\Big(i\Big|X(i)_{r-1} < x(i) \leq X(i)_r\Big)}{m}, \tag{5.3}$$

where # indicates the number of realized values in each of the cells; $r = 1, 2, ..., n$.

Calibration score

Following proposition due to Cooke (1991, chap. 12) explains the statistical background behind computing experts' calibration scores in the Cooke's classical model.

Proposition 1 *If $s \in S^n$ is the empirical distribution generated by m independent samples from $p \in S^n$, then the distribution of $2mI(s, p)$ approaches a chi − square distribution with $(n - 1)$ degrees of freedom as $m \to \infty$.*

The calibration score (Cal) for a given expert is defined in the Cooke's model using the above proposition as

$$Cal = 1 - F_{\chi^2_{n-1}}(2mI(s, p)), \tag{5.4}$$

where s denotes the empirical probability vector of p cells for the given expert. Here, $F_{\chi^2_{n-1}}$ denotes the cumulative distribution function of the χ^2_{n-1} distribution in standard

notation. Therefore, the calibration score (Cal) computes the probability of observing a value greater than the estimated value of $2mI(s,p)$ from the χ^2_{n-1} distribution for the considered expert. Following the definition of the relative information in equation 5.2 above, it can be shown that $I(s,p) \geq 0$ and $I(s,p) = 0$ if and only if $s = p$. Therefore, a lower value of $2mI(s,p)$ indicates better calibration with a small discrepancy between s and p on m quantities. Hence, a calibration score near 1 indicates better calibration and a calibration score near 0 indicates poor calibration.

The above discussed calibration score is a p-value calculated from the corresponding chi-square distribution for a given expert under the assumption that the considered expert is well-calibrated. It will be followed by using the computed calibration scores of experts to derive experts' weights to obtain aggregated probability distributions of unknown quantities as discussed in the next chapter. Higher weights are allocated to the experts with higher p-values indicating better calibration of elicited distributions if they have also obtained higher information scores of elicited distributions. From a statistical point of view, it implies that higher weights are allocated to the experts who are identified as well-calibrated from the calibration test as the p-values are computed under the assumption that the experts are well-calibrated. This forms a statistical hypothesis testing with the null hypothesis that a given expert is well-calibrated against the alternative hypothesis that the considered expert is not well-calibrated.

It is important to note that unable to reject the null hypothesis based on a higher p-value does not necessarily imply that the considered expert is well-calibrated in this instance. It implies that the data do not provide sufficient evidence to reject the null hypothesis as discussed in the preceding chapter. Hence, there is a possibility to allocate higher weights even to not well-calibrated experts from the process. Here, we assume that the error of allocating higher weights to not well-calibrated experts is more serious than the error of not allocating weights to well-calibrated experts. Therefore, it is useful to assess the properties of power of the Cooke's calibration test to correctly identify not-well calibrated experts and study how the weights are allocated to not well-calibrated experts.

Section 5.3: Methodology of the study

Following the discussion above, it is of interest to perform a simulation study to assess the properties of power of the Cooke's calibration test to correctly identify not-well calibrated experts at two of the commonly used specifications of number of percentiles; eliciting $5\%, 50\%$, and 95% percentiles and eliciting $5\%, 25\%, 50\%, 75\%$, and 95% percentiles with varied number of m quantities. We first consider the case of eliciting $5\%, 50\%$, and 95% percentiles from the probability distributions of quantities. It follows from the discussion above that the theoretical probability vector for this context is $p = (0.05, 0.45, 0.45, 0.05)$ with cell probabilities $p_r; r = 1, 2, 3, 4$. We refer this probability vector as p_1 in the analysis.

When eliciting percentiles, the experts assume that they are being asked for $5\%, 50\%$, and 95% percentiles from the probability distributions of given quantities. However, the elicited percentiles can be different to $5\%, 50\%$, and 95% in practice without the knowledge of experts. Hence, true probability vectors (t_1) of experts' elicited percentiles can be varied from the intended theoretical probability vector (p_1) in this instance. Therefore, we make a simulation to assess the power of the Cooke's calibration test to correctly identify not well-calibrated experts at some selected true probability vectors of experts' elicited percentiles that are different to the theoretical vector $p_1 = (0.05, 0.45, 0.45, 0.05)$ of interest. It will be followed by estimating the probabilities of rejecting the null hypothesis to obtain the power of the calibration test to correctly identify not well-calibrated experts by repeating the simulation 100000 times at each selected combination of true probability vector (t_1) and number of quantities (m). It is expected to increase the number of quantities at each selected true probability vector until a target value of 0.90 power to detect not well-calibrated experts is achieved. Furthermore, average calibration scores will be computed when the null hypothesis is not rejected at each combination above to study how the weights are allocated to not-well calibrated experts.

It can be very rare to observe situations where the true probability vectors of experts' elicited percentiles match exactly with the intended theoretical probability vector of interest in practice. Therefore, similar to the argument we considered in the previous

section of the analysis, it requires to set a certain total acceptable margin of deviations between the cell probabilities of the theoretical probability vector of interest and true probability vectors of elicited percentiles that is reasonable enough to consider a given expert as well-calibrated. Following the acceptable margin of deviation of 0.05 that was used in the previous analysis of the equivalence test of a single binomial proportion, 0.05 deviation of each cell will be considered to define a cut off margin for comparing four probability cells. It is equivalent to the Euclidean distance of $\sqrt{4 \times 0.05^2} = 0.1$ between four probability cells. Hence, we consider an expert as not well-calibrated if the Euclidean distance between probability vectors is greater than or equal 0.1 in this analysis.

Section 5.4: The analysis

We consider important to assess the power of the calibration test at true probability vectors that are adequately deviated from the theoretical probability vector p_1. If they are not deviated much, interest may not be there to assess the power of the test to distinguish true and theoretical vectors in this context. Therefore, the analysis will be carried out by assessing the power of the calibration test to correctly identify not well-calibrated experts at some selected true probability vectors of experts' elicited percentiles that are considerably deviated from p_1. First consider the following symmetric heavy tail true probability vector $t_{11} = (0.15, 0.35, 0.35, 0.15)$ with some considerable deviations between the probability cells. The significance level of the test α was taken equal to 0.05 and the null hypothesis of a given expert is well-calibrated was rejected if the calibration score less than or equal to 0.05. Table 5.1 provides the values of power of the Cooke's calibration test to correctly identify a given expert as not well-calibrated (probabilities of rejecting the false null hypothesis that a given expert is well-calibrated) and the average calibration scores when the false null hypothesis is not rejected with varied number of quantities (m).

It is reasonable to assume that a true probability vector with this amount of deviation from the theoretical probability vector of interest should be identified as statistically different with considerably higher value of power from the test. Observe that 50

TABLE 5.1: Cooke's test results for $t_{11} = (0.15, 0.35, 0.35, 0.15)$

m	power	average calibration score
25	0.60948	0.25835
30	0.69762	0.24469
35	0.75720	0.22183
40	0.82273	0.21623
45	0.86851	0.20712
50	0.89537	0.18858

quantities should be elicited to obtain the target value of 0.90 power approximately. Hence, reasonably higher type II errors (probabilities of not rejecting the false null hypothesis) are obtained (1-power) even for reasonably large number of elicited quantities. Therefore, the Cooke's calibration test will tend to declare experts with this amount of deviation between probability vectors as well-calibrated for small number of quantities. More importantly the average calibration scores are reasonably high when the false null hypothesis is not rejected for the considered range of number of elicited quantities. Considering the fact that observations are fluctuated above and below the average value of a distribution, it can be concluded that larger calibration scores above the given average calibration scores can be obtained for some experts. Hence, there is a possibility to allocate larger weights even to not well-calibrated experts from the process.

Now consider the following non-symmetric heavy tail true probability vector $t_{12} = (0.15, 0.35, 0.4, 0.1)$. According to table 5.2, approximately 80 elicited quantities are required to obtain 0.90 power to identify a given expert as not well-calibrated here.

TABLE 5.2: Cooke's test results for $t_{12} = (0.15, 0.35, 0.4, 0.1)$

m	power	average calibration score
25	0.43087	0.30676
50	0.72624	0.23572
75	0.88405	0.18612
80	0.90394	0.17754

We also considered the non-symmetric light tail true probability vector in the form of $t_{13} = (0, 0.3, 0.7, 0)$. Table 5.3 shows an interesting pattern of obtaining a higher value of power at 25 elicited quantities and rapidly increasing the power over the increase of number of quantities with comparatively lower average calibration scores when the false null hypothesis is not rejected.

TABLE 5.3: Cooke's test results for $t_{13} = (0, 0.3, 0.7, 0)$

m	power	average calibration score
25	0.67698	0.09630
30	0.84062	0.07097
35	0.96536	0.05650

The analysis was further continued with two selected positive and negative skewed true probability vectors of $t_{14} = (0.2, 0.4, 0.3, 0.1)$ and $t_{15} = (0, 0.3, 0.5, 0.2)$, respectively. Table 5.4 for the positive skewed case shows that approximately 45 quantities should be elicited to obtain the expected 0.90 power to identify a given expert as not well-calibrated. The required number of elicited intervals to attain the expected power is 35 for the considered negative skewed case as given in table 5.5.

TABLE 5.4: Cooke's test results for $t_{14} = (0.2, 0.4, 0.3, 0.1)$

m	power	average calibration score
25	0.67175	0.24266
30	0.76343	0.23332
35	0.81586	0.20938
40	0.87232	0.20497
45	0.90708	0.19229

TABLE 5.5: Cooke's test results for $t_{15} = (0, 0.3, 0.5, 0.2)$

m	power	average calibration score
25	0.71181	0.15529
30	0.83673	0.13918
35	0.89231	0.11120

The results above show that when the difference between probability vectors increases, the number of quantities required decreases to conclude a given expert as not well-calibrated with the target 0.90 level of power using the Cooke's calibration test. However, the average calibration scores are still reasonably high enough when the false null hypothesis of a given expert is well-calibrated is not rejected with possibilities to allocate higher weights to some experts. Therefore, we consider important to find a method that identifies true probability vectors of experts' elicited percentiles outside a certain acceptable margin of deviation from the intended probability vector of interest with higher accuracy to provide a safe guard for allocating higher weights to not well-calibrated experts from the Cooke's classical model.

We mentioned at the beginning of this analysis that concluding not well-calibrated

experts as well-calibrated can be considered as the most serious error in this context than the error of concluding other way around. The above discussed issue of the Cooke's calibration test is caused by this error. Therefore, we are interested to find a solution to the above issue using the multinomial equivalence test by defining the null hypothesis as a given expert is not well-calibrated and conclude the opposite if the data provide sufficient evidence.

Section 5.5: Multinomial equivalence test

We now review the multinomial equivalence test due to Frey (2009) as follows. Suppose we collect m independent observations, each of which will fall into exactly one of k categories. If we count the number of observations in each category; $m_1, m_2, ., m_k$, it leads to multinomial data $(m_1, m_2, ., m_k) \sim$ Multinomial(m, P), where $P = (P_1, P_2, .., P_k)$. Suppose, we are interested to test whether the observed multinomial data are consistent with a given specified vector $P_0 = (P_{10}, P_{20}, .., P_{k0})$ for the underlying probabilities of k categories. A goodness-of-fit approach to this problem involves testing $H_0 : P = P_0$ against $H_0 : P \neq P_0$ for a given P_0. Similar to the case discussed in the previous section for a single binomial proportion, even if we do not reject H_0, it does not imply that $P = P_0$. In fact we do not have enough evidence to reject H_0. We discussed above that considering two probability vectors as equal when they are different will be considered as the most serious error in our context of testing calibration of experts. Therefore, now we review the multinomial equivalence test that defines the null hypothesis as two probability vectors are different and concludes the opposite if the data provide sufficient evidence.

The above mentioned scenario is discussed in Wellek (2010, chap. 9) under the testing for equivalence of a single multinomial distribution with a fully specified reference distribution. The ordinary Euclidean distance between the corresponding parameter vectors P and P_0 are considered to formulate the following hypothesis testing of the goodness of fit of multinomial (M) distributions $M(m; P_1, P_2, ..., P_k)$ to M(m;

$\mathrm{P}_{10}, P_{20}, .., P_{k0})$ as

$$H_0 : d^2(P, P_0) \geq \varepsilon^2 \text{ versus } H_1 : d^2(P, P_0) < \varepsilon^2, \tag{5.5}$$

where $d^2(P, P_0) = \sum_{j=1}^{k} (P_j - P_{j0})^2$ denotes the square of the Euclidean distance between two vectors P and P_0. The test rejects H_0 or the lack of fit of $M(m; P_1, P_2, ..., P_k)$ to $\mathrm{M(m;\ P}_{10}, P_{20}, .., P_{k0})$ under α level of significance if $d^2(\hat{P}, P_0) < \varepsilon^2 - u_{1-\alpha}\,\nu(\hat{P}, P_0)/\sqrt{m}$. Note that $u_{1-\alpha} = \Phi^{-1}(1-\alpha)$ and $\nu^2(\hat{P}, P_0)$ is the variance estimator of the statistic given by $\nu^2(\hat{P}, P_0) = 4\Big[\sum_{j=1}^{k}(\hat{P}_j - P_{j0})^2\hat{P}_j - \sum_{j_1=1}^{k}\sum_{j_2=1}^{k}(\hat{P}_{j_1} - P_{j_1 0})(\hat{P}_{j_2} - P_{j_2 0})\hat{P}_{j_1}\hat{P}_{j_2}\Big]$. Furthermore, $\hat{P} = \Big(\frac{m_1}{m}, \frac{m_2}{m}, .., \frac{m_k}{m}\Big)$.

Section 5.6: Applying the multinomial equivalence test

We now apply the multinomial equivalence test for testing experts' calibration for $p_1 = (0.05, 0.45, 0.45, 0.05)$ at the true probability vectors; $t_{11}, t_{12}, t_{13}, t_{14}$ and t_{15}, considered above. It was discussed above to consider an expert as not well-calibrated if the Euclidean distance between probability vectors is greater than or equal 0.1 in this analysis. Therefore, we choose 0.1 as the cut off margin to the Euclidean distance ε between probability vectors for applying the multinomial test in the analysis. It ensures that Euclidean distances between all the selected true probability vectors and p_1 are above the cut off margin. Recall that the null hypothesis of the multinomial equivalence test for testing experts' calibration is defined as a given expert is not well-calibrated. Therefore, what requires here is to obtain lower probabilities of rejecting the null hypothesis for all the selected true probability vectors that are different from p_1. Figure 5.1 shows that the estimated type I error probabilities of the multinomial equivalence test are very small and less than the significance level 0.05 of the test within the considered range of number of quantities.

