Judgment and Decision Making, Vol. xx, No. x, Month 20xx, pp. XX–XX

Using Cross-Domain Expertise to Aggregate Forecasts When Within-Domain Expertise is Unknown

Marcellin Martinie* Tom Wilkening† Piers Howe‡

Abstract

In recent years, a number of crowd aggregation approaches have been proposed to combine the predictions of different forecasters in problems where decision makers do not have records of forecasters' past performance in a related domain. In these types of problems, decision makers can often obtain a measure of forecasters' past performance in an unrelated domain where the outcomes to questions are known and forecasters' performance can be easily assessed. The current paper explores the extent to which forecasters' relative expertise in one domain can be used to leverage their forecasts in an unrelated domain. Over three experiments comprising a range of decision problems from art, science, sport, and a test of emotional intelligence, we compare the performance of aggregation approaches that do not use forecasters' past performance to those that weight by forecasters' past performance on questions from the same domain (within-domain weighting) or from a different domain (cross-domain weighting). Our results show that although within-domain weighting generally outperforms all other aggregation approaches, cross-domain weighting can be as effective as within-domain weighting when aggregation weights are estimated from forecasters' performance on multiple unrelated domains and there is clear variability in forecasters' expertise in each domain. Our results demonstrate the potential of cross-domain weighting in problems where records of forecasters' past performance in a domain of interest are unavailable.

Keywords: forecast aggregation, wisdom of crowds, expertise, forecasting, decision-making

1 Introduction

State-of-the-art forecasting algorithms harness the Wisdom of Crowds by aggregating predictions from multiple forecasters to generate more accurate predictions than can be obtained from an individual forecaster (Surowiecki, 2005). When aggregating predictions, decision-makers typically weight forecasters' predictions according to forecasters' past performance in a domain related to the one for which predictions are required, so as to obtain more accurate predictions than can be obtained by simply averaging the predictions of all individual forecasters (Armstrong, 2001; Budescu & Chen, 2015; Clemen, 1989; Cooke, 1991; Winkler, 1989). But what should one do if one does not have a record of the forecasters' past performance in a related domain?

Here we propose a simple yet surprisingly effective solution: use each forecaster's past performance in domains unrelated to the one for which forecasts are required. Surprisingly, we found that this technique often produced forecasts as accurate as those based on records of past performance within the domain of interest and, even when it did not, always produced forecasts at least as good as and usually better than the unweighted aggregation of the individual forecasts. Additionally, we found that this technique produced at least as good, and sometimes better, forecasts than the existing, best performing algorithms that are designed to optimally combine the predictions of multiple forecasters when their past performance is unknown.

The only two caveats to our proposed solution is that one needs a record of the forecasters' past performance in more than one unrelated domain and there needs to be clear evidence that in each domain some forecasters have greater expertise than others. Cross-domain estimates of expertise can be unreliable if they are based only on a single domain or use domains where all forecasters have similar expertise. However, if an individual performs well in multiple unrelated domains relative to their peers, then they can be expected to perform well relative to their peers in the domain of interest.

In the rest of this introduction, we will first summarise the literature on traditional forecast aggregation algorithms and explain why we use the CWM algorithm as the basis for our cross-domain weighting studies (Budescu & Chen, 2015). We will then summarise the literature on algorithms that are designed to aggregate forecasts when the forecasters' past performance is unknown. Such algorithms are sometimes referred to as single-question algorithms because

*Corresponding author, Email: marcellin.martinie@gmail.com Melbourne School of Psychological Sciences, The University of Melbourne. ORCID: 0000-0002-1289-1467

†Department of Economics, The University of Melbourne. ORCID: 0000-0001-8037-9951

‡Melbourne School of Psychological Sciences, The University of Melbourne. ORCID: 0000-0001-6171-1381

performance on previous questions of the same type is unavailable, so the task is to make a prediction based solely on responses to a single question (Prelec et al., 2017). We will explain which of these single-question algorithms typically perform the best and use these as points of comparison for our cross-domain weighting algorithm. We will end by summarising the structure of the rest of the article.

1.1 Traditional Forecast Aggregation Algorithms

Because the errors of different individuals are usually not perfectly correlated, averaging the forecasts of a group of individuals will tend to cancel out their individual errors and consequently will typically produce a more accurate forecast than the forecast of the average individual (Clemen, 1989; Davis-Stober et al., 2014; Soll & Larrick, 2009; Surowiecki, 2005). The predictive accuracy of the resultant forecasts can be increased further by weighting individuals according to their expertise. For example, Cooke (1991) developed a classical model drawing from statistical hypothesis testing, where forecasters' weights are derived from their calibration performance on a set of seed questions with known outcomes. Forecasters whose performance on the seed questions are below a theoretical threshold are assigned weights of zero, and their predictions are removed from the crowd. Such an approach is common in the forecast aggregation literature (Armstrong, 2001; Clemen, 1989; Winkler, 1989).

More recently, the contribution weighted model (CWM) proposed by Budescu and Chen (2005) adopted a different approach. The key insight of this model was that combining the forecasts of the highest performing individuals may not always produce the best aggregate forecast. Instead, this algorithm analyses to what extent, in the past, each individual contributed to making the aggregate prediction more accurate. Individuals are then weighted based on their previous contribution. Budescu and Chen were able to show that an aggregate forecast based on these weights outperformed an aggregate forecast where individuals were weighted based on their individual past performance. For this reason, we utilised the CWM algorithm in our study.

1.2 Single-Question Forecasting Algorithms

In recent years, a number of approaches have been developed to aggregate forecasts in domains where forecasters' past performance on related questions is unknown, the so called single-question problem (Kurvers et al., 2019; McCoy & Prelec, 2017; Palley & Soll, 2019; Prelec et al., 2017; Satopää et al., 2016). Some of these algorithms can only make binary (yes/no) forecasts, so cannot be applied to the problem sets that we consider (Kurvers et al., 2019; Prelec et al., 2017; Wilkening et al., 2021). However, more recent single-question algorithms can predict the *probability* that an event will occur (Martinie et al., 2020; Palley & Soll, 2019). For example, Palley and Soll (2019) developed a theoretical framework to address the problem that when simple averaging is used and multiple forecasters base their forecasts on similar information, shared information will become overweighted in forecasters' reports and unique information that is available to only a small subset of forecasters will become under-weighted. The authors argued that the information shared between forecasters can be estimated using the forecasters' meta-predictions about the average forecasts of others. They further proposed an aggregation algorithm, called the minimal pivoting procedure, that uses the difference between the average forecast and the average meta-prediction to adjust the average forecast. Palley and Soll demonstrated theoretically that, under idealised circumstances, this minimal pivoting procedure would completely remove the overweighting of the shared information. In more realistic circumstances, the authors showed that this procedure should reduce, but not completely eliminate, the over-weighting of the shared information, so would still lead to a more accurate forecast than that obtained by simple averaging. Empirically, Palley and Soll found that this minimal pivoting (MP) algorithm did indeed outperform simple averaging on a range of real-world forecasting problems.