It was identified that the multinomial equivalence test could produce unstable estimates when the number of observations are small. In order to overcome this problem and to make sure that a reasonably large number of observation are used to satisfy the large sample requirement of the multinomial equivalence test, we consider a minimum of 25

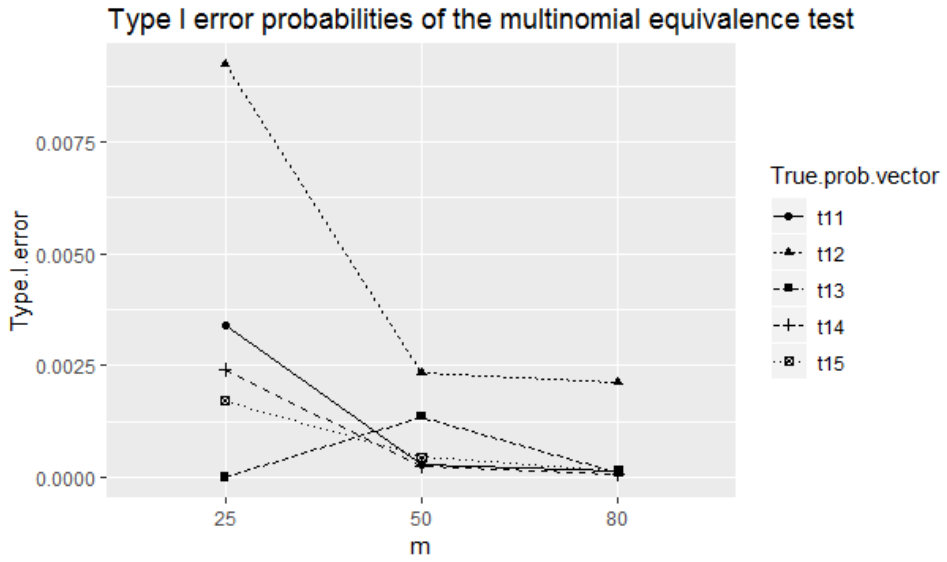


FIGURE 5.1: The estimated type I error probabilities of the multinomial equivalence test for testing experts' calibration for $p_1 = (0.05, 0.45, 0.45, 0.05)$

elicited quantities for the analysis. Furthermore, a general range of 25 to 80 quantities is considered for all the true probability vectors to monitor the pattern of behavior of type I error probabilities. Here, the range of 25 to 80 quantities is decided by the maximum range of quantities required to obtain 0.9 power for the true probability vector t_{12} above.

It is important to note that the multinomial equivalence test provides the required lower type error probabilities to reject the null hypothesis at 25 quantities for all the true probability vectors considered. It should also be noted that the type I error probabilities decrease as the number of quantities increases in general except for observing a lower type I error at 25 quantities for the true probability vector t_{13} probably by chance.

Now consider the case of testing experts' calibration on eliciting $5\%, 25\%, 50\%, 75\%$, and 95% percentiles of the probability distributions of quantities. Here, the theoretical probability vector is $p = (0.05, 0.2, 0.25, 0.25, 0.2, 0.05)$. We refer this as p_2 in the analysis. Similar to the above analysis, we consider 0.05 deviation of each cell that is equivalent to the Euclidean distance of $\sqrt{6 \times 0.05^2} = 0.1225$ as the cut off margin for comparing six probability cells. We also consider the importance of assessing the power of the calibration test at true probability vectors that are adequately deviated

from the theoretical probability vector p_2. Note that arranging different combinations of probability cells to form different distributional patterns is not straight forward here with six probability cells. Therefore, we restrict our attention to a non-symmetric heavy tail true probability vector (t_{21}), a symmetric light tail true probability vector (t_{22}), and a positive skewed probability vector (t_{23}) in this analysis.

First consider the non-symmetric heavy tail true probability vector $t_{21} = (0.1, 0.25, 0.15, 0.1, 0.3, 0.1)$. Table 5.6 provides the estimated power of the Cookes' calibration test to identify this difference between probability vectors to conclude a given expert as not well-calibrated and the average calibration scores when the false null hypothesis of a given expert is well-calibrated is not rejected as in the above analysis. In addition to that, estimated type I error probabilities of rejecting the null hypothesis of the multinomial equivalence test to conclude a given expert as well-calibrated are also provided. Observe that same patterns of obtaining lower power of the Cookes' calibration test and reasonably high average calibration scores are shown even for reasonably large number of elicited quantities as above. Furthermore, lower type I error probability estimates are obtained from the multinomial equivalence test as expected.

TABLE 5.6: Calibration test results for $t_{21} = (0.1, 0.25, 0.15, 0.1, 0.3, 0.1)$

m	power	average calibration score	type I error
25	0.56022	0.23766	0.00008
30	0.64293	0.21528	0.00015
35	0.72247	0.19992	0.00011
40	0.78824	0.18607	0.00011
45	0.83805	0.17264	0.00007
50	0.87870	0.16139	0.00005
55	0.91152	0.15265	0.00004

The results of the analysis given in table 5.7 for the case of the symmetric light tail true probability vector $t_{22} = (0, 0.1, 0.4, 0.4, 0.1, 0)$ shows similar type of pattern of obtaining a higher value of power at 25 elicited quantities and rapidly increasing the power over the increase of number of quantities with comparatively lower average calibration scores when the false null hypothesis is not rejected as in the analysis of the true probability vector $t_{13} = (0, 0.3, 0.7, 0)$ above. However, still reasonably large number of quantities are required to attain the required level of power to

identify a given experts as not well-calibrated. Observe that the estimated type I error probabilities of the multinomial equivalence test to conclude a given expert as well-calibrated are small as required.

TABLE 5.7: Calibration test results for $t_{22} = (0, 0.1, 0.4, 0.4, 0.1, 0)$

m	power	average calibration score	type I error
25	0.74133	0.11807	0.00000
30	0.87425	0.09401	0.00005
35	0.95710	0.08209	0.00002

Finally, the overall pattern of results of the analysis of the considered positive skewed probability vector $t_{23} = (0.1, 0.4, 0.2, 0.15, 0.1, 0.05)$ given in table 5.8 can also be interpreted similarly to the case of non-symmetric heavy tail true probability vector t_{21} above.

TABLE 5.8: Calibration test results for $t_{23} = (0.1, 0.4, 0.2, 0.15, 0.1, 0.05)$

m	power	average calibration score	type I error
25	0.58849	0.24701	0.00012
30	0.66875	0.22481	0.00028
35	0.73777	0.20811	0.00026
40	0.79696	0.19401	0.00013
45	0.84582	0.18213	0.00016
50	0.88509	0.16999	0.00006
55	0.91550	0.16348	0.00009

Section 5.7: Discussion

We considered important to assess the power of the Cooke's calibration test to correctly identify not well-calibrated experts at true probability vectors of experts' elicited percentiles that are adequately deviated from the theoretical probability vectors of intended percentiles to be elicited. The observed type I error probabilities of identifying a not well-calibrated expert as well-calibrated of the multinomial equivalence test are very small and considerably lower than the chosen significance level 0.05 of the analysis. An acceptable margin of deviation of 0.05 between each of cell probabilities was considered to compute a cut-off margin of Euclidian distance between true and theoretical probability vectors to consider an expert as well-calibrated or not in the analysis. Computed Euclidian distances between the considered true probability

vectors and theoretical probability vectors were higher than the corresponding cut-off Euclidian distances of the analysis. It was the reason to obtain very small type I error probabilities of identifying a not well-calibrated expert as well-calibrated in the analysis.

It is a known fact that obtaining lower type I errors causes to obtain higher type II errors in hypothesis testing. We do not focus on the type II error probabilities (probabilities of identifying a well-calibrated expert as not well-calibrated) in this analysis. The properties of type II error probabilities can be assessed by performing a power analysis of the multinomial equivalence test. It is outside the scope as the focus of this analysis is to take a remedial action to the most serious error of identifying a not well-calibrated expert as well-calibrated and allocating weights inappropriately. It is flexible to choose an appropriate cut-off margin of Euclidian distance between true and theoretical probability vectors to reduce the type I error probabilities depending on the context of the analysis.

Section 5.8: Conclusion

Experts' weights are derived to combine experts' elicited subjective probability distributions to obtain aggregated probability distributions of unknown quantities (O'Hagan, 2019). Assessing the levels of experts' calibration on eliciting a specified number of percentiles from the probability distributions of quantities of some given seed questions (with known realized values of quantities to the researchers but not to the experts) using the calibration score component is a part of deriving experts' weights using the Cooke's classical model (Cooke, 1991). We refer the application of the calibration score component to assess the levels of experts' calibration as the calibration test in this analysis. Here, the focus of the analysis is to explore the properties of the calibration test in assessing the levels of experts' calibration on eliciting a specified number of percentiles from the probability distributions of quantities in deriving experts' weights.

The results of a conducted simulation study at some commonly specified number of percentiles to be elicited show that the calibration test fails to detect not well-calibrated experts with adequately higher values of power even for reasonably large number of

elicited quantities. Furthermore, average calibration scores are reasonably high when the calibration test fails to detect not well-calibrated experts. It indicates a possibility of allocating higher weights to some of the not well-calibrated experts from the process if they have also obtained higher information scores. Therefore, it is important to find a method that identifies true probability vectors of experts' elicited percentiles outside a certain acceptable margin of deviation from the intended probability vector of interest with higher accuracy to provide a safe guard for allocating higher weights to not well-calibrated experts from the Cooke's classical model, even though they have obtained higher information scores.

It is evident from the analysis that the multinomial equivalence test with the null hypothesis of a given expert is not well-calibrated against the alternative hypothesis of a given expert is well-calibrated can be applied to overcome the above issue with suitably chosen acceptable margin of deviation between the probability vectors of experts' elicited percentiles and the intended percentiles to be elicited that can be reasonable enough to consider a given expert as well-calibrated. Multinomial equivalence test can be applied to produce lower probabilities of identifying a not well-calibrated expert as well-calibrated with reasonably large number of elicited intervals that sufficiently satisfy the large sample approximation of the test. Therefore, once the null hypothesis of the multinomial equivalence test is rejected, it can be assumed with higher confidence that the considered expert is well-calibrated. Hence, we recommend to apply the multinomial equivalence test to identify well-calibrated experts and proceed to compute experts' weights using the Cooke's classical model to reduce the potential risk of allocating higher weights to not well-calibrated experts.

6. Different way of deriving experts' weights

Section 6.1: Introduction

Experts' weights play a very important role in group decision making to obtain aggregated probability distributions of unknown quantities. Aggregated probability distributions of unknown quantities are obtained by combining the experts' elicited probability distributions of corresponding quantities together with suitable weights for experts. When multiple experts are consulted, experiments are performed to derive experts' weights based on their performance on predicting the probability distributions of quantities of some given seed questions of which the true values of quantities are known to the researchers but not to the experts. Experts' performance in making judgments about the seed questions will be considered as an indicator of how good their judgements about the unknown quantities of interest (O'Hagan, 2019). Here, we consider the Cooke's classical model (Cooke, 1991) considered in the previous chapter that derives experts' weights based on their performance in making judgments about the seed questions for the analysis.

We do not know whether the derived experts' weights from experiments are precise enough in practice as they are random variables subject to uncertainty. Therefore, it is important to address this underlying uncertainty of the derived experts' weights from experiments in computing aggregated distributions of quantities. James-Stein shrinkage estimation technique discussed in James and Stein (1961) can be used to estimate the mean of a multivariate normal distribution with reduced mean squared errors. Therefore, it is of interest to explore the potential of applying the James-stein technique to obtain experts' weights with reduced mean squared errors. We apply

an empirical Bayes development of the James-Stein shrinkage estimation technique discussed in Zhao (2010) that shrinks variables differently depending on their variances (larger the variance more shrinkage there should be) on Cooke's weights to derive shrinkage weights with reduced mean squared errors in this analysis. Here, the focus is to study the impact of deriving shrinkage weights with reduced mean squared errors on the calibration and information scores of aggregated probability distributions of quantities.

The results of the analysis show that overall Decision Maker (DM) calibration and information scores of some selected testing questions with known realized values are different between the normalized typical and shrinkage Cooke's weights. Considering the purpose of applying the shrinkage estimation technique to reduce the mean squared errors of estimators of the mean of a multivariate normal distribution due to James and Stein (1961) and the procedure of the employed empirical Bayes shrinkage approach to derive weights by reducing the allocated weights for experts with larger variances of weights in Zhao (2010), we suggest that the outcome of the shrinkage weights can expect to be more accurate than the typically computed weights.

Section 6.2: Background of deriving weights

When decisions are made under uncertainty, relying on the knowledge of a single expert may not be advisable in practice. Hence, it is common to elicit judgements from more than one expert when formal expert elicitation is employed (O'Hagan, 2019). It follows the need of resolving the experts' judgments into a single distribution representing the combined knowledge of experts. This is known as the problem of aggregation. There are two principle approaches; Mathematical aggregation and Behavioral aggregation. In mathematical aggregation, separate judgments are elicited from the experts and a probability distribution is fitted to each expert's judgments. Then, separately fitted probability distributions are combined to form the aggregate distribution using a mathematical formula (a pooing rule). The behavioral approach proceeds in a way that the group of experts have to discuss their knowledge and opinions to produce group "consensus" judgments to which an aggregate distribution

will be fitted (O'Hagan, 2019). Here, we focus on the mathematical aggregation of combining separately fitted experts' probability distributions using a mathematical formula.

Suppose subjective probability distributions for an unknown quantity $\tilde{\theta}$ have been elicited from q experts. Let $f_i(\tilde{\theta})$ indicate the subjective distribution for $\tilde{\theta}$ that has been elicited from the i^{th} expert, where $i = 1, 2, ..., q$. Now consider the following two main methods of mathematical aggregation to obtain the aggregated distribution $f(\tilde{\theta})$ of $\tilde{\theta}$ due to O'Hagan et al. (2006). The first method is the Bayesian approach of aggregation. Here, the decision maker has to decide the prior distribution $f(\tilde{\theta})$ of $\tilde{\theta}$ that needs to be updated using the information obtained from experts' elicited distributions; $f_1(\tilde{\theta}), f_2(\tilde{\theta}), ..., f_q(\tilde{\theta})$, to obtain the posterior distribution $f(\tilde{\theta}|D)$ of $\tilde{\theta}$ with $D = [f_1(\tilde{\theta}), f_2(\tilde{\theta}), ..., f_q(\tilde{\theta})]$. Following equation 6.1 gives the general formula for deriving posterior distributions in Bayesian analyses.

$$f(\tilde{\theta}|D) \propto f(\tilde{\theta}).f(D|\tilde{\theta}), \tag{6.1}$$

where $f(D|\tilde{\theta})$ denotes the likelihood function of the observed data D conditional on the true unknown density function of $\tilde{\theta}$. The likelihood function has to be decided by the decision maker by formulating his beliefs on experts' elicited distributions conditional on the true unknown density of $\tilde{\theta}$.

The second method is the opinion pooling. It is a simple and widely used method in place of the above mentioned Bayesian approach with difficulties to implement in practice. In opinion pooling, the aggregated distribution $f(\tilde{\theta})$ is obtained as a function of experts' elicited distributions; $f_1(\tilde{\theta}), f_2(\tilde{\theta}), ..., f_q(\tilde{\theta})$. We consider linear opinion pooling, in which $f(\tilde{\theta})$ is obtained as a weighted average of experts' elicited distributions; $f(\tilde{\theta}) = \sum_{i=1}^{q} w_i f_i(\tilde{\theta})$. Note that w_i is the weight allocated to i^{th} expert elicited distribution $f_i(\tilde{\theta})$ for $i = 1, 2, 3, ..., q$, and the sum of weights; $\sum_{i=1}^{q} w_i = 1$. Suppose equal weights are allocated to all the experts' elicited distributions. Then, $w_i = \frac{1}{q}$ for all i and $f(\tilde{\theta})$ will be the simple average of $f_i(\tilde{\theta}); i = 1, 2, 3, ..., q$.

Cooke (1991) explained that the levels of calibration and informativeness of experts' elicited distributions can be different in practice. Therefore, applying equal weights

for experts' elicited distributions may not be appropriate to obtain aggregated distributions of unknown quantities in such circumstances. To this end, Cooke's protocol (Cooke, 1991) is used in expert elicitation context to derive performance based experts' weights considering the levels of calibration and informativeness of experts' elicited distributions. In Cooke's protocol, separate judgements on the uncertain quantities of interest and on some seed questions (whose true values are known to the researchers) are obtained from experts. The seed questions are selected so that they are similar as much as possible to the uncertain quantities of interest. Experts' performance on making judgments on seed questions will be used as an indicator to assess how good their judgements about the uncertain quantities of interest. It will be followed by applying a special pooling rule that is known as "the classical model" to derive experts' weights based on the performance on the seed questions (O'Hagan, 2019).

Cooke et al. (2014) and Cooke and Goossens (2008) have analysed some applications of Cooke's method and showed that weighted pooling from the classical model performs better than equal weighted pooling. Therefore, we use Cooke's weights for the analysis and study whether deriving shrinkage Cooke's weights with reduced mean squared errors has an impact on the calibration and information scores of aggregated distributions of quantities.

6.2.1 Cooke's classical model

The background on computing the calibration scores of experts' elicited percentiles of quantities using the Cooke's classical model was reviewed in section 5.2 of the preceding chapter. Now we review the background on computing the information scores of experts' elicited percentiles of quantities for deriving experts' weights due to Cooke (1991, chap. 12) as follows.

Information score

The informativeness of an expert's elicited distribution of an uncertain quantity $X(i)$ can be assessed by comparing the elicited distribution with the uniform background distribution $U(i)$ on $[X(i)_0, X(i)_n]$. The values of $X(i)_0$ and $X(i)_n$ are chosen as the lowest

$(X(i)_1)$ and highest $(X(i)_{n-1})$ elicited percentiles plus a 10% overshoot. If a realized value of the quantity $X(i)$ is known, it can also be taken into consideration. In such a situation, $l = \min(X(i)_1, \text{realized value})$ and $h = \max(X(i)_{n-1},$ realized value $)$ give $X(i)_0 = l - 10\%(h-l)$ and $X(i)_n = h + 10\%(h-l)$.

In order to avoid continuous distributions and to simplify calculations, the interval $[X(i)_0, X(i)_n]$ is assumed to be divided into a large number $\tilde{k}$ of evenly spaced points. Furthermore, we consider all mass functions on this set D_i of points. Hence, the uniform mass function $U(i)$ at point $j \in D_i$ is assumed to be given by

$$U(i)(j) = \frac{1}{\tilde{k}}. \tag{6.2}$$

It is also assumed that all expert's elicited percentiles coincide with one of these points. Note that we only know $(n-1)$ percentiles from the expert's elicited distribution of $X(i)$. Therefore, interpolation between percentiles is required to find the expert's distribution $f(i)$ of $X(i)$ to be compared with the uniform distribution $U(i)$. Following proposition due to Cooke (1991, chap. 12) is useful to find a mass function $f(i)$ which adds as little information as possible to the expert's elicited percentiles relative to the uniform mass function.

Proposition 2 *The mass function f(i) on* $D_i \subset [X(i)_0, X(i)_n]$ *which minimizes* $I(f(i), U(i))$ *and which agrees with the expert's percentiles is* :

$$f(i) = \frac{p_r}{\#\left(x \middle| X(i)_r \geq x > X(i)_{r-1}\right)} \qquad r = 1, 2, ..., n. \tag{6.3}$$

Here, the probability mass function of each cell r is estimated as a ratio between the intended theoretical probability p_r and the number of points from the set of D_i has been fallen in each cell. Note that $f(i)$ is normalized to satisfy the requirement of total mass over all points in D_i equals to one. The overall information score of an expert should be computed by combining the individual measures of informativeness

of all m elicited quantities. Thus, the information score for a given expert can be stated as

$$Inf = \sum_{i=1}^{m} \frac{I(f(i), U(i))}{m}. \tag{6.4}$$

The information score measures the degree to which experts' elicited distributions are concentrated with respect to the background uniform distributions of quantities. Thus, high values of Inf indicate that the considered expert has added more information relative to the uniform background distributions of quantities. Therefore, in comparing the informativeness of experts' elicited distributions of quantities, the highest information score implies that more concentrated distributions of quantities compared to the background uniform distributions have been elicited from the corresponding expert.

Weights of the Cooke's classical model

Now consider the computation of weights using the Cooke's classical model. According to Cooke (1991), the weights of the classical model are proportional to the product of the above discussed calibration and information scores. For $\alpha > 0$, let $1_\alpha(x) = 1$ if $x \geq \alpha$, and 0 otherwise. Then for all $\alpha > 0$, the performance based weight for a given expert can be given as

$$w = Cal \times 1_\alpha(Cal) \times Inf. \tag{6.5}$$

Observe that an imposed threshold value of α in the calibration component (Cal) through $1_\alpha(Cal)$ ensures that elicited distributions are not only concentrated but also calibrated within an acceptable margin. It emphasizes the fact that eliciting concentrated distributions alone is not useful if they are not adequately calibrated. Hence, weights will be assigned only to experts whose calibration scores exceed the threshold value α (O'Hagan et al., 2006).

It follows from the discussion in section 5.2 of the preceding chapter that the calibration component is a p-value from a chi-square distribution of which the degrees of freedom

depends on the number of elicited quantities m . Thus, it is a constant probability value. The indicator function $1_\alpha(Cal)$ produces either 1 or 0 based on a comparison between two constant values α and Cal. Therefore, it is also a constant value. In addition to that the information component is also a constant value. Therefore, overall the computed experts' weights are considered as fixed values. However, it should be noted that experts' weights can randomly vary depending on experimental conditions. Nature of seed questions, number of seed questions to be answered, time limitations, and other experimental conditions can impact on estimated weights to be varied between experiments. The seed questions that are used to derive weights are selected so that they are similar as much as possible to the uncertain quantities of interest. Suppose we repeat the experiment under similar experimental conditions with the same number of new seed questions from the same background. Note that the computed experts' weights will still be varied between experiments. Therefore, from a statistical point of view, it is important to consider experts' weights as random variables.

If experts' weights are considered as random variables, then it follows the need of estimating the mean weights of corresponding probability distributions. This forms a multivariate mean estimating problem. Stein (1956) first showed that shrinkage estimation technique can be used to obtain estimators of the mean of a multivariate normal distribution with reduced mean squared errors. Here, the interest is not to obtain an unbiased estimator but an estimator with reduced mean squared error. James-Stein shrinkage estimator discussed in James and Stein (1961) was proved to dominate the ordinary least squares estimator with lower mean squared error in this context. Therefore, we are interested to apply the James-Stein shrinkage estimation technique on Cooke's weights to obtain shrinkage Cooke's weights with reduced mean squared errors. Shrinkage estimation technique is not restricted only to the estimation of mean of a multivariate normal distribution. Since 1956, a large number of papers have been published on discussing the application of shrinkage estimation technique to obtain improved estimators of parameters for several statistical models (Voinov and Nikulin, 1995). We focus on its application to the estimation of mean of a multivariate distribution as the research interest of this study.

Section 6.3: Shrinkage estimation of weights

The idea of considering the estimated experts' weights from an experiment as random variables leads to a claim that they are a set of realized values from the corresponding random distributions of experts' weights that are created by repeating the experiment under similar experimental conditions. It follows the need of estimating the unknown population mean weights of experts using the observed set of estimated weights. Thus, if we consider $w_1, w_2, ..., w_q$ as the estimated weights for q experts and $\theta_1, \theta_2, ..., \theta_q$ as the unknown population mean weights, then the problem of interest is to estimate the vector of population mean weights; $\theta = \{\theta_1, \theta_2, ..., \theta_q\}$ using a single realization of weights; $w = \{w_1, w_2, ..., w_q\}$. Stein (1956) showed that this is a small sample situation of which, in terms of mean squared error, the usual ordinary least squares estimates of mean parameters are not suboptimal and the James-Stein shrinkage estimator (James and Stein, 1961) can dominate the ordinary least squares estimator with lower mean squared error in this context. Here, we assume a given set of raw weights from the Cooke's classical model $w = \{w_1, w_2, ..., w_q\}$ as a single realization from a q-variate random variable $W = \{W_1, W_2, W_3, \ldots, W_q\}$ and proceed into deriving James-Stein shrinkage weights. We now review James-Stein shrinkage estimator in general due to James and Stein (1961) as follows.

Suppose X is a $q-$variate normally distributed random variable with a vector θ of unknown means and a known covariance matrix $\sigma^2 I$. Here, I is the $(q \times q)$ identity matrix and σ^2 is the assumed constant variance for q variables of X. It follows that $X \sim N(\theta, \sigma^2 I)$. Now consider a situation that requires to obtain an estimate $\hat{\theta}$ of θ using a $q-$variate single observation of X. Note that $\hat{\theta}_{LS} = X$ is the ordinary least square (OLS) estimator of θ in this situation considering the observations as estimates of θ themselves. This estimator is suboptimal in terms of mean squared error. It led to the development of the following James-Stein (JS) shrinkage estimator for θ by shrinking the OLS estimator towards the origin 0. JS estimator $(\hat{\theta}_{JS})$ for known σ^2 is given by

$$\hat{\theta}_{JS_1} = \Big(1 - \frac{(q-2)\sigma^2}{\| X \|^2}\Big)X, \tag{6.6}$$

and it dominates $\hat{\theta}_{LS}$ in terms of lower mean squared error for any $q \geq 3$.

If σ^2 is unknown and an estimator $S_{\tilde{n}}$ of σ^2 independent of X and distributed as $\sigma^2\chi^2_{\tilde{n}}$ is available, then JS estimator of θ is given by

$$\hat{\theta}_{JS_2} = \Big(1 - \frac{(q-2)S_{\tilde{n}}}{(\tilde{n}+2)\| X \|^2}\Big)X. \tag{6.7}$$

In the most realistic situation $X \sim N(\theta, \Sigma)$, Σ being unknown. Suppose there is an independent estimator S of Σ which is distributed as $W_{\tilde{n}-1}(s, \Sigma)$, a Wishart distribution. James & Stein proposed the following JS estimator in this situation

$$\hat{\theta}_{JS_3} = \Big(1 - \frac{(q-2)}{(\tilde{n}-q+3)X^T S^{-1} X}\Big)X. \tag{6.8}$$

If we review the above discussed JS estimators in equations; 6.6, 6.7, and 6.8, it can be identified that shrinkage estimators are obtained by multiplying the original observations by a constant value. Thus, if we apply this technique on a set of derived experts' weights, then the resulting shrinkage weights will be proportional to the original weights. Thus, normalized weights before and after shrinkage will be identical.

Baranchik (1964) first introduced the concept that the positive part estimators in the form of $(x)_+ = \text{maximum}(0, x)$ can dominate James-Stein estimators. Here, $+$ indicates the positive part estimator in standard notation. Therefore, we can define the positive part James-Stein estimators of the above discussed original definitions as

$$\hat{\theta}_{JS_1^+} = \Big(1 - \frac{(q-2)\sigma^2}{\| X \|^2}\Big)_+ X, \tag{6.9}$$

$$\hat{\theta}_{JS_2^+} = \Big(1 - \frac{(q-2)S_{\tilde{n}}}{(\tilde{n}+2)\| X \|^2}\Big)_+ X, \text{ and} \tag{6.10}$$

$$\hat{\theta}_{JS_3^+} = \Big(1 - \frac{(q-2)}{(\tilde{n}-q+3)X^T S^{-1} X}\Big)_+ X, \text{ respectively.} \tag{6.11}$$

Observe that positive part James-Stein estimators do not overcome the above discussed issue of producing identical normalized weights before and after shrinkage in our

context. Therefore, we need to employ a shrinkage procedure that shrinks weights differently depending on a suitable factor. To this end, from a statistical point of view, it seems useful to shrink weights differently depending on their variances as a measure of uncertainty.

Now consider the following empirical Bayes approach of obtaining shrinkage estimators of multivariate mean discussed in Efron and Morris (1975). Suppose $X_i|\theta_i \sim N(\theta_i, \sigma_i^2)$, independently, $i = 1, 2, 3, ..., q$, where σ_i^2 are known, but are different from one another. Also assume that $\theta_i \sim N(0, \sigma_\theta^2)$, independently, $i = 1, 2, 3, ..., q$, where σ_θ^2 is an unknown constant. It follows that

$$\theta_i|X_i \sim N((1 - B_i)X_i, (1 - B_i)\sigma_i^2);\ i = 1, 2, 3, ..., q, \tag{6.12}$$

where $B_i = \sigma_i^2/(\sigma_\theta^2 + \sigma_i^2)$. Here, the empirical Bayes shrinkage estimator of θ_i is the posterior mean $E(\theta_i|X_i) = (1 - B_i)X_i$ with the Bayes risk $V(\theta_i|X_i) = (1 - B_i)\sigma_i^2$ being less than the risk σ_i^2 of the least square estimator $\hat{\theta}_i = X_i$. The shrinkage factor $1 - B_i = \sigma_\theta^2/(\sigma_\theta^2 + \sigma_i^2)$ avoids the above discussed problem of shrinking towards the origin by the same factor. Here, larger σ_i^2 is the more shrinkage there should be.

Efron and Morris (1975) recommended to use the estimated shrinkage factor $\hat{\sigma}_\theta^2/(\hat{\sigma}_\theta^2 + \sigma_i^2)$ discussed in Efron and Morris (1973) that reduces to the Stein's rule when all σ_i^2 are equal (refer Efron and Morris (1973) for the details of deriving an estimator for unknown σ_θ^2). We consider important to have different variances of random variables to obtain different shrinkage factors between variables in this model. However, practically they are unknown. Furthermore, mean of the distribution of θ_i is taken equal to a specific value zero. Therefore, we are interested to a find a more general Bayes approach that deals with unknown mean and variance of the distribution of θ_i and different and unknown variances of q-variables for the analysis.

Zhao (2010) discussed an empirical Bayes approach of obtaining shrinkage estimators of multivariate mean with unknown and unequal variances of q-variables. The proposed estimator shrinks not only means but also variances as well. Therefore, it is called the double shrinkage estimator. According to the author, extensive numerical studies indicate that the double shrinkage estimator has lower Bayes risk than the shrinkage

estimator of means alone and the naive estimator with no shrinkage at all. In this approach, each variable $X_i (i = 1, 2, 3, ..., q)$ is assumed to follow a normal distribution with mean θ_i and unknown variance σ_i^2 which differ across all the variables and each $\theta_i (i = 1, 2, 3, ..., q)$ is assumed to follow a common prior distribution $N(\mu, \tau^2)$ with unknown mean μ and variance τ^2.

Suppose we derive the posterior distribution of θ_i using the model; $X_i|\theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2)$ and $\theta_i \sim N(\mu, \tau^2)$ for $i = 1, 2, 3, ..., q$ assuming all the parameters of the model are known. Then, the posterior distribution of θ_i will be derived as

$$\theta_i|X_i, \sigma_i^2 \sim N(M_i X_i + (1 - M_i)\mu, M_i\sigma_i^2), \tag{6.13}$$

where $M_i = \tau^2/(\tau^2 + \sigma_i^2)$. Therefore, for the known σ_i^2 case, the estimator for θ_i is $M_i X_i + (1 - M_i)\mu$, which is the posterior expectation of θ_i given X_i and σ_i^2. This estimator shrinks X_i towards mean μ and the shrinkage factor $M_i = \tau^2/(\tau^2 + \sigma_i^2)$ depends on the variance σ_i^2 of X_i. However, σ_i^2, μ and τ^2 are assumed unknown in this context. Therefore, Zhao (2010) derived the following double shrinkage estimator for θ in the form of

$$\hat{\theta}_i = \hat{M}_i X_i + (1 - \hat{M}_i)\hat{\mu} \tag{6.14}$$

with $\hat{M}_i = \hat{\tau}^2/(\hat{\tau}^2 + \hat{\sigma}_i^2)$. We review the procedure of deriving the double shrinkage estimator due to Zhao (2010) as follows.