More recently, Martinie et al. (2020) developed the Meta-Probability Weighting (MPW) algorithm as an alternative way of using the forecasters' meta-predictions about the average forecasts of others to improve the average forecast. The key insight of this algorithm is that the difference between the forecaster's own prediction and their meta-prediction of the average prediction others can be used to reliably estimate their expertise. The argument, originally proposed by Prelec et al. (2017), is that someone without any knowledge of the problem (i.e., a non-expert) has no reason to suppose that their prediction would be any different from the prediction of others. Thus, the difference between their prediction and their meta-prediction of the average prediction of other people should always be small. Conversely, someone who is knowledgeable about the problem (i.e., an expert) is also likely to be knowledgeable about common misconceptions. Thus, their meta-prediction of the average prediction of others might be very different from their own prediction. Consequently, the larger the difference between a forecaster's prediction and meta-prediction, the more likely they are to be an expert. The MPW algorithm consequently weights forecasters by the difference between their predictions and meta-predictions. Martinie et al. (2020) showed that the MPW algorithm empirically outperformed simple averaging, as well as a number of other single-question aggregation algorithms, across a range of decision problems varying in difficulty.

While these single-question aggregation approaches generally perform well relative to simpler aggregation approaches such as majority voting and simple averaging, their

key advantage is that they can be applied when records of forecasters' past performance are unavailable. To date, few comparisons have been made between single-question approaches and aggregation approaches that use forecasters past performance to identify expertise, and thus, it remains unclear whether these single-question approaches will outperform these other, more-traditional approaches when it is possible to obtain records of forecasters' past performance, albeit in unrelated domains. A central aim of this paper is therefore to examine the efficacy of single-question approaches relative to weighting by forecasters' performance on training questions with known outcomes and to determine to what extent performance varies depending on whether the questions are from the domain of interest as opposed to being drawn from an unrelated domain.

1.3 Summary of our Approach

Over three experiments, we evaluated the efficacy of different forecast-aggregation approaches. Specifically, we compared the efficacy of weighting by forecasters' past performance on questions from the same domain (i.e., within-domain weighting) to weighting forecasters by their performance on questions from other, unrelated domains (i.e., cross-domain weighting) using the CWM algorithm (Budescu & Chen, 2015). We also examined whether cross-domain weighting would provide any benefits over single-question aggregation approaches, specifically the MPW algorithm (Martinie et al., 2020) and the Minimal Pivoting algorithm (Palley & Soll, 2019). Performing this comparison is important because cross-domain weighting requires decision makers to expend resources eliciting additional responses to questions with known outcomes. Decision-makers might only be willing to incur this expense if cross-domain weighting can be shown to outperform single-question algorithms.

The rest of this paper is structured as follows. To motivate our experimental design, we first construct a theoretical model that (1) illustrates how cross-domain forecasting might improve forecasts and (2) shows how answers from multiple cross domains may improve on forecasts that use only a single cross domain. We then conduct three experiments. In Experiment 1, we show that within-domain weighting significantly outperforms cross-domain weighting when a single unrelated domain is used but that cross-domain weighting performs just as well as a simple average prediction across all forecasters. In Experiment 2, we elicit forecasters' responses to questions from a third domain and show that cross-domain weighting calculated using two separate domains is superior to simple averaging but not as effective as within-domain weighting. We attribute the latter finding to the lack of clear evidence for relative expertise in one of the domains (the NFL domain). In Experiment 3, we replace the NFL domain with a domain where forecasters show relative expertise, with some forecasters demonstrating more expertise than others. We find that cross-domain weighting again significantly outperforms simple averaging and now performs equally well as within-domain weighting.

2 Theoretical Model

In this section, we develop a theoretical model that illustrates how a forecaster's performance in a domain of interest, as measured by their Brier score measured on a set of questions in that domain, relates to their average performance across a number of other domains.

Let $d_1, \ldots, d_s$ be a set of s knowledge domains. A forecaster's performance, x_i, in domain d_i is modelled as a combination of (1) their general expertise, g, which we assume is constant across all domains and represents factors such as diligence, engagement, and general intelligence, (2) their specific expertise for that domain, e_i, which represents factors such as domain-specific knowledge, and (3) a noise term, n_i, which models the expected variation in performance on repeated tests in the same domain by the same forecaster. Thus,

$$x_i := g + e_i + n_i. \tag{1}$$

For each domain, we define the zero point as the average performance. We define the forecaster's general expertise, the forecaster's domain-specific expertise, and the noise term as latent variables that are normally distributed around zero with variances σ_g^2, $\sigma_{e_i}^2$, and $\sigma_{n_i}^2$ respectively, (i.e., $g \sim \mathcal{N}(0, \sigma_g^2)$, $e_i \sim \mathcal{N}(0, \sigma_{e_i}^2)$ for all i, and $n_i \sim \mathcal{N}(0, \sigma_{n_i}^2)$ for all i). We assume that at least some variation exists in general expertise and domain-specific expertise (i.e., $\sigma_g^2 > 0$ and $\sigma_{e_i}^2 > 0$ for all i), and that there is some noise (i.e., $\sigma_{n_i}^2 > 0$ for all i). We further assume that a forecaster's domain-specific expertise in one domain is independent of their domain-specific expertise in all other domains, such that $\mathbb{E}(e_i.e_j) = 0$ for all $i \neq j$, and that general expertise is also independent of domain-specific expertise such that $\mathbb{E}(g.e_i) = 0$ for all i. We also assume that the noise is independent of both general expertise and domain-specific expertise, such that $\mathbb{E}(g.n_i) = 0$ for all i and $\mathbb{E}(e_i.n_j) = 0$ for all i and j.

Using this model, we can determine the extent to which the forecaster's expertise in one domain is expected to be linearly correlated with their average expertise across a number of other domains. Let d_θ represent the domain of interest and d_Ω represent all other domains in the set of s domains (i.e., $d_\Omega := \{d_1, \ldots, d_s\} \setminus \{d_\theta\}$). We define x_θ as the forecaster's performance in the domain of interest and x_Ω as the forecaster's performance averaged across all other domains. Thus,

$$x_\Omega := g + \frac{1}{s-1} \sum_{i \neq \theta} (e_i + n_i). \tag{2}$$

The correlation between x_θ and x_Ω is given by

$$\rho = \frac{\mathbb{E}(x_\theta.x_\Omega) - \mathbb{E}(x_\theta)\mathbb{E}(x_\Omega)}{\sqrt{\mathbb{E}(x_\theta^2) - (\mathbb{E}(x_\theta))^2}\sqrt{\mathbb{E}(x_\Omega^2) - (\mathbb{E}(x_\Omega))^2}}. \quad (3)$$

Noting that $\mathbb{E}(x_\theta) = \mathbb{E}(x_\Omega) = 0$, this equation simplifies to

$$\rho = \frac{\mathbb{E}(x_\theta.x_\Omega)}{\sqrt{\mathbb{E}(x_\theta^2)}\sqrt{\mathbb{E}(x_\Omega^2)}}. \quad (4)$$

Recalling that $\mathbb{E}(e_i.e_j) = 0$ for all $i \neq j$, $\mathbb{E}(g.e_i) = 0$ for all i, $\mathbb{E}(g.n_i) = 0$ for all i, and $\mathbb{E}(e_i.n_j) = 0$ for all i and j, this equation simplifies to

$$\rho = \frac{\mathbb{E}(g^2)}{\sqrt{\mathbb{E}(g^2 + e_\theta^2 + n_\theta^2)}\sqrt{\mathbb{E}(g^2 + \frac{1}{(s-1)^2}\sum_{i\neq\theta}(e_i^2 + n_i^2))}}. \quad (5)$$

Because $\sigma_g^2 = \mathbb{E}(g^2) - \mathbb{E}(g)^2$, $\sigma_{e_i}^2 = \mathbb{E}(e_i^2) - \mathbb{E}(e_i)^2$, $\sigma_{n_i}^2 = \mathbb{E}(n_i^2) - \mathbb{E}(n_i)^2$ and $\mathbb{E}(g) = 0$, $\mathbb{E}(e_i) = 0$ and $\mathbb{E}(n_i) = 0$ for all i, this equation can be rewritten as:

$$\rho = \frac{\sigma_g^2}{\sqrt{\sigma_g^2 + \sigma_{e_\theta}^2 + \sigma_{n_\theta}^2}\sqrt{\sigma_g^2 + \frac{1}{(s-1)^2}\sum_{i\neq\theta}(\sigma_{e_i}^2 + \sigma_{n_i}^2)}}. \quad (6)$$

To simplify this equation, we define a variable α_i as:

$$\alpha_i := \frac{\sigma_{e_i}^2 + \sigma_{n_i}^2}{\sigma_g^2}, \quad (7)$$

where $0 < \alpha_i < \infty$. Using equation 7, equation 6 rearranges to:

$$\rho = \frac{1}{\sqrt{1 + \alpha_\theta}\sqrt{1 + \frac{1}{(s-1)^2}\sum_{i\neq\theta}\alpha_i}}. \quad (8)$$

We define $\overline{\alpha_\Omega}$ as the average α_i in set d_Ω:

$$\overline{\alpha_\Omega} := \frac{1}{(s-1)}\sum_{i\neq\theta}\alpha_i. \quad (9)$$

Using equation 9, we can simplify equation 8 as follows:

$$\rho = \frac{1}{\sqrt{1 + \alpha_\theta}\sqrt{1 + \frac{1}{(s-1)}\overline{\alpha_\Omega}}}. \quad (10)$$

From equation 10, we can see that the correlation between a forecaster's performance in one domain and their average performance across all other domains increases as either α_θ or $\overline{\alpha_\Omega}$ decreases. Thus, the correlation between a forecaster's performance in one domain and their average performance across all other domains increases when α_i decreases for any $i \in \{1, \ldots, s\}$. This occurs if (1) σ_g^2 increases or (2) $\sigma_{e_i}^2$ or $\sigma_{n_i}^2$ decreases.

Equation 10 highlights three other important aspects of the relationship between a forecaster's performance in the domain of interest and their performance averaged over other forecasting domains. First, the correlation between a forecaster's performance in a domain of interest and their performance averaged across all other domains will, on average, be positive because $\alpha_i < \infty$ for all i. This means that an aggregated prediction from a group of forecasters, where forecasters are weighted by their performance on other domains, will, on average, outperform assigning equal weights to all forecasters.

Second, because $\alpha_i > 0$ for all i, ρ will always be less than 1. Consequently, performance in the domain of interest will be better predicted by a forecaster's past performance in the domain of interest than by that forecaster's performance on other domains.

Third, if the variance for domain-specific expertise and noise are both equal across all domains, ρ will increase as s increases. This implies that as the number of domains in d_Ω increases, the correlation between the average performance across these other domains and the performance in the domain of interest, ρ, will increase. In other words, the prediction of a forecaster's performance in the domain of interest will improve as the number of other domains in which their performance can be measured increases.

If the variance for domain-specific expertise and noise does differ across domains, then the value of adding another domain depends on (1) the number of domains that currently exist and (2) the level of domain-specific knowledge and noise in the candidate domain compared to the domains already in the cross-domain set. In particular, suppose that domain d_{s+1} becomes available, we define the average α_i in the cross-domain set when this new cross-domain is included as:

$$\overline{\alpha_{\Omega+1}} = \frac{1}{s}\left[\left(\sum_{i\neq\theta}\alpha_i\right) + \alpha_{s+1}\right]. \quad (11)$$

Then, ρ will increase if

$$\overline{\alpha_{\Omega+1}} < \frac{s}{s-1}\overline{\alpha_\Omega}.$$

Thus, in general, the correlation between a forecaster's performance in one domain and their average performance across other domains is predicted to increase with the addition of a new domain to the cross-domain set if the amount of domain-specific knowledge and noise in the new domain is not substantially higher than the average of the existing cross-domain set.

We now summarise a series of experiments designed to explore how a cross-domain weighting algorithm performs in forecasting environments using the three results described above as a guide for our experimental design. As discussed in Section 1.2, for these experiments we utilised the CWM algorithm as this is one of the best performing weighting algorithms in the literature. This algorithm weights forecasters not by their individual past performance but their contribution towards improving the aggregated group forecast. How-

ever, under the assumption that a forecaster's contribution weight as determined by the CWM algorithm is positively correlated with their individual performance, the above three predictions should continue to hold. As we demonstrate below, this proves to be the case, with all three predictions being borne out by the empirical data.

3 Experiment 1

Experiment 1 examined forecasters' performance on two domains: American NFL trivia and general-knowledge Science trivia. From the discussion above, we predicted: (1) that cross-domain weighting would be generally less effective than within-domain weighting and (2) that cross-domain weighting would still outperform an unweighted average of forecasters' responses. Without a formal model, it was hard to predict how cross-domain weighting would compare in terms of performance to the single-question aggregation approaches. In particular, these two approaches draw on different forms of information and it hard to compare the relative predictive power of these different forms of information. Ignoring this issue and basing our prediction solely on the fact that cross-domain weighting has access to more information than the single-question approaches have access to, we also predicted (3) that cross-domain weighting would be generally more effective than the two existing single-question aggregation approaches in the literature – the MPW algorithm (Martinie et al., 2020) and the Minimal Pivoting algorithm (Palley & Soll, 2019).

3.1 Methods

We collected people's responses to 50 questions from the NFL trivia domain and 50 questions from the Science Trivia domain. NFL questions were adapted from trivia questions on the www.funtrivia.com website, and then converted into true or false statements. Science trivia questions were taken from the Grades dataset from Martinie et al. (2020), which comprised moderate difficulty general science questions from Biology, Chemistry, Geography, and Physics.

We recruited 100 respondents from Amazon Mechanical Turk; only respondents inside the US were able to participate in the experiment. Participants were paid a flat fee of $4.00 for completing the survey. The survey was conducted on the Qualtrics platform. Before beginning the experiment, participants were first required to answer three basic logic questions, which we used to identify and exclude any non-human agents from the survey. Participants were then asked to answer each question as honestly as they could and without cheating (e.g., by looking up any of the questions online). Two individuals who reported at the end of the experiment that they cheated on the task were excluded from the analyses; analyses were conducted on the data of the remaining 98 participants.