Assume that there exists a statistic S_i^2 independent of X_i that contains information of σ_i^2. It follows to assume in general that $S_i^2|\sigma_i^2 \sim \sigma_i^2 \frac{\chi^2_{\tilde{d}_i}}{\tilde{d}_i}$ where $\tilde{d}_i$ represents the degrees of freedom corresponding to the i^{th} statistic S_i^2. Now approximate $\log\left(\frac{\chi^2_{\tilde{d}_i}}{\tilde{d}_i}\right)$ to follow $N(\tilde{m}_i, \sigma^2_{ch,i})$ distribution with mean $\tilde{m}_i = E\left(\log\left(\frac{\chi^2_{\tilde{d}_i}}{\tilde{d}_i}\right)\right) = \psi(\frac{\tilde{d}_i}{2}) - \log(\frac{\tilde{d}_i}{2})$ and variance $\sigma^2_{ch,i} = V\left(\log\frac{\chi^2_{\tilde{d}_i}}{\tilde{d}_i}\right) = \frac{d}{dx}\psi(\frac{\tilde{d}_i}{2})$, where $\psi(x) = \frac{d}{dx}\log(\Gamma(x))$ is known as the digamma function. Assumption of $S_i^2|\sigma_i^2 \sim \sigma_i^2 \frac{\chi^2_{\tilde{d}_i}}{\tilde{d}_i}$ and the approximation of $\log\left(\frac{\chi^2_{\tilde{d}_i}}{\tilde{d}_i}\right)$ to follow $N(\tilde{m}_i, \sigma^2_{ch,i})$ distribution follows that

$$\log(S_i^2)|\log(\sigma_i^2) \sim N(\tilde{m}_i + \log(\sigma_i^2), \sigma^2_{ch,i}). \tag{6.15}$$

Furthermore, this model assumes that $\log(\sigma_i^2)$ is a normal random variable with unknown mean μ_v and variance τ_v^2. Thus,

$$\log(\sigma_i^2) \sim N(\mu_v, \tau_v^2). \tag{6.16}$$

Combining the information from equations 6.15 and 6.16 provides the following equation for $\log(\sigma_i^2)|\log(S_i^2)$ as

$$\log(\sigma_i^2)|\log(S_i^2) \sim N\Big(M_{v,i}(\log(S_i^2) - \tilde{m}_i) + (1 - M_{v,i})\mu_v,\ M_{v,i}\,\sigma_{ch,i}^2\Big), \tag{6.17}$$

where $M_{v,i} = \tau_v^2/(\tau_v^2 + \sigma_{ch,i}^2)$. Thus, the shrinkage variance estimate of σ_i^2 can be obtained as the posterior mean from equation 6.17 as

$$\sigma_i^2 = \exp\Big(M_{v,i}(\log(S_i^2) - \tilde{m}_i) + (1 - M_{v,i})\mu_v\Big) \tag{6.18}$$

under the assumption that both μ_v and τ_v^2 are known. However, μ_v and τ_v^2 are assumed unknown in the model.

Now focus on obtaining an empirical Bayes estimator $\hat{\sigma}_i^2$ using the estimated $\hat{\mu}_v$ and $\hat{\tau}_v^2$ from the data. It can be obtained from equation 6.15 that

$$\log(S_i^2) - \tilde{m}_i|\log(\sigma_i^2) \sim N(\log(\sigma_i^2), \sigma_{ch,i}^2). \tag{6.19}$$

It is known from equation 6.16 that

$$\log(\sigma_i^2) \sim N(\mu_v, \tau_v^2). \tag{6.20}$$

Thus, using conditional expectation and conditional variance, it can be shown that $E(\log(S_i^2) - \tilde{m}_i) = \mu_v$ and $E(\log(S_i^2) - \tilde{m}_i)^2 = \mu_v^2 + \tau_v^2 + \sigma_{ch,i}^2$. Hence, the model estimates μ_v by $\hat{\mu}_v = \frac{1}{q}\sum_{i=1}^{q}(\log(S_i^2) - \tilde{m}_i)$ and τ_v^2 by

$\hat{\tau}_v^2 = \Big(\frac{1}{q}\Big(\sum_{i=1}^{q}(\log(S_i^2) - \tilde{m}_i)^2 - \hat{\mu}_v^2 - \sigma_{ch,i}^2\Big)\Big)_+$. Hence, $M_{v,i}$ can be estimated as $\hat{M}_{v,i} = \hat{\tau}_v^2/(\hat{\tau}_v^2 + \sigma_{ch,i}^2)$ and the empirical Bayes estimator of σ_i^2 can be derived as

$$\hat{\sigma}_i^2 = \exp\Big(\hat{M}_{v,i}(\log(S_i^2) - \tilde{m}_i) + (1 - \hat{M}_{v,i})\hat{\mu}_v\Big). \tag{6.21}$$

Next step is to estimate μ and τ^2 of the assumed distribution of θ_i for $i = 1, 2, 3, ..., q$. Revise that the model begins with assuming $X_i|\theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2)$ and $\theta_i \sim N(\mu, \tau^2)$ for $i = 1, 2, 3, ..., q$. It can be shown following the same procedure discussed above for estimating μ_v and τ_v^2 that

$$\mathrm{E}(X_i|\sigma_i^2) = \mu \text{ and} \tag{6.22}$$

$$\mathrm{E}(X_i - \mu)^2|\sigma_i^2 = \sigma_i^2 + \tau^2. \tag{6.23}$$

From equation 6.22, μ can be estimated by the weighted average as $\hat{\mu} = \sum_{i=1}^{q} \frac{Xi/\hat{\sigma}_i^2}{\sum_{i=1}^{q} 1/\hat{\sigma}_i^2}$. Equation 6.23 can be used to estimate τ^2 as $\hat{\tau}^2 = \Big(\frac{\sum_{i=1}^{q}(X_i - \hat{\mu})^2 - \hat{\sigma}_i^2}{q}\Big)_+$. According to Zhao (2010), this estimator $\hat{\tau}^2$ can be inconsistent for τ^2 as $q \to \infty$. Therefore, following estimator $\hat{\tau}^2 = \Big(\frac{\sum_{i=1}^{q}(X_i - \hat{\mu})^2 - S_i^2\exp(-\tilde{m}_i - \sigma_{ch,i}^2/2)}{q}\Big)_+$ is suggested in the model. Two estimators of $\hat{\tau}^2$ and $\hat{\sigma}_i^2$ are used to derive double shrinkage estimator of θ_i in the equation 6.14 above. This estimator shrinks X_i towards the weighted average $\hat{\mu}$. Additionally, the estimator $\hat{\sigma}_i^2$ is a variance shrinkage estimator. It shrinks observation $S_i^2/\exp(\tilde{m}_i); i = 1, 2, 3, ..., q$ towards their geometric mean. Therefore, this empirical Bayes shrinkage estimator is called the double shrinkage estimator. We will use this approach to derive shrinkage estimators of experts' weights in this analysis.

Section 6.4: Methodology of the study

It was discussed in the section 6.2.1 above that the Cooke's model imposes a threshold value to select experts whose calibration scores exceed a certain cut-off value for allocating weights. However, following the analysis of the previous chapter, it can be suggested that this weight optimization procedure does not guarantee that weights are not allocated to not well-calibrated experts. Therefore, we recommend to apply

the multinomial equivalence test to provide a safe guard for allocating weights for not well-calibrated experts by selecting a set of experts whose true probability vectors of elicited percentiles remain within an acceptable margin of deviation from the intended probability vector of elicited percentiles. It will be followed by applying the Cookes' classical model to derive weights for the selected set of experts without imposing the weight optimization as the selected experts' true probability vectors of elicited percentiles are within an acceptable margin of deviation from the intended probability vector of elicited percentiles. Here, we derive shrinkage weights from normalized Cooke's weights without applying the optimization with the intention of further improving the accuracy of deriving aggregated distributions of quantities.

The expert judgment data base that is maintained by the Delft University of Technology, Netherlands (Cooke and Goossens, 2008) will be used to select suitable data sets for the analysis. It was observed that most of the data sets in Delft data base contain around 10 calibration or seed questions with known realized values to derive experts' weights. We focus on deriving weights with usual 10 seed questions in this analysis considering the potential difficulties of using more seed questions to derive weights in practice. In order to assess the performance of the derived weights, it is also required to have a set of testing questions with known realized values. Hence, some suitable data sets with more seed questions will be selected and the first 10 questions of each data set will be used to derive normalized Cooke's weights. The remaining questions of each data set will be used to assess the performance of derived weights before and after applying the shrinkage technique discussed above. We will refer first 10 questions as training questions and the rest as testing questions for the analysis.

If we consider applying the above discussed empirical Bayes shrinkage procedure due to Zhao (2010) to derive shrinkage weights, it is required to obtain a statistic S_i^2 independent of the ith normalized Cooke's weight w_i derived from the training questions that contains the information about the variance σ_i^2 of w_i. Therefore, we will obtain 10 random samples of 10 questions each from the testing questions to obtain $S_i^2 = \frac{1}{9}\sum_{i=1}^{10}(w_i - \overline{w})^2$ as the sample variance of weights, where w_i is the normalized Cooke's weight derived from the ith sample; $i = 1, 2, 3, ..., 10$, and $\overline{w} = \frac{1}{10}\sum_{i=1}^{10} w_i$. Therefore, the degrees of freedom $\tilde{d}_i$ will be equal to 9 for each statistic S_i^2 for

$i = 1, 2, 3, ..., q$; where q is the number of derived experts' weights in the analysis.

The user define weights option of the Excalibur package (Cooke and Solomatine, 1992) can be applied with an assigned set of experts' weights to compute the overall calibration and information scores of any given set of questions with experts' elicited percentiles and the known realized values of questions. Therefore, the impact of deriving shrinkage weights can be assessed by comparing the overall calibration and information scores of testing questions computed using the normalized typical and shrinkage Cooke's weights. Overall, the steps of the analysis can be stated as

1. Select some data sets with more seed questions.
2. Consider the first 10 seed questions of each data set as training questions for deriving normalized Cooke's weights.
3. Consider the remaining questions of each data set as testing questions of the analysis.
4. Estimate the sample variances of normalized Cooke's weights using 10 randomly selected samples of 10 questions each from the testing questions.
5. Derive shrinkage weights from the normalized Cooke's weights derived from the training questions using the above estimated sample variances of normalized Cooke's weights.
6. Obtain normalized shrinkage weights.
7. Compute the overall calibration and information scores of testing questions using the normalized typical and shrinkage Cooke's weights by applying the user define weights option of the Excalibur package as discussed above.
8. Compare the overall calibration and information scores above to assess the impact of deriving shrinkage weights.

Section 6.5: The data

We used two data sets; $PBINTDOS$ and $RETURNafter$, from the Delft data base with reasonably large number of seed questions to perform the analysis. Data contain

experts' elicited 5% , 50%, and 95% percentiles of the distributions of quantities of some given questions in both data sets. The experts who have answered majority of seed questions were selected to the analysis from each data set. Data from the first 35 seed questions from the 5 experts with their identity numbers from 2-6 were gathered from the $PBINTDOS$ data set. Data from all 31 seed questions and all 5 experts were gathered from the $RETURNafter$ data set. First 10 questions from each data set were used to derive weights and the remaining questions were used to estimate variances of weights and to compare the overall calibration and information scores of typical and shrinkage weights as discussed above.

Section 6.6: Analysis of data

The Excalibur package was used to obtain the given outputs of derived Cooke's weights without Decision Maker (DM) optimization from the training data in Appendix K. The Output 1 and Output 2 provide details about the derived weights from $PBINTDOS$ and $RETURNafter$ data, respectively. The normalized Cooke's weights are given under the name of " Normaliz.w - without DM" in each output. Here, the decision maker indicates a combination of experts' assessments (Colson and Cooke, 2018). We mentioned above that the optimization of the decision maker through imposing a threshold value to select experts whose calibration scores exceed a certain cut-off value for allocating weights will not be applied in this analysis. Therefore, the decision maker assigns each expert a constant weight depending on the experts' average calibration and information scores over all seed questions.

Now consider the analysis of $PBINTDOS$ data. The normalized Cooke's weights from Output 1 in Appendix K were used to derive shrinkage weights of the analysis. In order to estimate the sample variances of weights, we randomly selected 10 samples of 10 questions each from the testing data and estimated the normalized Cooke's weights for each expert separately for each sample using the Excalibur package. The resulting weights were used to compute the empirical Bayes shrinkage weights of the analysis.

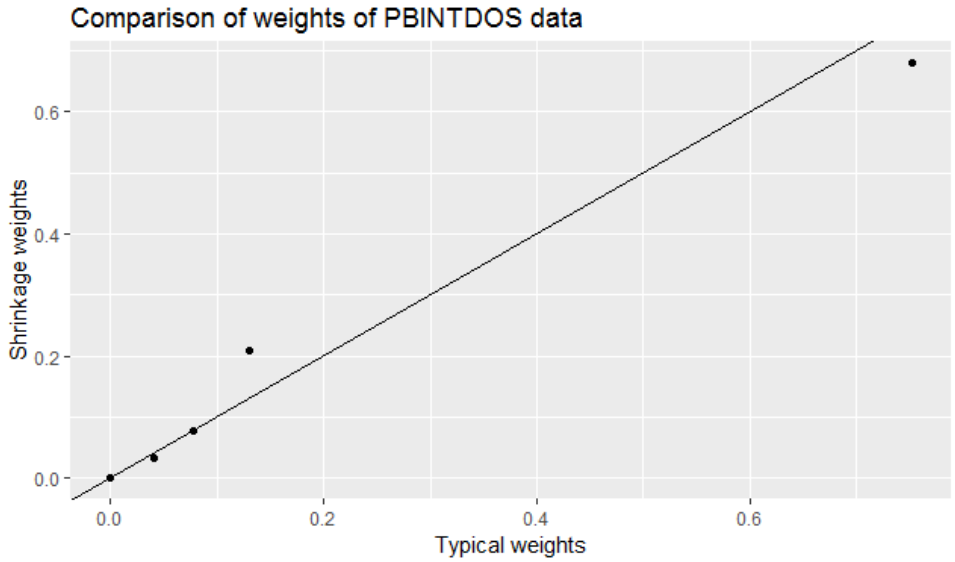


FIGURE 6.1: Normalized typical and shrinkage Cooke's weights with added x=y line from PBINTDOS training data

It was mentioned in section 6.2.1 that the seed questions to derive weights are selected so that they are as similar as possible to the uncertain quantities of interest in practice. If we refer the derived Cooke's weights for experts from different random samples of size 10 from the testing questions (PBINTDOS_sample _weights.csv) given in Appendix L, it can be noted that experts' weights will be varied randomly even they are derived from similar number of questions from a given background. Therefore, it is important to incorporate this underlying uncertainty into the process of deriving weights. Here, we implement this using the empirical Bayes shrinkage approach discussed above.