Participants completed 100 trials each, with each trial comprising one statement which was either true or false. Half the statements in each domain were true, and the other half were false. Participants were asked to provide their predictions about (1) whether the statement was more likely to be true or false, (2) the probability that the statement was "true" (from 0–100), and (3) what they think is the average probability estimated by other people on question (2) (from 0–100). Participants' probability forecasts were restricted to between 0 and 50 if they reported that the statement was more likely to be false and between 50 to 100 if they reported that the statement was more likely to be true. Thus all participants were required to provide binary predictions consistent with their probability forecasts.

Participants were presented each question from the set of 100 questions in a randomised order. The full list of questions used in this experiment and the responses for each participant are available for download in the supplementary materials.

3.2 Analyses

In this section, we provide a formal definition of the contribution metric and the transformed Brier scoring rule used to assess the performance of each algorithm. We also provide a description and justification for each forecasting algorithm.

We use the contribution metric proposed by Budescu & Chen (2015) to provide a measure of forecasters' expertise relative to other forecasters in the crowd, as this contribution metric has been shown to be more effective than forecasters' past performance, such as forecasters' mean Brier scores, at identifying and extracting forecasters' expertise (Budescu & Chen, 2015; Chen et al., 2016). Formally, the contribution of the jth forecaster is equal to the difference between the transformed Brier score of the crowd's average forecast, $\overline{S}$, and the transformed Brier score of the crowd's average forecast without that forecaster, $\overline{S}_{\neg j}$, averaged over all N_j forecasts made by that forecaster in the training set:

$$C_j := \sum_{i=1}^{N_j} \frac{\overline{S} - \overline{S}_{\neg j}}{N_j} \tag{12}$$

The transformed Brier score for the jth forecaster, used to assess the accuracy of their forecast, is given by:

$$S_j := 100 - 100 \sum_{k=1}^{K_j} \frac{(O_k - P_k)^2}{K_j}, \tag{13}$$

where $O_k = 1$ if is the correct outcome is "true" for the kth event and $O_k = 0$ otherwise, K_j is the set of events forecasted by forecaster j (i.e., $k \in \{1, \ldots, K_j\}$), and P_k is the probability assigned to the outcome being "true" by the

jth forecaster. This linear transformation of the Brier score retains the same functional form as the original Brier score proposed by Brier (1950), and is strictly proper (Murphy & Winkler, 1970). Further, it has a straightforward interpretation where scores range from 0 to 100, with 100 being a perfect forecast over all events and 75 being the score that is generated from an uninformed forecasts of $P_k = 0.5$ in all questions. We use the transformed Brier score to assess the performance of each aggregation approach in this paper.

Our main comparisons of interest were between cross-domain weighting and four other aggregation approaches: the simple average, the MPW algorithm, the Minimal Pivoting algorithm, and within-domain weighting. We assessed the mean score of each approach across all 100 questions in the dataset, and on each of the two domains.

For each statistical comparison in this paper, we report a Bayes factor (BF_{10}) calculated using a paired-samples Bayesian t-test in JASP (Wagenmakers et al., 2018), where model predictions are paired at the event level. By convention, we used the default Cauchy prior in JASP with a scale parameter of 0.707. The Bayes factor provides an indication as to whether the null hypothesis (i.e. that the two models being compared produce equally accurate forecasts) or the alternative hypothesis (i.e. that the two models don't produce equally accurate forecasts) is better supported by the data. We interpret these Bayes factors in accordance with the recommendations of Kass & Raftery (1995), summarised in Table 1.

TABLE 1: Interpretations for Bayes Factors (BF_{10})

Lower Bound	Upper Bound	Favoured Hyp.	Strength
1	3	Alternative	Weak
3	20	Alternative	Positive
20	150	Alternative	Strong
150	∞	Alternative	Very Strong
.333	1	Null	Weak
.05	.333	Null	Positive
.007	.05	Null	Strong
$-\infty$	.007	Null	Very Strong

3.3 Aggregation Algorithms

The predictions of each of the five aggregation algorithms were obtained as follows:

1. For the simple average (SA), the forecast for each event was calculated by taking an unweighted average of the probability forecasts of all forecasters in the crowd for that question.

2. For the MPW algorithm, a weight was constructed for each participant using the participants' probabilistic forecast that the event was true, $P_{i,k}$, and their probabilistic meta-prediction – their forecast about the average probability estimated by other people, $M^P_{i,k}$. Let $j = \{1, \ldots, N_k\}$ be the set of forecasters who answered forecasting problem k. A forecaster i's weight on decision problem k was given by

$$w_{i,k} = \frac{|P_{i,k} - M^P_{i,k}|}{\sum_{j=1}^{N_k} |P_{j,k} - M^P_{j,k}|},$$

where the numerator is the absolute difference between forecaster i's forecast and their probabilistic meta-prediction and the denominator is the sum of this absolute difference over all N_k forecasters. By construction, the weights assigned to the forecasters add up to 1. The forecast generated by the MPW algorithm for event k was the weighted average of each participants' probabilistic forecast:

$$T_{MPW}(X_k) = \sum_{i=1}^{N_k} w_{i,k} P_{i,k}.$$

3. For the Minimal Pivoting (MP) algorithm, the forecast for each event was the average forecast of the participants plus a correction term that was equal to the difference between the average forecast and the average meta-prediction (Palley & Soll, 2019). For example, if the average forecast was 0.8 and the average meta-prediction was 0.7, then the prediction from the minimal pivoting algorithm would be 0.9. Conversely, if the average forecast was 0.7 and the average meta-prediction was 0.8, then the prediction from the minimal pivoting algorithm was 0.6. The algorithm's forecasts were always bounded between 0 and 1.

4. For within-domain weighting, we applied the CWM algorithm (equation 12). To estimate forecasters' weights on each question, we used leave-one-out cross-validation. For example, for the NFL dataset, one of the 50 NFL questions was selected (the 'test' question) and the remaining 49 NFL questions were used to calculate the forecasters' mean contributions. We then normalised the forecasters' contributions (so that the total of all forecasters' contributions summed to one) and applied them as linear weights to forecasters' probability forecasts on the test question in order to generate an aggregated prediction for the test question. As with the original implementation of the CWM algorithm, forecasters whose contributions were negative were assigned weights of zero (Budescu & Chen, 2015). This procedure was repeated for all 50 NFL questions.

5. For cross-domain weighting, we also used the CWM algorithm, but now trained the algorithm on data obtained from one or more other domains. We will refer to this as the xCWM algorithm. Following the example above, the 50 NFL questions in Experiment 1 were used to calculate the weights used to aggregate forecasts in the Science trivia domain. Similarly, the 50 Science trivia questions were used to calculate the weights used to aggregate forecasts in the NFL trivia domain. As before, forecasters whose contributions were negative were assigned weights of zero. Since the training set was identical for each question in a particular domain, this meant that the exact same set of contribution weights was used for all forecasts within the same domain.

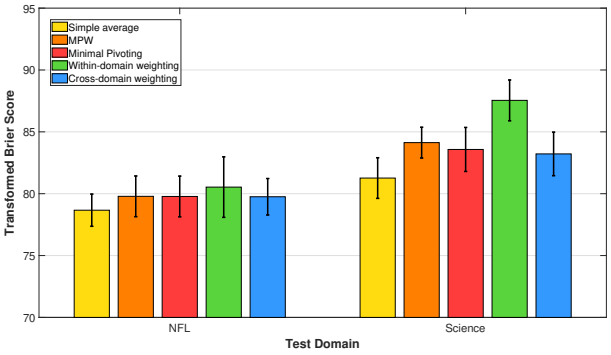

FIGURE 1: Results from Experiment 1 showing the mean score for each algorithm on the NFL Trivia (left) and Science Trivia domain (right). Error bars show the standard error.

3.4 Results

Figure 1 shows the mean performance of each algorithm separately for the NFL Trivia and Science Trivia domains. Table 2 shows the Bayes Factor comparing the evidence for the alternative hypothesis over the null hypothesis between each comparison algorithm and cross-domain weighting as implemented by the xCWM algorithm. The comparisons are first made in the NFL trivia (NFL) and Science Trivia (Sci) domains separately, and then shown with both domains combined (All). For all comparisons, the null hypothesis is that the two models produce equally accurate forecasts and the alternative hypothesis is that one of the models produce more accurate forecasts than the other. Because Bayes' factor is calculated using a stochastic algorithm, different calculation runs can result in different values for Bayes' factor. The column "Error %" shows the expected percentage difference in Bayes' Factor between different calculation runs. In all cases the expected error is very small, showing that the calculation of the Bayes' Factor is very stable.

In the NFL trivia domain, there was very little difference in performance between the xCWM algorithm and the other algorithms. In the Science trivia domain, cross-domain weighting performed approximately as well as the MPW algorithm and Minimal Pivoting algorithm. There was some evidence that the xCWM algorithm outperformed simple averaging and very strong evidence that it performed worse than the CWM algorithm.

Overall, there was some evidence that the xCWM algorithm performed better than simple averaging and approximately equally as well as the MPW algorithm and the Minimal Pivoting algorithm. There was strong evidence that the xCWM algorithm performed worse than the CWM algorithm. There therefore appeared to be very little difference in performance between cross-domain weighting and single-question aggregation approaches in both domains.

TABLE 2: Experiment 1 – Bayes factors for model comparisons

Algorithm 1	Algorithm 2	Domain	BF_{10}	Error %
SA	xCWM	NFL	0.351	3.110e-6
MPW	xCWM	NFL	0.154	1.966e-6
MP	xCWM	NFL	0.154	1.978e-6
CWM	xCWM	NFL	0.185	2.412e-6
SA	xCWM	Sci	9.246	4.893e-7
MPW	xCWM	Sci	0.250	2.899e-6
MP	xCWM	Sci	0.207	2.630e-6
CWM	xCWM	Sci	7518.929	2.521e-11
SA	xCWM	All	6.140	2.701e-6
MPW	xCWM	All	0.156	7.950e-5
MP	xCWM	All	0.125	9.608e-5
CWM	xCWM	All	20.116	8.370e-7

3.5 Discussion

Consistent with our first prediction, we found that cross-domain weighting was generally less effective than within-domain weighting. Consistent with our second prediction, cross-domain weighting outperformed simple averaging. These results continued to hold when considering only the Science data but did not hold when considering only the NFL data. For the NFL data, there was only weak evidence that the cross-domain weighting was superior to simple averaging and there was no evidence that it was worse than within-domain weighting. So, why did we obtain different results for the NFL data as opposed to the Science data?

The key point is that because the performance of the xCWM algorithm cannot exceed that of the CWM algorithm, the xCWM can only perform well if the CWM algorithm per-

forms well. Because the CWM algorithm performed poorly for the NFL data set, doing little better than simple averaging, the xCWM also could not perform significantly better than simple averaging. So why didn't the CWM algorithm perform well for the NFL data set? The fact that the CWM algorithm did not perform well implies that it was unable to identify experts for this data set. According to our model, this implies that the within-domain noise n_i was large. If the within-domain noise is large, this means that a person's prior performance is not a reliable indicator of their future performance. It follows from equation 10 and the definition of α_i given in equation 7 that if the within-domain noise is large, the correlation between performance in the NFL domain and performance in the Science domain will be small. In other words, because within-domain noise makes the measurement of expertise in the NFL domain unreliable, expertise in the NFL domain cannot reliably be used to predict expertise in the Science domain. Consequently, the xCWM algorithm will perform poorly in the Science domain if it uses the weights obtained from the NFL domain, as these are unreliable. This explain why the performance of the xCWM algorithm was so much less than that of the CWM algorithm in the Science domain.

Turning our attention to our third prediction, we found that cross-domain weighting offered no advantage over either the MPW algorithm or the Minimal Pivoting algorithm in either domain. Intuitively, this makes sense. For the reasons outlined above, the xCWM could not perform well in this setting. It is therefore not surprising that it could not outperform these two single-question aggregation approaches.

4 Experiment 2

As discussed in section 2, our theoretical analysis predicts that cross-domain weighting will be more effective when more domains are used to calculate the cross-domain weights. Experiment 2 tested this prediction. Alongside the two domains NFL and Science from Experiment 1, we added an additional domain, Emotional Intelligence (EI). Expertise in this domain was measured using questions adapted from the Situational Test of Emotional Understanding and Situational Test of Emotion Management (MacCann & Roberts, 2008). We chose EI questions because they theoretically tap into a different fundamental set of skills and characteristics compared to NFL Trivia and Science Trivia (e.g., Cattell, 1963; Lam & Kirby, 2002; Mayer et al., 2001).

We make the following three predictions: (1) Because the xCWM algorithm cannot outperform the CWM algorithm, the xCWM algorithm will still perform poorly on the NFL questions, because the CWM algorithm performs poorly on these questions, due to the large amount of noise in this domain. (2) Assuming that the CWM algorithm significantly outperforms simple averaging in the EI domain, then we should be able to use questions from the EI domain to reliable identify generalised expertise, g. Consequently, when the xCWM is trained using data obtain, in part, from this domain, it should perform well in the Science domain. (3) We should be able to use performance in the Science domain to reliably identify generalised expertise g. Consequently, assuming that CWM algorithm significantly outperforms simple averaging in the EI domain, training the xCWM algorithm using data that includes questions from the Science domain should allow it to perform well in the EI domain.

4.1 Methods

We used the same experimental paradigm and set of questions as in Experiment 1, but included an additional set of 50 trials where participants were presented with EI statements. The EI questions were adapted from the Situational Test of Emotional Understanding and Situational Test of Emotion Management developed and validated by MacCann & Roberts (2008). We chose to adapt questions from this source due to the fact the test relies on questions with objectively-correct answers rather than self-report scales with no objective answer. Fifty questions, which were originally in the form of four-alternative multiple choice, were randomly selected from these two tests and converted into two-alternative questions after removing two of the three possible incorrect options. While adapting these tests in such a way may reduce their validity as a measure of EI, these questions should still capture to some extent participants' expertise in the EI domain. The full list of questions used in this experiment and the responses for each participant are available to download in the supplementary materials.

Participants were paid USD $6.00 for participating in the experiment and responded to 150 trials in total over 3 domains: 50 trials about NFL Trivia, 50 trials about Science Trivia, and 50 trials about EI. As before, we collected responses from 100 participants. Only respondents inside the US were able to participate, and participants from any of our previous experiments was prevented from participating in this experiment. Three participants reported that they cheated during the experiment or failed to complete the survey and were therefore removed from the analyses. The analyses were conducted on the remaining 97 people.