TABLE 6.1: Overall Decision Maker scores of testing questions

Data set	Types of weight	Calibration score	Information score
PBINTDOS	Typical	0.7496	1.044
	Shrinkage	0.7587	1.077
RETURNafter	Typical	0.01487	0.2433
	Shrinkage	0.004452	0.2837

Figure 6.1 compares the normalized typical and shrinkage Cooke's weights from PBINTDOS training data. Observe that there are some differences between weights for larger typical Cooke's weights. Table 6.1 indicates that the overall Decision Maker calibration and information scores are slightly higher for the shrinkage weights compared to the typical weights. Here, decision maker represents the overall calibration

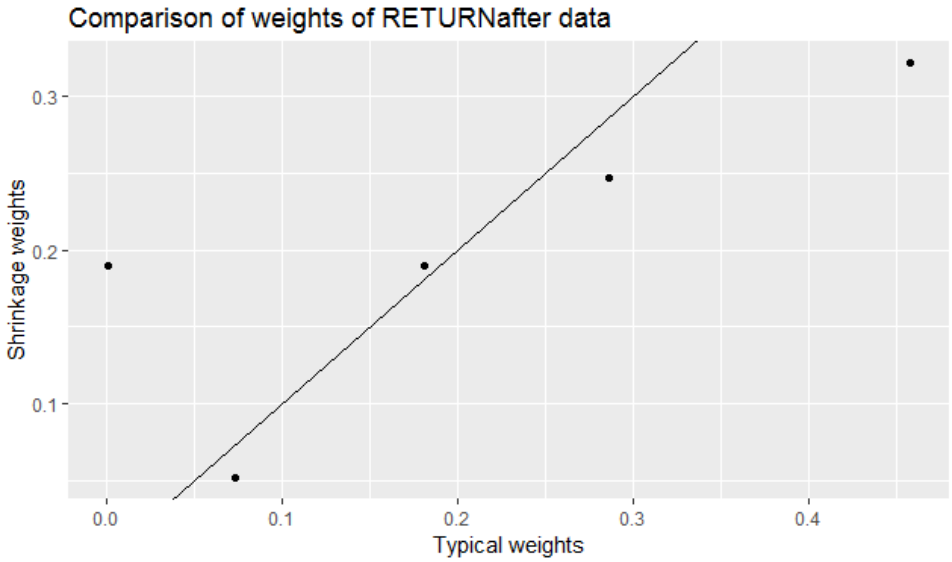


FIGURE 6.2: Normalized typical and shrinkage Cooke's weights with added x=y line from RETURNafter training data

and information scores of testing questions that are computed using the user define experts' weights for the typical and shrinkage weights. The results of the analysis of the PBINTDOS testing data using the user define normalized typical and shrinkage Cooke's weights are given in Output 1 and Output 2 in Appendix M, respectively.

The analysis was also carried out following the same procedure for the $RETURNafter$ data set. Figure 6.2 shows that the typical and shrinkage weights are varied for both larger and smaller typical weights. Table 6.1 indicates that shrinkage weights have produced a lower Decision Maker calibration score and a higher Decision Maker information score compared to the typical weights. Similar to the analysis of PBINTDOS data, the results from analysing the RETURNafter testing data using the user define normalized typical and shrinkage Cooke's weights are given in Output 1 and Output 2 in Appendix N, respectively.

The above results show that the shrinkage estimators of experts' weights have not produced a general result of either increasing or decreasing both of the overall calibration and information scores of quantities. Here, we do not focus on the direction of change of scores. We emphasize the importance of reducing the allocated weights for experts with larger variances of weights to enhance the precision of the derived aggregated distributions of quantities. Therefore, the accuracy of the obtained results can expected to be improved by applying the shrinkage experts' weights.

Section 6.7: Discussion

If we consider the context of deriving experts' weights, it can be observed that small number of seed questions are used in general. We acknowledge the potential difficulties of using more seed questions that are similar as much as possible to the uncertain quantities of interest in practice. Furthermore, the outcomes of the seed questions should not be known by the experts. Derivation of experts' weights with limited amount of information due to lack of seed questions can have an impact on experts' weights to be varied considerably from their unknown mean values of the random distributions of weights. Therefore, there is a possibility to have large variances of weights with small number of seed questions. We consider this as an important aspect for deriving shrinkage weights for possible reduction of the mean squared errors of estimated mean weights when limited number of seed questions are used to derive weights.

Sample variances of experts' weights were estimated using random samples of questions obtained from the testing questions as explained in section 6.4. Computation of weights from random samples to estimate the sample variances of weights was carried out manually by creating separate Excalibur data files with randomly selected questions. It caused practical difficulties to obtain more samples and to restrict our attention to data sets with small number of experts in this analysis. We acknowledge the fact that additional set of seed questions are not available to estimate the sample variances of weights to derive shrinkage weights in practice. To this end, we suggest to apply the non-parametric jackknife resampling method discussed in Efron and Stein (1981) for estimating the variances of experts' weights using a given set of seed questions by following a resampling technique.

Section 6.8: Conclusion

Experts' weights are derived from experiments to obtain aggregated probability distributions of unknown quantities by combining the experts' elicited probability distributions of corresponding quantities together with suitable weights for experts.

Therefore, they are random variables subject to uncertainty. The focus of this analysis is to address this underlying uncertainty of the derived experts' weights from experiments in computing aggregated distributions of quantities. James-Stein shrinkage estimation technique discussed in James and Stein (1961) can be used to estimate the mean of a multivariate normal distribution with reduced mean squared errors. We applied an empirical Bayes development of the James-Stein shrinkage estimation technique discussed in Zhao (2010) that shrinks variables differently depending on their variances to derive weights in this analysis. Hence, more shrinkage will be applied to derived weights with larger variances.

The results of the analysis show that overall Decision Maker (DM) calibration and information scores of some selected testing questions with known realized values are different between the normalized typical and shrinkage Cooke's weights for couple of data sets that have been chosen from the Delft data base (Cooke and Goossens, 2008). However, there is no evidence to identify a general pattern of either increasing or decreasing both of the overall calibration and information scores by employing shrinkage weights from the analysis. From a statistical point of view, we cannot expect to change the calibration and information scores in a particular direction by obtaining shrinkage estimates of weights. However, considering the purpose of applying the shrinkage estimation technique to reduce the mean squared errors of estimators of the mean of a multivariate normal distribution due to James and Stein (1961) and the procedure of the employed empirical Bayes shrinkage approach to derive weights by reducing the allocated weights for experts with larger variances of weights in Zhao (2010), it can be suggested that the outcome of the shrinkage weights can expect to be more accurate than the typically computed weights.

Following the results of the analysis, we recommend to carry out cross validation studies for data sets with more seed questions to further ensure the effectiveness of deriving shrinkage weights as a future research direction. Here, we suggest to consider the potential of deriving normalized original and shrinkage Cooke's weights from different number of seed questions to predict the distributions of some seed questions with known realized values.

7. Conclusions and potential future directions

It is important to note that different chapters have different research objectives under an overall research title of exploring the statistical aspects of expert elicited experiments in this study. Therefore, we provide a chapter-wise summary of conclusions and identified potential future directions of the study as follows.

Section 7.1: Chapter 2 - Aligning the analysis and the design of expert elicitation experiments

Experts' Brier scores are derived from expert elicited experiments to assess the prediction accuracy of experts on predicting probabilities of the occurrence of events. The focus of the analysis of this chapter is to improve the estimation of standard errors of experts' Brier scores by incorporating the potential correlation structure induced in the probability predictions by asking common questions from experts. It can be concluded based on the results of the analysis that the standard error estimates of experts' Brier scores from the fitted linear mixed-effects model with questions' effects as random effects are more accurate than the typical standard error estimates of experts' Brier scores that are computed by ignoring the potential correlations between the probability predictions due to common questions' effects.

7.1.1 Future directions

Mixed-effects models are developed to deal with multi-level of hierarchical data structures in a wide variety of experimental designs. Therefore, it is useful to extend the suggested mixed-effects model approach of computing experts' Brier scores and

their standard error estimates to hierarchical experimental designs with multiple layers in future studies. Here, the models can be extended by including necessary random effects to incorporate potential design-based correlated sources of predictions. Furthermore, extending this approach to compute Logarithmic and Spherical scores and their standard error estimates can also be useful to the expert elicitation context.

Section 7.2: Chapter 3 - Missing values, and ways of dealing with them

The focus of the analysis of this chapter is to show that estimating missing values of probability predictions of questions can be useful to enhance the comparability of experts' Brier scores to assess the prediction accuracy of experts. Experts' Brier scores that are computed without estimating missing values of probability predictions of questions are not adjusted for the effects of missing questions for some experts in a given analysis. Therefore, the comparisons between experts' Brier scores may not be accurate enough in such situations. We employed some selected missing value estimation methods to estimate missing probability predictions of questions in computing experts' Brier scores and found that multiple imputation method using a mixed-effects model with questions' effects as random effects can estimate missing probability predictions to compute experts' Brier scores with reduced errors compared to the typically computed Brier scores that ignore missing predictions.

It is important to note that the suggested multiple imputation method of estimating missing predictions uses all the available predictions to estimate missing predictions and compute a new set of experts' Brier scores with estimated missing predictions. This is more effective than throwing away some collected data to compute Brier scores using same sets of predictions for all the experts when there are missing predictions by experts. Therefore, we recommend to apply multiple imputation method using mixed-effects models with necessary random effects to bringing in any design-based correlated sources of predictions to compute Brier scores with estimated missing predictions.

7.2.1 Future directions

This simulation study was performed based on a specific data set. Therefore, there is a possibility that some specific characteristics of the data set may have caused multiple imputation method to work well. Therefore, it is an interesting future research direction to theoretically prove the ability of the multiple imputation method with relevant mixed-effects models to better estimate missing probability predictions for computing experts' Brier scores over the other considered methods under some specific conditions.

Section 7.3: Chapter 4 - Testing experts' calibration I

The focus of the analysis of this chapter is to assess the properties of testing experts' calibration on eliciting credible intervals for unknown quantities using the direct comparison of experts' hit rates; observed proportions of experts' elicited intervals that contain realized values of given quantities (McBride, Fidler, and Burgman, 2012), with the level of intended coverage probability of elicited credible intervals. The results of the conducted simulation study at some selected standard levels of intended coverage probabilities show that the direct comparison of hit rates has a property of obtaining lower values of power to correctly identify well-calibrated experts and more importantly, the power tends to decrease as the number of elicited intervals increases. This is a contradictory result from a statistical point of view as the power of the test to correctly identify well-calibrated experts can be expected to increase with increased number of elicited intervals.

We explored the potential of using the equivalence test of a single binomial proportion to overcome the above identified problems of the direct comparison of hit rates. The values of power of the equivalence test to correctly identify well-calibrated experts tend to increase as intuitively expected over the increase of number of elicited intervals and they are comparatively higher than the corresponding values of the direct test for large number of elicited intervals. Therefore, this avoids the need of performing the analysis with small number of elicited intervals with lower values of power of the direct comparison of hit rates. Hence, we recommend to apply the equivalence test of

a single binomial proportion with large number of elicited intervals to test experts' calibration in this context.

It was also observed that the power of the direct comparison of hit rates and equivalence test of a single binomial proportion to correctly identify well-calibrated experts tends to increase with increased levels of intended coverage probabilities. Therefore, we suggest to look into possibilities of optimizing the power to correctly identify well-calibrated experts by increasing the levels of intended coverage probabilities of elicited credible intervals of quantities, appropriately.

7.3.1 Future directions

The results of this simulation study show that a very large number of elicited intervals is required to obtain a reasonably large value of power to correctly identify well-calibrated experts from the equivalence test of a single binomial proportion. It is practically difficult to conduct elicitation experiments in which a large number of interval judgements are obtained from experts. Therefore, more reacher is needed before we understand what method strikes the right balance between practicability and the desired statistical power.

Section 7.4: Chapter 5 - Testing experts' calibration II

Experts' weights are derived to combine experts' elicited subjective probability distributions to obtain aggregated probability distributions of unknown quantities (O'Hagan, 2019). Assessing the levels of experts' calibration on eliciting a specified number of percentiles from the probability distributions of quantities of some given seed questions (with known realized values of quantities to the researchers but not to the experts) using the calibration score component is a part of deriving experts' weights using the Cooke's classical model (Cooke, 1991). We refer the application of the calibration score component to assess the levels of experts' calibration as the calibration test in this analysis. Here, the focus of the analysis is to explore the properties of the calibration test in assessing the levels of experts' calibration on eliciting a specified number of percentiles from the probability distributions of quantities in deriving experts' weights.

The results of a conducted simulation study at some commonly specified number of percentiles to be elicited show that the calibration test fails to detect not well-calibrated experts with adequately higher values of power even for reasonably large number of elicited quantities. Furthermore, average calibration scores are reasonably high when the calibration test fails to detect not well-calibrated experts. It indicates a possibility of allocating higher weights to some of the not well-calibrated experts from the process if they have also obtained higher information scores. Thus, it is important to provide a safe guard for allocating higher weights to not-well calibrated experts, even though they have obtained higher information scores.

It is evident from the analysis that the multinomial equivalence test with the null hypothesis of a given expert is not well-calibrated against the alternative hypothesis of a given expert is well-calibrated can be applied to produce lower probabilities of identifying a not well-calibrated expert as well-calibrated with reasonably large number of elicited intervals that sufficiently satisfy the large sample approximation of the test. Therefore, once the null hypothesis of the multinomial equivalence test is rejected, it can be assumed with higher confidence that the considered expert is well-calibrated. Hence, we recommend to apply the multinomial equivalence test to identify well-calibrated experts and proceed to compute experts' weights using the Cooke's classical model to reduce the potential risk of allocating higher weights to not well-calibrated experts.

Section 7.5: Chapter 6 - Different way of deriving experts' weights

Experts' weights are random variables subject to uncertainty as they are derived from experiments. The focus of this analysis is to address this underlying uncertainty of the derived experts' weights from experiments in computing aggregated distributions of quantities. James-Stein shrinkage estimation technique discussed in James and Stein (1961) can be used to estimate the mean of a multivariate normal distribution with reduced mean squared errors. We applied an empirical Bayes development of the James-Stein shrinkage estimation technique (James and Stein, 1961) discussed in

Zhao (2010) that shrinks variables differently depending on their variances to derive weights in this analysis. Hence, more shrinkage will be applied to derived weights with larger variances.

The results of the analysis show that overall Decision Maker (DM) calibration and information scores of some selected testing questions with known realized values are different between the normalized typical and shrinkage Cooke's weights with no evidence to identify a general pattern of either increasing or decreasing both of the overall calibration and information scores by employing shrinkage weights. From a statistical point of view, we cannot expect to change the calibration and information scores in a particular direction by obtaining shrinkage estimates of weights. However, considering the purpose of applying the shrinkage estimation technique to reduce the mean squared errors of estimators of the mean of a multivariate normal distribution due to James and Stein (1961) and the procedure of the employed empirical Bayes shrinkage approach to derive weights by reducing the allocated weights for experts with larger variances of weights in Zhao (2010), it can be suggested that the outcome of the shrinkage weights can expect to be more accurate than the typically computed weights.

7.5.1 Future directions

Following the results of the analysis, we suggest to carry out cross validation studies for data sets with more seed questions to further ensure the effectiveness of deriving shrinkage weights as a future research direction. Here, we propose to consider the potential of deriving normalized original and shrinkage Cooke's weights from different number of seed questions to predict the distributions of some seed questions with known realized values. Then, the accuracy of the predictions can be compared between two types of weights.

A. Typical computation of Brier scores

TABLE A.1: Computed Brier scores and their standard error estimates

ParticipantId	Brier score	Standard error estimate
3	0.12146	0.07384
4	0.14521	0.09081
5	0.10792	0.05157
20	0.14521	0.03818
21	0.12604	0.03704
22	0.08108	0.02768
23	0.06542	0.02795
27	0.19511	0.08302
28	0.07676	0.03567
29	0.09563	0.04513
33	0.19021	0.09252
34	0.13157	0.08843
35	0.15042	0.06480
36	0.07918	0.07492
38	0.15813	0.07538
61	0.09862	0.05413

B. Comparison of standard error estimates (linear models)

Table B.1: Standard error estimates of Brier scores

ParticipantId	Typical computation	Linear_c	Linear_nc
3	0.07384	0.06410	0.07384
4	0.09081	0.06410	0.09081
5	0.05157	0.06410	0.05157
20	0.03818	0.06410	0.03818
21	0.03704	0.06410	0.03704
22	0.02768	0.06410	0.02768
23	0.02795	0.06410	0.02795
27	0.08302	0.06410	0.08302
28	0.03567	0.06410	0.03567
29	0.04513	0.06410	0.04513
33	0.09252	0.06410	0.09252
34	0.08843	0.06410	0.08843
35	0.06480	0.06410	0.06480
36	0.07492	0.06410	0.07492
38	0.07538	0.06410	0.07538
61	0.05413	0.06410	0.05413

C. Comparison of standard error estimates (*Mixed_que* model)

Table C.1: Standard error estimates of Brier scores

ParticipantId	Typical computation	Mixed_que
3	0.07384	0.06164
4	0.09081	0.06898
5	0.05157	0.06007
20	0.03818	0.03982
21	0.03704	0.04913
22	0.02768	0.04790
23	0.02795	0.04115
27	0.08302	0.07093
28	0.03567	0.05215
29	0.04513	0.05810
33	0.09252	0.07763
34	0.08843	0.06665
35	0.06480	0.05209
36	0.07492	0.06122
38	0.07538	0.05664
61	0.05413	0.04993

D. Comparison of root mean squared errors (RMSE) of Brier scores (missing completely at random)

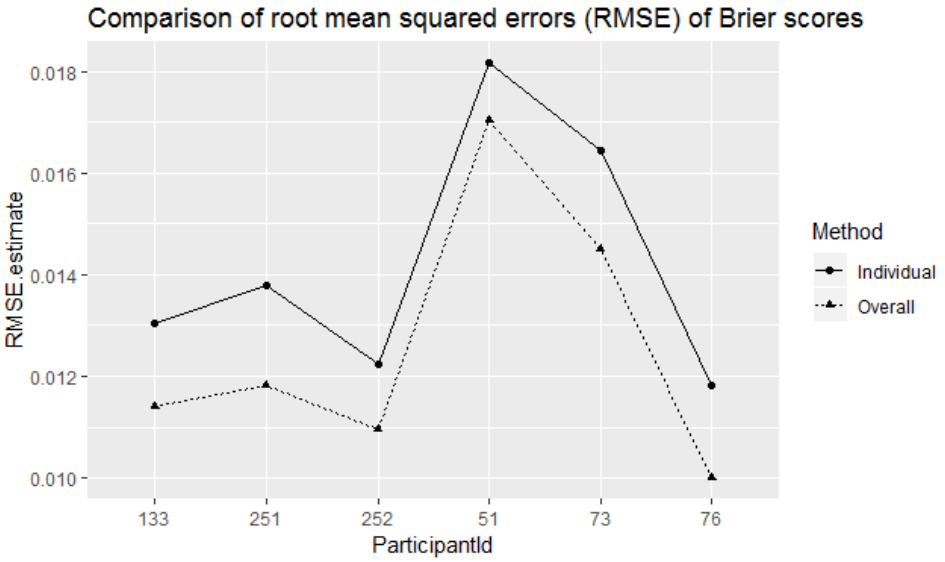


FIGURE D.1: Comparison of root mean squared errors of computed participants' Brier scores with overall and individual 10% missing values introduced completely at random

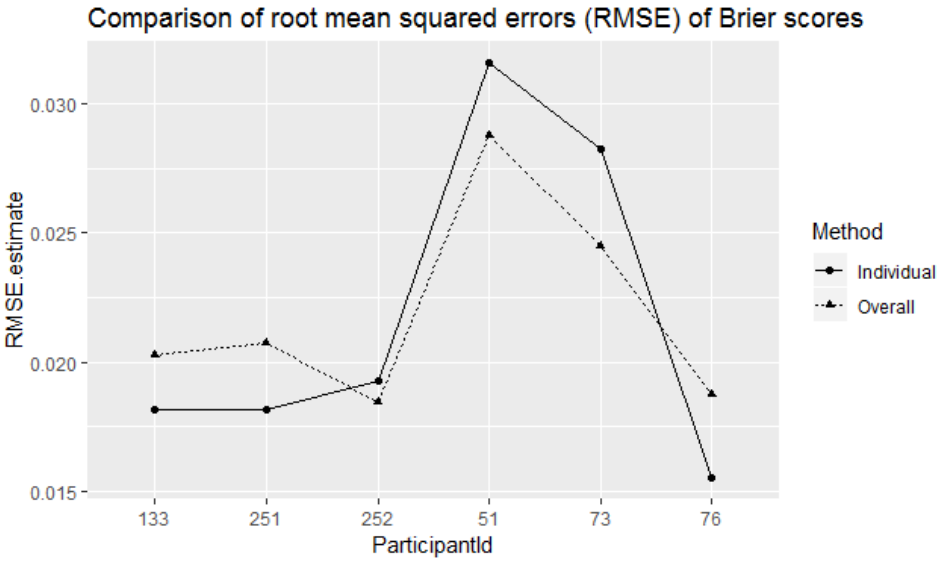


FIGURE D.2: Comparison of root mean squared errors of computed participants' Brier scores with overall and individual 25% missing values introduced completely at random

E. Confidence intervals for mean errors (individual missing values completely at random)

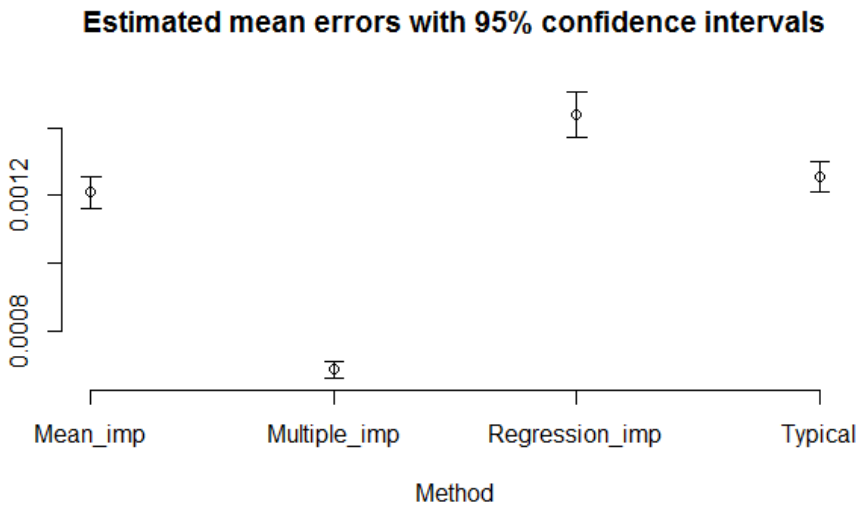


FIGURE E.1: Estimated mean errors with 95% confidence intervals for computing Brier scores with 10% individual missing values introduced completely at random

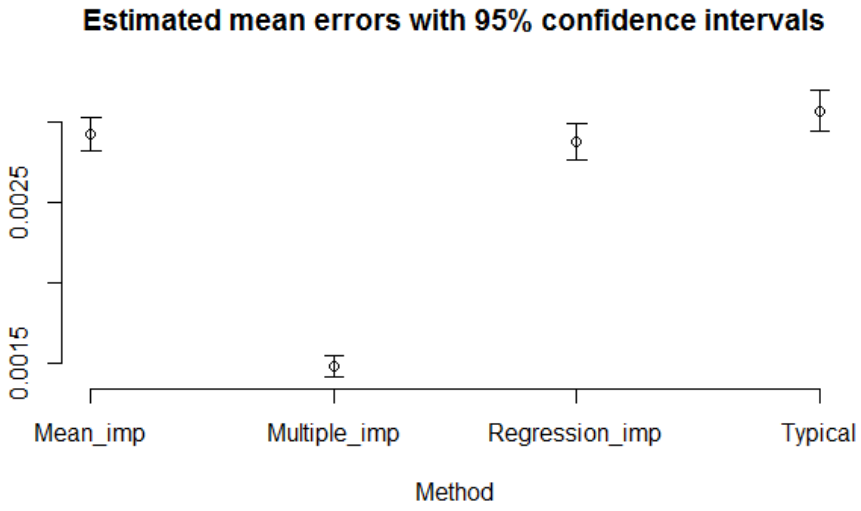


FIGURE E.2: Estimated mean errors with 95% confidence intervals for computing Brier scores with 25% individual missing values introduced completely at random

F. Multiple imputed Brier scores and standard error estimates (missing completely at random)

Table F.1: Brier scores and standard error estimates with 10 percent overall missing values

Participant Id	Original Brier score	Multiple imputed Brier score	Standard error estimate
51	0.14732	0.15185	0.03862
73	0.17228	0.17414	0.03859
76	0.16387	0.16559	0.03856
133	0.14272	0.14795	0.03867
251	0.14703	0.15112	0.03861
252	0.16968	0.17141	0.03862

Table F.2: Brier scores and standard error estimates with 25 percent overall missing values

Participant Id	Original Brier score	Multiple imputed Brier score	Standard error estimate
51	0.14732	0.16190	0.04206
73	0.17228	0.17781	0.04187
76	0.16387	0.17017	0.04180
133	0.14272	0.15638	0.04194
251	0.14703	0.15891	0.04195
252	0.16968	0.17518	0.04184

TABLE F.3: Brier scores and standard error estimates with 10 percent individual missing values

Participant Id	Original Brier score	Multiple imputed Brier score	Standard error estimate
51	0.14732	0.15399	0.03911
73	0.17228	0.17429	0.03903
76	0.16387	0.16615	0.03901
133	0.14272	0.14825	0.03905
251	0.14703	0.15210	0.03903
252	0.16968	0.17126	0.03902

TABLE F.4: Brier scores and standard error estimates with 25 percent individual missing values

Participant Id	Original Brier score	Multiple imputed Brier score	Standard error estimate
51	0.14732	0.15737	0.04194
73	0.17228	0.17575	0.04177
76	0.16387	0.17104	0.04148
133	0.14272	0.15530	0.04172
251	0.14703	0.15757	0.04174
252	0.16968	0.17076	0.04151

G. Confidence intervals for mean errors (individual missing values not at random)

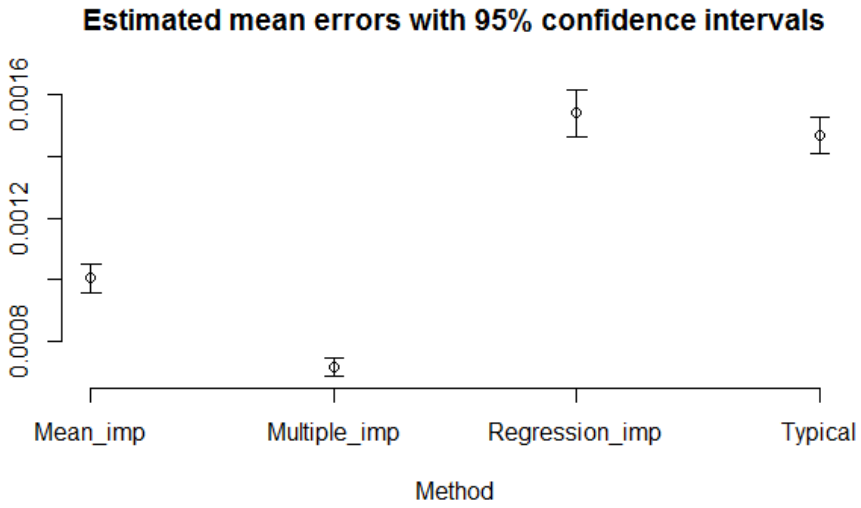


FIGURE G.1: Estimated mean errors with 95% confidence intervals for computing Brier scores with 10% individual missing values introduced not at random

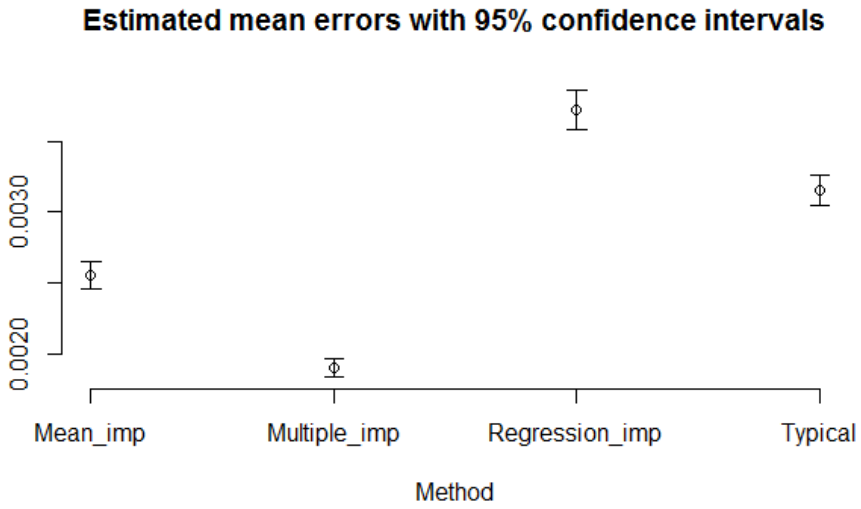


FIGURE G.2: Estimated mean errors with 95% confidence intervals for computing Brier scores with 25% individual missing values introduced not at random

H. Multiple imputed Brier scores and standard error estimates (missing not at random)

TABLE H.1: Brier scores and standard error estimates with 10 percent overall missing values

Participant Id	Original Brier score	Multiple imputed Brier score	Standard error estimate
51	0.14732	0.15069	0.03836
73	0.17228	0.16957	0.03844
76	0.16387	0.16349	0.03891
133	0.14272	0.14702	0.03898
251	0.14703	0.14789	0.03791
252	0.16968	0.17061	0.03844

TABLE H.2: Brier scores and standard error estimates with 25 percent overall missing values

Participant Id	Original Brier score	Multiple imputed Brier score	Standard error estimate
51	0.14732	0.15557	0.04101
73	0.17228	0.16841	0.04140
76	0.16387	0.16568	0.04217
133	0.14272	0.15151	0.04217
251	0.14703	0.15010	0.04001
252	0.16968	0.17256	0.04127

TABLE H.3: Brier scores and standard error estimates with 10 percent individual missing values

Participant Id	Original Brier score	Multiple imputed Brier score	Standard error estimate
51	0.14732	0.15312	0.03922
73	0.17228	0.17488	0.03951
76	0.16387	0.16363	0.03906
133	0.14272	0.14941	0.04008
251	0.14703	0.15460	0.04015
252	0.16968	0.17252	0.03972