4.2 Analyses

The analysis was the same as in Experiment 1 except that the xCWM was always trained on data from the two domains for which it was not making a prediction. For example, if it was making prediction for the NFL data set, it would be trained using both the Science data and the EI data. We note that by estimating weights from the other two domains, there

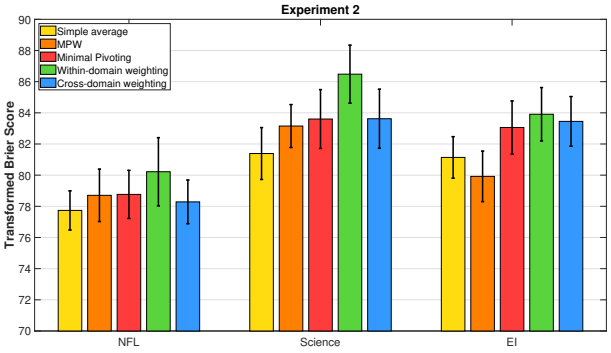

FIGURE 2: Results from Experiment 2 showing the mean score for each algorithm on questions from each domain. Error bars show the standard error.

are twice as many questions in the training set for cross-domain weighting compared to the training set for within-domain weighting. We show in Section 6 that our general findings hold even after ensuring both algorithms use the same number of training questions.

4.3 Results

Table 3 shows the Bayes Factor comparing the evidence for the alternative hypothesis over the null hypothesis for each comparison in the EI domain. Consistent with the finding from Experiment 1, there was positive evidence that the CWM algorithm outperformed the xCWM algorithm and there was very strong evidence that the xCWM algorithm outperformed simple averaging. However, there was no evidence that it outperformed either of the single question algorithms.

Figure 2 shows the mean performance of each algorithm in the Emotional Intelligence domain. As expected the xCWM algorithm did not perform well in the NFL domain, presumably because of the poor performance of the CWM algorithm in this domain. Also, as expected, in performed better in the Science and NFL domains.

TABLE 3: Experiment 2 – Bayes factors for model comparisons

Algorithm 1	Algorithm 2	BF_{10}	error %
SA	xCWM	2886.292	1.861e-8
MPW	xCWM	0.374	2.642e-4
MP	xCWM	0.091	0.001
CWM	xCWM	12.619	6.267e-6

4.4 Discussion

Our results were consistent with our expectations. The xCWM algorithm performed better than simple averaging but worse than the CWM algorithm. It performed relatively well in the Science and EI domains, but did not perform well in NFL domain, presumably because of the poor performance of the CWM algorithm in this domain. Crucially, there was no evidence that the xCWM outperformed either the MPW algorithm or the MP algorithm. Presumably, this was caused in part by its poor performance in the NFL domain but also because the noise in the NFL domain meant that performance in this domain could not be used to reliably predict performance in other domains, thereby reducing the performance of the xCWM algorithm in those domains.

5 Experiment 3

So far, we have failed to find any evidence that the xCWM algorithm outperforms either the MPW algorithm or the MP algorithm. We have argued that this is caused by the noise present in the NFL domain. This noise means that both the CWM algorithm and the xCWM algorithm perform poorly in this domain. In addition, this noise means that it is hard to use performance in this domain to estimate general expertise, g, which, in turn, reduces the performance of the xCWM algorithm in the Science and EI domains. In Experiment 3 we test the conjecture that the poor performance of the xCWM algorithm is caused by the noise in the NFL domain by replacing the NFL domain with another domain, one which we hoped would contain less noise. We reasoned that if this were to occur, then we expect that the xCWM algorithm will then be able to outperform both the MPW and the MP algorithms.

In Experiment 3, we replaced the NFL questions with questions requiring judgement of prices of professional and amateur artworks. We elicited forecasters' judgements about the prices of artworks because they require a fundamentally different set of skills compared to questions from the Science or EI domains. This is important because in the theoretical section we assumed that domain-specific expertise was not correlated between domains. We hoped that the Art domain would be less noisy than the NFL domain, to better allow us to estimate general expertise, g. We predicted that the xCWM algorithm would now outperform both the MPW algorithm and the MP algorithm.

5.1 Methods

We adopted the same methodology as the previous experiment. We used the same set of questions from the Science and EI domains, and replaced the set of questions from the NFL domain with a set of questions where participants were asked to judge the value of different artworks. On each of

the *Art* trials, participants were presented with an image of an artwork and asked a) whether the market price of the original version of that artwork would exceed USD $10,000, (2) the probability that their statement was "true" (from 0–100), and (3) what they thought the average probability estimated by other people on question (2) would be (from 0-100). As a test of the reliability of our findings, we conducted the experiment twice, using a separate sample of participants in each experiment and a small change in the elicitation procedure: In the first survey (Experiment 3a), participants were asked to provide votes and probability forecasts that were consistent (i.e., providing a probability forecast greater than 0.5 when predicting 'True' and a probability less than 0.5 when predicting 'False'), and we subsequently excluded any forecasts that were inconsistent from our analyses. In the second survey (Experiment 3b), participants were forced to provide votes and probability forecasts that were consistent in order to proceed.

The artworks presented to participants were taken from an online website listing professional artworks along with their prices (Sotheby auctions), online websites listing famous historical artworks, and online websites selling original amateur artworks (e.g., Etsy). After the 'low-priced' artwork images were selected, they were double-checked using Google reverse-image search to ensure that they were not sold or listed on any website for more than USD $10,000. The full list of questions and art images used in this experiment, as well as the data we collected, are available to download in the supplementary materials.

We reduced the number of questions from each domain to 40 per domain. Participants therefore completed 120 trials in total. All 120 questions were presented in a randomized order. Each participant was paid USD $4.00 for completing the experiment. Participants from any of our previous experiments was prevented from participating in this experiment. In both Experiments 3a and 3b, we collected responses for 100 participants, and we then excluded participants who reported cheating during the experiment as we did for the previous experiments. In Experiment 3a, 12 participants were excluded and analyses were conducted on the remaining 88 people. In Experiment 3b, 21 were excluded and analyses were conducted on the remaining 79 people.

We repeated the same set of analyses as before for both Experiment 3a and 3b. Our main comparison is between the mean transformed Brier score of within-domain weighting algorithm and cross-domain weighting, aggregated across all 120 questions in the dataset. As before, we computed 95% CIs for the mean difference in score between cross-domain weighting and within-domain weighting, generally across all 120 questions. Lastly, we also tested whether cross-domain weighting generally outperformed the MPW algorithm and the Minimal Pivoting algorithm, and examined the difference in improvement offered by cross-domain weighting relative to these single-question aggregation approaches. All algorithms were implemented in the same way as before.

5.2 Results

Figure 3 shows the mean performance of each algorithm separately on each of the three domains. Table 4 shows the Bayes Factor comparing the overall evidence for the alternative hypothesis over the null hypothesis between each comparison model and cross-domain weighting (xCWM) across all three domains. For example, when the xCWM algorithm was required to make predictions in the EI domain, the contribution weights were derived from the Science and Art domains. Similarly, when it was required to make predictions in the Art domain, the contribution weights were derived from the Science and EI domains. As we predicted, there was very strong evidence that the xCWM algorithm outperformed simple averaging and the Minimal Pivoting algorithm, and positive evidence that cross-domain weighting outperformed the MPW algorithm. As we also predicted, the performance of the xCWM improved relative to the CWM algorithm and there was now weak evidence supporting the null hypothesis that the performance of these two algorithms was equivalent.