TABLE H.4: Brier scores and standard error estimates with 25 percent individual missing values

Participant Id	Original Brier score	Multiple imputed Brier score	Standard error estimate
51	0.14732	0.16405	0.04308
73	0.17228	0.17890	0.04349
76	0.16387	0.16556	0.04275
133	0.14272	0.15931	0.04438
251	0.14703	0.16455	0.04469
252	0.16968	0.17646	0.04385

I. R program for computing the critical regions of the equivalence test

```
# R program for computing the critical regions of the equivalence
# test of a single binomial proportion (the original program 'bi1st'
# was used in Wellek (2010))
# Significance level of the test
alpha <- .05
# Level of intended coverage probability of credible intervals
conf_level <- assigned_conf_level
# Assigned acceptable margin of deviation around the referenced value
epsilon <- assigned_epsilon
# Number of elicited intervals
n <- num.intervals
P1 <- conf_level - epsilon
P2 <- conf_level + epsilon
P0 <- (P1 + P2) / 2 # middle value of the equivalence range
K <- trunc(n/2)
indiP1K <- 0
indiS6 <- 0
if (2*K >= n || P2 == (1-P1))
 { P1K <- pbinom(K,n,P1) - pbinom(K-1,n,P1)
   if (P1K >= alpha)
    { C1 <- K
      C2 <- K
```

```
      gam1 <- alpha/P1K
      gam2 <- gam1
      POWNONRD <- 0
      POK <- pbinom(K,n,.5) - pbinom(K-1,n,.5)
      POW <- gam1 * POK
      indiP1K <- 1  } }
if (indiP1K != 1)
{
P0 <- (P1+P2) / 2
K1 <- max(trunc(n*P1),1)
K2 <- max(trunc(n*P0)-2,K1-1)
repeat
{
FBINP1C1 <- pbinom(K1-1,n,P1)
alpha1 <- 0
FBINP2C1 <- pbinom(K1-1,n,P2)
alpha2 <- 0
while(max(alpha1,alpha2) <= alpha)
  {  alpha1 <- pbinom(K2,n,P1) - FBINP1C1
     alpha2 <- pbinom(K2,n,P2) - FBINP2C1
     K2 <- K2 + 1  }
K2 <- K2 - 2
if(K2 < K1)
  { INCL <- 1
    INCR <- 1  } else
   { K1 <- K1 + 1
     INCL <- 0
     INCR <- 1 }
repeat
 {
alpha1 <- pbinom(K2,n,P1) - pbinom(K1-1,n,P1)
alpha2 <- pbinom(K2,n,P2) - pbinom(K1-1,n,P2)
```

```
delalph1 <- alpha - alpha1
delalph2 <- alpha - alpha2
b11 <- pbinom(K1-1,n,P1) - pbinom(max(K1-2,0),n,P1) * sign(1+sign(K1-2))
b12 <- pbinom(K2+1,n,P1) - pbinom(K2,n,P1)
b21 <- pbinom(K1-1,n,P2) - pbinom(max(K1-2,0),n,P2) * sign(1+sign(K1-2))
b22 <- pbinom(K2+1,n,P2) - pbinom(K2,n,P2)
gam1 <- (b22*delalph1 - b12*delalph2) / (b11*b22 - b12*b21)
gam2 <- (b11*delalph2 - b21*delalph1) / (b11*b22 - b12*b21)
if ((min(gam1,gam2)<0 || max(gam1,gam2) >= 1) && INCL == 0 && INCR == 1)
   { K1 <- K1 - 1
     K2 <- K2 - 1
     INCL <- 1
     INCR <- 0  } else
if ((min(gam1,gam2)<0 || max(gam1,gam2) >= 1) && INCL == 1 && INCR == 0)
   { K2 <- K2 +1
     INCL <- 1
     INCR <- 1  } else
if ((min(gam1,gam2)<0 || max(gam1,gam2) >= 1) && INCL == 1 && INCR == 1)
   { K1 <- K1 +1
     break    } else
   { indiS6 <- 1
     break    }   }
if (indiS6 == 1)
    break          }
C1 <- K1 - 1 # lower bound of the rejection region
C2 <- K2 + 1 # upper bound of the rejection region
}
```

J. Experts' calibration on eliciting 80% credible intervals

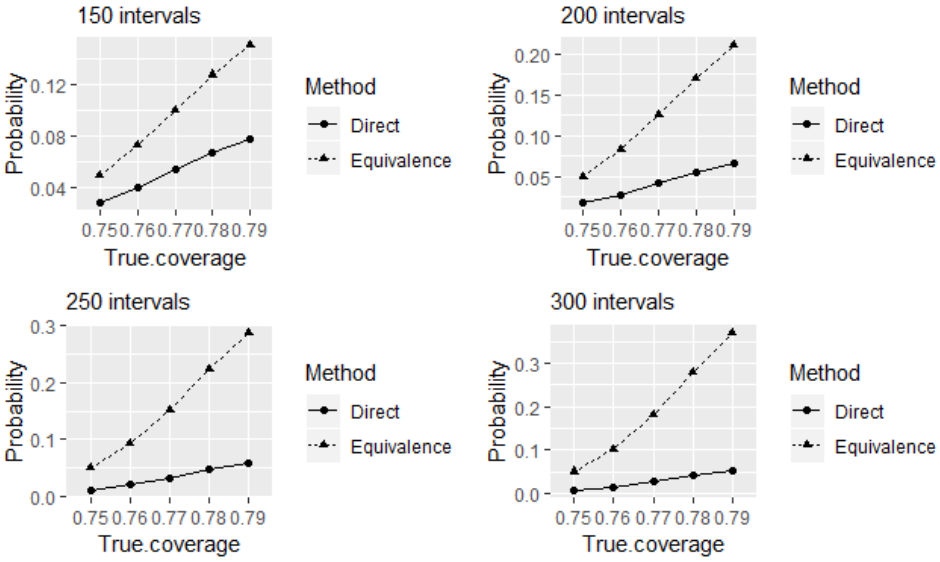


FIGURE J.1: The probabilities of the direct and equivalence tests to identify the experts as 80% well-calibrated when true levels of coverage of elicited intervals are less than 80%

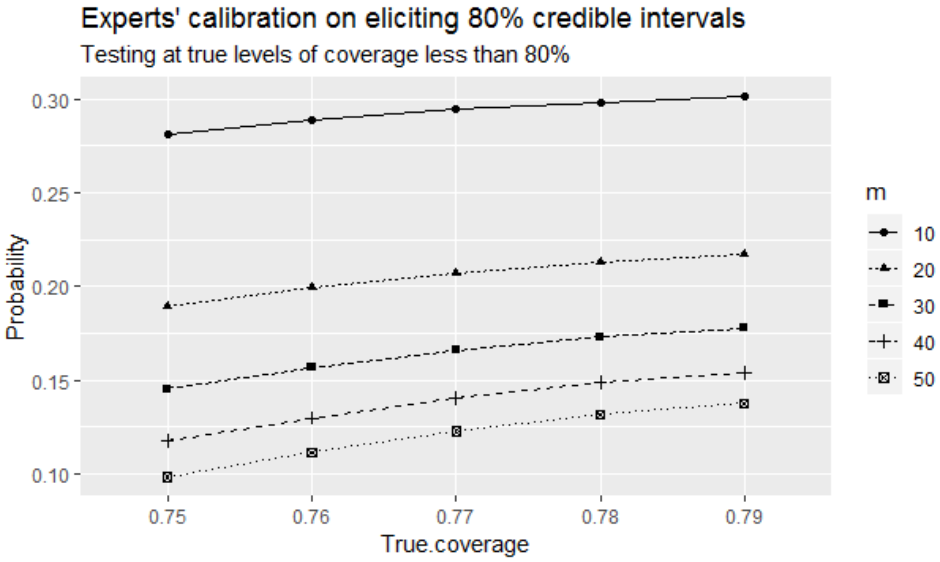


FIGURE J.2: The probabilities of the direct test to identify the experts as 80% well-calibrated when true levels of coverage of elicited intervals are less than 80% for small number of elicited intervals

K. Derived Cooke's weights from training data

Output 1: Derived Cooke's weights without Decision Maker (DM) optimisation
from PBINTDOS data set
Case name : PBINTDOS_training 7/10/2019 CLASS version W4.0

Results of scoring experts
Bayesian Updates: no Weights: global DM Optimisation: no
Significance Level: 0 Calibration Power: 1

Nr.	Id	Calibr.	Mean relat total	Mean relat realizatii	Numb real	UnNormaliz weight	Normaliz.w without DM	Normaliz.w with DM
1	exp 2 +	0.4735	0.9791	0.9791	10	0.4636	0.7523	0.5542
2	exp 3 +	0.02367	1.07	1.07	10	0.02532	0.04109	0.03026
3	exp 4 +	0.06085	1.315	1.315	10	0.08	0.1298	0.09562
4	exp 5 +	1.543E-7	2.363	2.363	10	3.646E-7	5.916E-7	4.358E-7
5	exp 6 +	0.06085	0.7775	0.7775	10	0.04731	0.07676	0.05654
6	DMaker 1	0.4735	0.4654	0.4654	10	0.2204		0.2634

(c) 1999 TU Delft

Output 2: Derived Cooke's weights without Decision Maker (DM) optimisation
from RETURNafter data set
Case name : RETURNafter_training 7/10/2019 CLASS version W4.0

Results of scoring experts
Bayesian Updates: no Weights: global DM Optimisation: no
Significance Level: 0 Calibration Power: 1

Nr.	Id	Calibr.	Mean relat total	Mean relat realizatii	Numb real	UnNormaliz weight	Normaliz.w without DM	Normaliz.w with DM
1	Portfol1	0.3135	0.866	0.866	10	0.2715	0.2866	0.2636

```
     2|Portfol2|    0.4735|    0.9158|    0.9158|  10|    0.4337|    0.4578|     0.421
     3|Portfol3|    0.3946|    0.1751|    0.1751|  10|   0.06909|   0.07294|   0.06707
     4|Riskan1 | 0.0007994|     1.443|     1.443|  10|  0.001153|  0.001218|   0.00112
     5|Riskan2 |    0.3005|    0.5716|    0.5716|  10|    0.1718|    0.1814|    0.1668
     6|DMaker 1|    0.2894|    0.2866|    0.2866|  10|   0.08295|          |   0.08052
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
                                                          (c) 1999 TU Delft
```

L. Sample weights of PBINTDOS data

```
##    Expert_id Sample_number Expert_weight
## 1          2             1     2.058e-01
## 2          3             1     7.223e-01
## 3          4             1     5.152e-02
## 4          5             1     1.120e-04
## 5          6             1     2.030e-02
## 6          2             2     6.960e-05
## 7          3             2     9.993e-01
## 8          4             2     6.340e-04
## 9          5             2     1.680e-07
## 10         6             2     1.100e-05
## 11         2             3     5.121e-01
## 12         3             3     4.165e-01
## 13         4             3     7.750e-04
## 14         5             3     3.650e-03
## 15         6             3     6.693e-02
## 16         2             4     7.358e-01
## 17         3             4     1.384e-02
## 18         4             4     1.895e-02
## 19         5             4     3.600e-04
## 20         6             4     2.311e-01
## 21         2             5     1.796e-01
## 22         3             5     8.086e-01
## 23         4             5     4.490e-03
```

```
## 24          5           5     2.230e-04
## 25          6           5     7.110e-03
## 26          2           6     9.150e-02
## 27          3           6     8.717e-01
## 28          4           6     1.760e-03
## 29          5           6     2.175e-02
## 30          6           6     1.332e-02
## 31          2           7     1.811e-01
## 32          3           7     5.414e-01
## 33          4           7     8.172e-02
## 34          5           7     5.360e-03
## 35          6           7     1.904e-01
## 36          2           8     1.134e-01
## 37          3           8     8.579e-01
## 38          4           8     4.800e-06
## 39          5           8     1.760e-06
## 40          6           8     2.869e-02
## 41          2           9     1.353e-01
## 42          3           9     1.404e-01
## 43          4           9     4.480e-05
## 44          5           9     1.410e-04
## 45          6           9     7.242e-01
## 46          2          10     1.090e-03
## 47          3          10     9.986e-01
## 48          4          10     2.150e-04
## 49          5          10     5.930e-06
## 50          6          10     3.680e-05
```

M. Results from PBINTDOS testing data

Output 1: Results from PBINTDOS testing data using the user define normalized Cooke's weights obtained from training data

Case name : PBINTDOS_testing 7/12/2019 CLASS version W4.0

Results of scoring experts

Bayesian Updates: no Weights: user DM Optimisation: no

Significance Level: 0 Calibration Power: 1

Nr.	Id	Calibr.	Mean relat total	Mean relat realizatii	Numb real	UnNormaliz weight	Normaliz.w without DM	Normaliz.w with DM
1	exp 2 +	2.755E-5	2.022	2.022	25	5.57E-5	0.7523	5.017E-5
2	exp 3 +	0.3749	0.8748	0.8748	25	0.3279	0.04109	0.2953
3	exp 4 +	1.487E-9	2.768	2.768	25	4.116E-9	0.1298	3.707E-9
4	exp 5 +	4.069E-13	2.475	2.475	25	1.007E-12	5.916E-7	9.072E-13
5	exp 6 +	3.433E-8	1.429	1.429	25	4.905E-8	0.07681	4.418E-8
6	DMaker 1	0.7496	1.044	1.044	25	0.7823		0.7046

Output 2: Results from PBINTDOS testing data using the user define normalized shrinkage Cooke's weights obtained from training data

Case name : PBINTDOS_testing 7/12/2019 CLASS version W4.0

Results of scoring experts

Bayesian Updates: no Weights: user DM Optimisation: no

Significance Level: 0 Calibration Power: 1

Nr.	Id	Calibr.	Mean relat total	Mean relat realizatii	Numb real	UnNormaliz weight	Normaliz.w without DM	Normaliz.w with DM
1	exp 2 +	2.755E-5	2.022	2.022	25	5.57E-5	0.681	4.865E-5

```
     2|exp 3 + |    0.3749|    0.8748|    0.8748|  25|    0.3279|   0.03276|    0.2864
     3|exp 4 + |  1.487E-9|     2.768|     2.768|  25|  4.116E-9|    0.2094|  3.595E-9
     4|exp 5 + | 4.069E-13|     2.475|     2.475|  25| 1.007E-12|  1.263E-5| 8.797E-13
     5|exp 6 + |  3.433E-8|     1.429|     1.429|  25|  4.905E-8|    0.0768|  4.284E-8
     6|DMaker 1|    0.7587|     1.077|     1.077|  25|    0.8169|          |    0.7135
-------------------------------------------------------------------------------
-------------------------------------------------------------------------------
                                                          (c) 1999 TU Delft
```

N. Results from RETURNafter testing data

Output 1: Results from RETURNafter testing data using the user define normalized Cooke's weights obtained from training data

Case name : RETURNafter_testing 7/12/2019 CLASS version W4.0

Results of scoring experts

Bayesian Updates: no Weights: user DM Optimisation: no

Significance Level: 0 Calibration Power: 1

Nr.	Id	Calibr.	Mean relat total	Mean relat realizatii	Numb real	UnNormaliz weight	Normaliz.w without DM	Normaliz.w with DM
1	Portfol1	0.1312	0.721	0.721	21	0.09462	0.2866	0.8914
2	Portfol2	6.019E-9	0.8251	0.8251	21	4.966E-9	0.4578	4.678E-8
3	Portfol3	4.786E-6	0.1524	0.1524	21	7.295E-7	0.07294	6.873E-6
4	Riskan1	0.004916	1.471	1.471	21	0.007233	0.001218	0.06814
5	Riskan2	0.0007734	0.8819	0.8819	21	0.000682	0.1814	0.006425
6	DMaker 1	0.01487	0.2433	0.2433	21	0.003616		0.03407

Output 2: Results from RETURNafter testing data using the user define normalized shrinkage Cooke's weights obtained from training data

Case name : RETURNafter_testing 7/12/2019 CLASS version W4.0

Results of scoring experts

Bayesian Updates: no Weights: user DM Optimisation: no

Significance Level: 0 Calibration Power: 1

Nr.	Id	Calibr.	Mean relat total	Mean relat realizatii	Numb real	UnNormaliz weight	Normaliz.w without DM	Normaliz.w with DM
1	Portfol1	0.1312	0.721	0.721	21	0.09462	0.2466	0.9116

```
     2|Portfol2|  6.019E-9|    0.8251|    0.8251|  21|  4.966E-9|    0.3219|  4.785E-8
     3|Portfol3|  4.786E-6|    0.1524|    0.1524|  21|  7.295E-7|   0.05179|  7.028E-6
     4|Riskan1 |  0.004916|     1.471|     1.471|  21|  0.007233|    0.1902|   0.06969
     5|Riskan2 | 0.0007734|    0.8819|    0.8819|  21|  0.000682|    0.1895|  0.006571
     6|DMaker 1|  0.004452|    0.2837|    0.2837|  21|  0.001263|          |   0.01217
-------------------------------------------------------------------------------------
-------------------------------------------------------------------------------
                                                          (c) 1999 TU Delft
```

Bibliography

Ambler, Gareth, Rumana Z Omar, and Patrick Royston (2007). "A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome". In: *Statistical methods in medical research* 16.3, pp. 277–298.

Andersen, Scott W and Brian A Millen (2013). "On the practical application of mixed effects models for repeated measures to clinical trial data". In: *Pharmaceutical statistics* 12.1, pp. 7–16.

Baranchik, Alvin J (1964). *Multiple regression and estimation of the mean of a multivariate normal distribution.* Tech. rep. STANFORD UNIV CALIF.

Bates, Douglas et al. (2014). *lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-7).*

Bates, Douglas et al. (2015). "Package 'lme4'". In: *Convergence* 12.1, p. 2.

Bennett, Derrick A (2001). "How can I deal with missing data in my study?" In: *Australian and New Zealand journal of public health* 25.5, pp. 464–469.

Bolker, Benjamin M et al. (2009). "Generalized linear mixed models: a practical guide for ecology and evolution". In: *Trends in ecology & evolution* 24.3, pp. 127–135.

Brewer, Mark J, Adam Butler, and Susan L Cooksley (2016). "The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity". In: *Methods in Ecology and Evolution* 7.6, pp. 679–692.

Brier, Glenn W (1950). "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1, pp. 1–3.

Budescu, David V and Timothy R Johnson (2011). "A model-based approach for the analysis of the calibration of probability judgments". In: *Judgment and Decision Making* 6.8, pp. 857–869.

Candille, G and O Talagrand (2005). "Evaluation of probabilistic prediction systems for a scalar variable". In: *Quarterly Journal of the Royal Meteorological Society* 131.609, pp. 2131–2150.

Colson, Abigail R and Roger M Cooke (2018). "Expert elicitation: using the classical model to validate experts' judgments". In: *Review of Environmental Economics and Policy* 12.1, pp. 113–132.

Cooke, Roger (1991). *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press on Demand.

Cooke, Roger M and Louis LHJ Goossens (2008). "TU Delft expert judgment data base". In: *Reliability Engineering & System Safety* 93.5, pp. 657–674.

Cooke, Roger M and D Solomatine (1992). "EXCALIBR Integrated System for Processing Expert Judgements version 3.0". In: *Delft University of Technology and SoLogic Delft, Delft*.

Cooke, Roger M et al. (2014). "Out-of-sample validation for structured expert judgment of Asian carp establishment in Lake Erie". In: *Integrated Environmental Assessment and Management* 10.4, pp. 522–528.

Delattre, Maud, Marc Lavielle, and Marie-Anne Poursat (2014). "A note on BIC in mixed-effects models". In: *Electronic journal of statistics* 8.1, pp. 456–475.

Demidenko, Eugene (2004). *Mixed models : Theory and Applications*. Hoboken, N.J. : Wiley.

— (2013). *Mixed models: theory and applications with R*. John Wiley & Sons.

Demirtas, Hakan (2004). "Simulation driven inferences for multiply imputed longitudinal datasets". In: *Statistica Neerlandica* 58.4, pp. 466–482.

Dong, Yiran and Chao-Ying Joanne Peng (2013). "Principled missing data methods for researchers". In: *SpringerPlus* 2.1, p. 222.

Efron, Bradley and Carl Morris (1973). "Stein's estimation rule and its competitors—an empirical Bayes approach". In: *Journal of the American Statistical Association* 68.341, pp. 117–130.

— (1975). "Data analysis using Stein's estimator and its generalizations". In: *Journal of the American Statistical Association* 70.350, pp. 311–319.

— (1977). "Stein's paradox in statistics". In: *Scientific American* 236.5, pp. 119–127.

Efron, Bradley and Charles Stein (1981). "The jackknife estimate of variance". In: *The Annals of Statistics* 9.3, pp. 586–596.

Finch, W Holmes, Jocelyn E Bolin, and Ken Kelley (2016). *Multilevel modeling using R*. Crc Press.

Frey, J. (2009). "An exact multinomial test for equivalence". In: *Canadian Journal of Statistics* 37.1, pp. 47–59.

Gelman, Andrew and Jennifer Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Graham, John W (2009). "Missing data analysis: Making it work in the real world". In: *Annual review of psychology* 60, pp. 549–576.

Hanea, AM et al. (2017). "I nvestigate D iscuss E stimate A ggregate for structured expert judgement". In: *International journal of forecasting* 33.1, pp. 267–279.

Hanea, AM et al. (2018). "Classical meets modern in the IDEA protocol for structured expert judgement". In: *Journal of Risk Research* 21.4, pp. 417–433.

Harel, Ofer and Xiao-Hua Zhou (2007). "Multiple imputation: review of theory, implementation and software". In: *Statistics in medicine* 26.16, pp. 3057–3077.

Harrison, Xavier A et al. (2018). "A brief introduction to mixed effects modelling and multi-model inference in ecology". In: *PeerJ* 6, e4794.

Hemming, Victoria et al. (2018). "A practical guide to structured expert elicitation using the IDEA protocol". In: *Methods in Ecology and Evolution* 9.1, pp. 169–180.

Herzog, Thomas N and Donald B Rubin (1983). "Using multiple imputations to handle nonresponse in sample surveys". In: *Incomplete data in sample surveys* 2, pp. 209–245.

Horton, Nicholas J and Ken P Kleinman (2007). "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models". In: *The American Statistician* 61.1, pp. 79–90.

Hox, J. J. (2002). *Multilevel analysis : techniques and applications*. Mahwah, N.J. : Lawrence Erlbaum Associates, 2002.

Hox, Joop J and Timo M Bechger (1998). "An introduction to structural equation modeling". In: *Family Science Review* 11, pp. 354–373.

James, W and Charles Stein (1961). "Estimation with quadratic loss." In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 361–379.

Janssen, Kristel JM et al. (2010). "Missing covariate data in medical research: to impute is better than to ignore". In: *Journal of clinical epidemiology* 63.7, pp. 721–727.

Jiang, Jiming (2007). *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media.

Kaiser, Jiri (2014). "Dealing with Missing Values in Data". In: *Journal of Systems Integration* 5.1, pp. 42–51.

Kang, Hyun (2013). "The prevention and handling of the missing data". In: *Korean journal of anesthesiology* 64.5, pp. 402–406.

Klayman, Joshua, Peter Juslin, and Anders Winman (2006). "Subjective confidence and the sampling of knowledge". In: *Information sampling and adaptive cognition*, pp. 153–82.

Laird, Nan M and James H Ware (1982). "Random-effects models for longitudinal data". In: *Biometrics* 38.4, pp. 963–974.

Lakshminarayan, Kamakshi, Steven A Harp, and Tariq Samad (1999). "Imputation of missing data in industrial databases". In: *Applied intelligence* 11.3, pp. 259–275.

Lichtenstein, Sarah and Baruch Fischhoff (1980). "Training for calibration". In: *Organizational Behavior and Human Performance* 26.2, pp. 149–171.

Littell, Ramon C et al. (1996). *SAS system for mixed models*. Cary, N.C. : SAS Institute, Inc.

Little, Roderick J. A. and Donald B. Rubin (2002). *Statistical analysis with missing data*. Hoboken, N.J. : Wiley, c2002.

McBride, Marissa F (2013). "Expert knowledge for conservation: tools for enhancing the quality of expert judgment". PhD thesis. School of Botany, University of Melbourne.

McBride, Marissa F, Fiona Fidler, and Mark A Burgman (2012). "Evaluating the accuracy and calibration of expert predictions under uncertainty: predicting the outcomes of ecological research". In: *Diversity and Distributions* 18.8, pp. 782–794.

Merkle, Edgar C. et al. (2016). "Item response models of probability judgments: Application to a geopolitical forecasting tournament". In: *Decision* 3.1, pp. 1–19.

Meteyard, Lotte and Robert AI Davies (2020). "Best practice guidance for linear mixed-effects models in psychological science". In: *Journal of Memory and Language* 112, p. 104092.

Mulford, Matthew et al. (1998). "Physical attractiveness, opportunity, and success in everyday exchange". In: *American journal of sociology* 103.6, pp. 1565–1592.

Noortgate, Wim Van den, Marie-Christine Opdenakker, and Patrick Onghena (2005). "The effects of ignoring a level in multilevel analysis". In: *School Effectiveness and School Improvement* 16.3, pp. 281–303.

O'Hagan, Anthony et al. (2006). *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons.

O'Hagan, Anthony (2019). "Expert Knowledge Elicitation: Subjective but Scientific". In: *The American Statistician* 73.sup1, pp. 69–81.

Peng, Chao-Ying Joanne et al. (2006). "Advances in missing data methods and implications for educational research". In: *Real data analysis* 3178.

Peters, Ellen and Paul Slovic (2000). "The springs of action: Affective and analytical information processing in choice". In: *Personality and Social Psychology Bulletin* 26.12, pp. 1465–1475.

Pinheiro, Jose et al. (2017). "R Core Team (2017) nlme: linear and nonlinear mixed effects models. R package version 3.1-131". In: *Computer software Retrieved from https://CRAN. R-project. org/package= nlme*.

Pinheiro, José C and Douglas M Bates (2000). *Mixed effects models in S and S-PLUS*. New York : Springer, c2000.

Plous, Scott (1993). *The psychology of judgment and decision making*. Mcgraw-Hill Book Company.

Predd, Joel B et al. (2008). "Aggregating probabilistic forecasts from incoherent and abstaining experts". In: *Decision Analysis* 5.4, pp. 177–189.

Robinson, Andrew P and Jeff D Hamann (2010). *Forest analytics with R: an introduction*. Springer Science & Business Media.

Robinson, GK (1991). "That BLUP Is a Good Thing: The Estimation of Random Effects". In: *Statistical Science* 6.1, pp. 15–32.

Rodrıguez, Germán and Irma Elo (2003). "Intra-class correlation in random-effects models for binary data". In: *Stata J* 3.1, pp. 32–46.

Rubin, Donald B (1978). "Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse". In: *Proceedings of the survey research methods section of the American Statistical Association*. Vol. 1. American Statistical Association, pp. 20–34.

Rubin, Donald B (1986). "Statistical matching using file concatenation with adjusted weights and multiple imputations". In: *Journal of Business & Economic Statistics* 4.1, pp. 87–94.

Rubin, Donald B. (1988). "An Overview of Multiple Imputation". In: *In Proceedings of the Survey Research Section, American Statistical Association*, pp. 79–84.

Rubin, Donald B (2004a). *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons.

Rubin, Donald B (2004b). "The design of a general and flexible system for handling nonresponse in sample surveys". In: *The American Statistician* 58.4, pp. 298–302.

Rubin, Donald B and Nathaniel Schenker (1987). "Interval estimation from multiply-imputed data: a case study using census agriculture industry codes". In: *Journal of Official Statistics* 3.4, p. 375.

Scheffer, Judi (2002). "Dealing with Missing Data". In: *Research Letters in the Information and Mathematical Sciences*. Vol. 3. Citeseer, pp. 153–160.

Soll, Jack B and Joshua Klayman (2004). "Overconfidence in interval estimates." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30.2, 299–314.

Speelman, Dirk, Kris Heylen, and Dirk Geeraerts (2018). *Mixed-effects regression models in linguistics*. Springer.

Speirs-Bridge, Andrew et al. (2010). "Reducing overconfidence in the interval judgments of experts". In: *Risk Analysis* 30.3, pp. 512–523.

Steenbergen, Marco R and Bradford S Jones (2002). "Modeling multilevel data structures". In: *american Journal of political Science*, pp. 218–237.

Stein, Charles (1956). "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution." In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 197–206.

Stockard, Jean, Ellen Peters, et al. (2007). "The use of mixed models in a modified Iowa Gambling Task and a prisoner's dilemma game". In: *Judgment and Decision Making* 2.1, pp. 9–22.

Sullivan, Lisa, Kimberly A. Dukes, and Elena Losina (1999). "Tutorial in biostatistics. An introduction to hierarchical linear modelling." In: *Statistics in medicine* 18 7, pp. 855–88.

Tabachnick, Barbara G and Linda S Fidell (2013). "Using multivariate statistics, 6th edn Boston". In: *Ma: Pearson*.

Teigen, Karl Halvor and Magne Jørgensen (2005). "When 90% confidence intervals are 50% certain: On the credibility of credible intervals". In: *Applied Cognitive Psychology* 19.4, pp. 455–475.

Treiman, Donald J, William T Bielby, and Man-Tsun Cheng (1988). "Evaluating a multiple-imputation method for recalibrating 1970 US census detailed industry codes to the 1980 standard". In: *Sociological Methodology*, pp. 309–345.

Venables, W. N. and Brian D Ripley (2002). *Modern applied statistics with S*. New York : Springer.

Voinov, Vassiliy G and Mikhail S Nikulin (1995). "A review of the results on the Stein approach for estimators improvement". In: *Qüestiió: quaderns d'estadística i investigació operativa* 19.1.

Wellek, Stefan (2010). *Testing statistical hypotheses of equivalence and noninferiority*. Chapman and Hall/CRC.

Wintle, B et al. (2012). "The Intelligence Game: Assessing Delphi Groups and Structured Question Formats". In: *2012 SECAU Security and Intelligence Congress*. SRI Security Research Institute, Edith Cowan University, pp. 1–26.

Yarkiner, Zalihe et al. (2013). "Applications of mixed models for investigating progression of chronic disease in a longitudinal dataset of patient records from general practice". In: *Journal of Biometrics and Biostatistics* S9, p. 001.

Zhao, Zhigen (2010). "Double shrinkage empirical Bayesian estimation for unknown and unequal variances". In: *Statistics and Its Interface* 3.4, pp. 533–541.