Experiment 3b found the exact same pattern of results, with even stronger evidence favouring the cross-domain weighting model compared to all other models. Thus, across our results in both Experiments 3a and 3b, cross-domain weighting offers a similar level of performance as within-domain weighting, and also consistently outperformed both single-question aggregation approaches.

TABLE 4: Experiment 3 – Bayes factors for model comparisons

Exp.	Algorithm 1	Algorithm 2	BF_{10}	Error %
3a	SA	xCWM	1047.591	4.497e-9
3a	MPW	xCWM	12.274	2.957e-8
3a	MP	xCWM	431.860	7.788e-9
3a	CWM	xCWM	0.398	1.513e-5
3b	SA	xCWM	166376.666	7.542e-11
3b	MPW	xCWM	81.892	1.451e-8
3b	MP	xCWM	19682.746	2.351e-8
3b	CWM	xCWM	0.340	1.888e-5

5.3 Discussion

Results from this experiment provide compelling evidence of the effectiveness of cross-domain weighting. Overall, cross-domain weighting offered a similar level of performance as

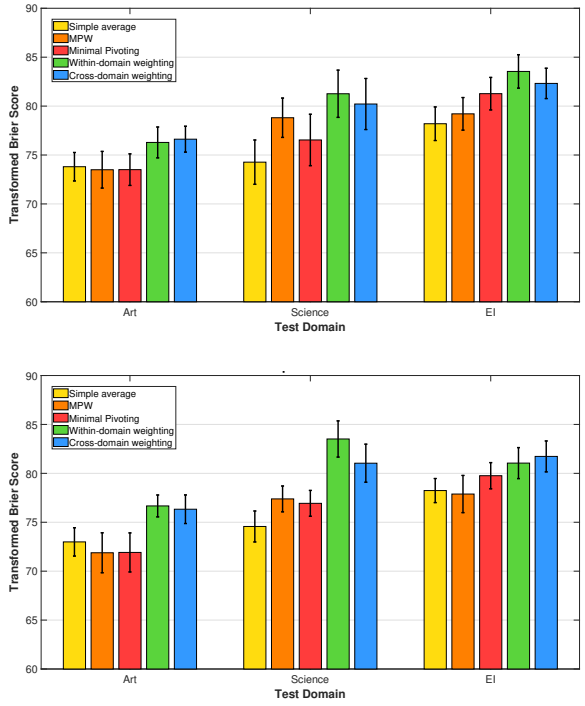

FIGURE 3: The mean score for each model in Experiment 3a (top panel) and Experiment 3b (bottom panel) on the Art questions (left), Science Trivia questions (centre), and Emotional Intelligence questions (right). Error bars show the standard error. Both experiments show a similar pattern of results: cross-domain weighting (light-blue bar) performs generally as well as within-domain weighting (light-green bar).

within-domain weighting, indicating that forecasters' expertise could be estimated almost as effectively from unrelated training domains as from each test domain. Given that expertise in the judgement of art prices, science trivia, and EI questions all theoretically rely on fundamentally different sets of skills and knowledge, our results suggest that forecasters' contributions are a highly generalised metric of expertise. Furthermore, these results suggest that the relationship between the test and training domains have a relatively small effect on our ability to estimate forecasters' expertise.

Our results also indicate that the cross-domain weighting approach can provide substantial improvements in forecasting performance over contemporary single-question aggregation approaches. By quantifying the improvement offered by cross-domain weighting over these single-question aggregation approaches, our results therefore capture the trade-off in performance between eliciting responses from forecasters on an unrelated set of questions with known outcomes and simply estimating their expertise using single-question approaches.

6 Simulating changes in training set size

A potential limitation of our results thus far is that cross-domain weighting appears to require a larger set of training problems in order to obtain a similar level of performance as within-domain weighting. In this section, we explore how the CWM algorithm and the xCWM algorithm compare when we restrict the two algorithms to use the same number of training questions. We use a bootstrapping method to simulate changes in the training set size for each dataset from all our experiments, and compared the performance of within-domain weighting and cross-domain weighting for different training set sizes.

We simulated the change in mean transformed Brier score for within-domain weighting and cross-domain weighting over different sized training sets by using the bootstrap (Efron & Tibshirani, 1994) to resample data from each of the three experiments. For each test question (i.e. for each question for which we wanted to make a prediction), we divided forecasters' forecasts on training questions into (1) a pool for questions from same domain as the test question, or (2) a pool for questions from other domain(s). We then resampled questions from each pool, while varying the number of training questions used on each iteration.

To ensure that both within-domain and cross-domain pools contained the same number of training questions, on each bootstrap iteration a subset of questions equal to the number of training questions in the within-domain pool (since it contains fewer questions than the cross-domain pool) was randomly selected without replacement from the cross-domain pool of training questions. Forecasters' responses to the remaining questions in the cross-domain pool were removed for that iteration and this ensured the training set for within-domain weighting and cross-domain weighting were both drawn from pools with the same number of potential training questions. A random set of k training questions was then randomly sampled with replacement from each pool. We repeated this process 1,000 times for each training set size in the range of $\{10, 20, 30, 40, 49\}$ for data from Experiments 1 and 2, and for Experiments 3, the range was $\{10, 20, 30, 39\}$ because the datasets for this experiment were smaller.

On each of the 1,000 iterations, we calculated the performance of within-domain weighting and cross-domain weighting, which we then averaged across the 1,000 iterations to obtain a single score for each algorithm on that test event, calculated using k training questions. We then averaged each model's performance across each test event in the dataset to obtain that model's mean score for that experiment. We applied the same simulation procedure for each dataset. As a reference, we also calculated the mean score of the simple average on the original sample in each dataset. As an inferential test for the difference in score between within-domain weighting and cross-domain weighting, we

computed the overall performance of cross-domain weighting and within-domain weighting over all the experiments using the largest training set size for the simulations from the respective dataset.

Figure 4 shows the mean score of cross-domain weighting and within-domain weighting across different training set sizes for each dataset. We can see that in all four figures, the difference in score between within-domain weighting (red line) and cross-domain (blue line) weighting was fairly consistent regardless of training set size. Computing a Bayes factor for the comparison between within-domain weighting and cross-domain weighting using all four training sets, we find positive evidence for the alternative hypothesis that within-domain weighting yields more accurate forecasts than cross-domain weighting ($BF_{10} = 4.44$), which is consistent with our theoretical model. However, the difference in score between within-domain weighting and cross-domain weighting appears to be particularly small on smaller set sizes. Our results demonstrate that the efficacy of cross-domain weighting can therefore be expected to generalise to decision and forecasting problems with few training events.

More importantly, we can see there was very little difference in score between cross-domain weighting and within-domain weighting in all four training sets, regardless of the number of training events used. Thus, while some of our original cross-domain estimates used more training events to estimate forecasters' contributions, our results here suggest that the benefits of this larger training sample are small.

Overall, these simulation results suggest that forecasters' expertise can be estimated efficiently using questions from other domains and the cross-domain weighting model also appears to be highly robust under different training set sizes. These results are consistent with those of Chen et al. (2016), who demonstrated the robustness of the standard CWM across different training set sizes. Our results in this section show that the robustness of the contribution metric also extends to contributions estimated from domains that are theoretically unrelated to the test domain.

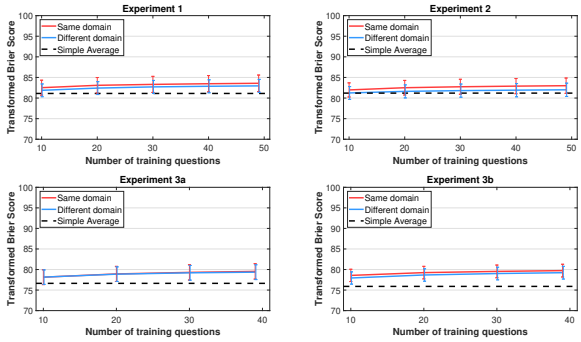

FIGURE 4: Simulations using data from Experiments 1 – 3 showing the mean transformed Brier score for within-domain weighting (red) compared to cross-domain weighting (blue) over different training set sizes. Error bars show the standard error. The performance of the simple average (dashed line), which does not use training data, is shown for reference.

7 General Discussion

The aim of the current paper was to investigate whether forecasters' expertise could be accurately identified using their performance on decision problems in unrelated domains. Over three experiments, we examined the performance of cross-domain weighting relative to within-domain weighting, the MPW algorithm (Martinie et al., 2020), and the Minimal Pivoting algorithm (Palley & Soll, 2019), using several different question domains, including questions from NFL trivia, Science trivia, a test of Emotional Intelligence, and judgments about the price of professional and amateur artworks. Our results demonstrated that cross-domain weighting can be almost as effective as within-domain weighting providing that (1) multiple domains are used to calculate the cross-domain weights and (2) these domains are not too noisy (i.e., differences in forecasters' responses reflect genuine variability in expertise). If these domains are too noisy, then they are not helpful for estimating generalised expertise, g. Both these findings follow directly from our theoretical model discussed in Section 2. As indicated by that model, the correlation between performance in the domain of interest and performance averaged across all other domains increases as the set size increases providing $\overline{\alpha_\Omega}$ does not increase. Adding a domain that is too noisy will increase $\overline{\alpha_\Omega}$ so may not improve your estimate of generalised expertise. We argued that this is what occurred with the NFL data. This data was so noisy that the CWM algorithm could not reliable identify expertise, and therefore performed similarly to simple averaging. Performance in this domain was therefore not helpful in estimating generalised expertise. The key insight is that when estimating generalised expertise, use as many unrelated domains as possible but only use domains where there is clear evidence of differing levels expertise. If the data in a domain is so noisy that you cannot reliably identify higher performing individuals, then that domain should not be used in the estimation of forecasters' expertise.

Our results provide novel insight into the generality of expertise across a range of decision-making domains. Previous applications of expertise-identification approaches such as the CWM algorithm (Budescu & Chen, 2015) have been largely limited to estimating forecasters' expertise by their performance within similar or identical domains to the questions of interest (Cooke, 1991; Mellers et al., 2015). While the CWM's robustness across different crowd sizes, crowd compositions, and number of training questions have been demonstrated previously (Budescu & Chen, 2015; Chen et al., 2016), no study to date has examined the extent to which forecasters' contributions can be estimated from their

performance on unrelated domains. Here, we have shown that forecasters' contributions can be estimated effectively using their performance on unrelated domains, compared to their performance on similar domains. We have therefore demonstrated the contribution-metric to be even more versatile for extracting and identifying expertise than what has been shown in the existing literature.

Our results also show that cross-domain weighting may be more effective than existing single-question aggregation approaches when multiple cross-domains are available. While previous research has shown that single-question aggregation approaches can be useful for leveraging expertise when forecasters' performance on the relevant domains are unknown (Martinie et al., 2020), the current results suggest that a better alternative might be to estimate forecasters' expertise on unrelated domains. As our simulations show, forecasters' expertise can be estimated effectively with as few as 20 training questions, regardless of domain, and provides an improvement over simple averaging that is consistently several times larger than that provided by single-question aggregation approaches. The cross-domain weighting approach is therefore an attractive and effective alternative for decision makers seeking to improve forecasts on novel problems for which there are no records of forecasters expertise.

Based on our findings, future research might wish to examine other ways at improving cross-domain weighting approaches. For example, it may be possible to improve the forecasts of cross-domain weighting further by selecting a crowd of fewer but better-performing experts, as have been demonstrated for the CWM algorithm (Chen et al., 2016). While this was beyond the purview of our current study, we hope that our results will inspire future researchers to examine the efficacy cross-domain weighting approaches more generally, for example, in other problem domains or by combining it with other aggregation approaches.

References

Armstrong, J. S., Ed. (2001). *Principles of forecasting: A handbook for researchers and practitioners*, volume 30. Springer Science & Business Media.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Budescu, D. V. & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1.

Chen, E., Budescu, D. V., Lakshmikanth, S. K., Mellers, B. A., & Tetlock, P. E. (2016). Validating the contribution-weighted model: Robustness and cost-benefit analyses. *Decision Analysis*, *13*(2), 128–152.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583.

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

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79.

Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kurvers, R. H., et al. (2019). How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances*, *5*(11), eaaw9011.

Lam, L. T. & Kirby, S. L. (2002). Is emotional intelligence an advantage? An exploration of the impact of emotional and general intelligence on individual performance. *Journal of Social Psychology*, *142*(1), 133–143.

MacCann, C. & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: theory and data. *Emotion*, *8*(4), 540.

Martinie, M., Wilkening, T., & Howe, P. D. (2020). Using meta-predictions to identify experts in the crowd when past performance is unknown. *PLOS One*, *15*(4), e0232058.

Mayer, J., Salovey, P., Caruso, D., & Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion (Washington, DC)*, *1*(3), 232.

McCoy, J. & Prelec, D. (2017). A statistical model for aggregating judgments by incorporating peer predictions. *arXiv preprint arXiv:1703.04778*.

Mellers, B., et al. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, *10*(3), 267–281.

Murphy, A. H. & Winkler, R. L. (1970). Scoring rules in probability assessment and evaluation. *Acta Psychologica*, *34*, 273–286.

Palley, A. B. & Soll, J. B. (2019). Extracting the wisdom of crowds when information is shared. *Management Science*, *65*(5), 2291–2309.

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532.

Satopää, V. A., Pemantle, R., & Ungar, L. H. (2016). Modeling probability forecasts via information diversity. *Journal of the American Statistical Association*, *111*(516), 1623–1633.

Soll, J. B. & Larrick, R. P. (2009). Strategies for revis-

ing judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Wagenmakers, E.-J., et al. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76.

Wilkening, T., Martinie, M., & Howe, P. D. (2021). Hidden experts in the crowd: Using meta-predictions to leverage expertise in single-question prediction problems. *Management Science*.

Winkler, R. L. (1989). Combining forecasts: A philosophical basis and some current issues. *International Journal of Forecasting*, *5*(4), 605–609.