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

Piers Howe* Marcellin Martinie† Tom Wilkening‡

October 26, 2022

Abstract

In recent years, a number of crowd aggregation approaches have been proposed to combine the judgments of different individuals in problems where decision makers do not have records of the individuals' past performance in that domain. However, it is often possible to obtain a measure of the individuals' past performance in other domains. The current paper explores the extent to which individuals' relative expertise in one domain can be used to weight their judgments in another 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 individuals' past performance to those that weight by individuals' 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 in some circumstances. We present a simple model of the relationship between within-domain and cross-domain performance and discuss the conditions under which cross-domain weighting is likely to be effective. Our results demonstrate the potential of cross-domain weighting in problems where records of individuals' past performance in the domain of interest are unavailable.

Keywords: wisdom of crowds, decision-making, aggregating judgments, expertise

*Melbourne School of Psychological Sciences, The University of Melbourne. ORCID: 0000-0001-6171-1381. Email: pdhowe@unimelb.edu.au Corresponding author.

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

‡Department of Economics, The University of Melbourne. ORCID: 0000-0001-8037-9951. Email: Tom.Wilkening@unimelb.edu.au

1 Introduction

Aggregation algorithms harness the 'Wisdom of Crowds' by aggregating judgments from multiple individuals to generate more accurate judgments than can be obtained from a single individual (Surowiecki, 2005). When aggregating judgments, such algorithms typically weight the individuals' judgments according to the individuals' past performance in the test domain (i.e. the domain for which judgments are needed), so as to obtain more accurate judgments than can be obtained by simply averaging the judgments of all individuals (Cooke, 1991; Armstrong, 2001; Clemen, 1989; Winkler, 1989; Budescu & Chen, 2015). But what should one do if one does not have a record of the individuals' past performance in the test domain?

Previous work has suggested that, in addition to domain-specific expertise, people have a general ability to make accurate judgments, reflecting attributes such as diligence, open-mindedness and intelligence (Mellers et al., 2015). Building on this previous work we investigated to what extent we could use each individual's past performance in domains other than the test domain to predict their performance in the test domain. We refer to our technique as *cross-domain weighting* since we weight judgments in the test domain based on the individuals' performances in the other domains.

We found that, in some circumstances, this cross-domain weighting technique produced estimates with similar accuracy as within-domain weighting and, even when it did not, always produced judgments at least as good as and usually better than the unweighted aggregation of the individual judgments. Additionally, we found that this technique produced at least as good, and sometimes better, judgments than the existing, prominent algorithms that are designed to optimally combine the judgments of multiple individuals when their past performance is unknown. Such algorithms are sometimes referred to as single-question algorithms because performance on previous questions of the same type is unavailable, so the task is to make a judgment based solely on responses to a single question (Prelec et al., 2017).

In the rest of this introduction, we will first summarise the literature on traditional aggregation algorithms and explain why we use the Top 5 algorithm (Mannes et al., 2014) and the Contribution Weighted Model (CWM) algorithm (Budescu & Chen, 2015) as the basis for our cross-domain weighting studies. We will then summarise the literature on single-question algorithms. We will explain which of these single-question algorithms typically perform well and use these algorithms as points of comparison for our cross-domain weighting algorithms. We will end by summarising the structure of the rest of the article.

1.1 Traditional Aggregation Algorithms

Because the errors of different individuals are usually not perfectly correlated, averaging the judgments of a group of individuals will tend to cancel out their individual errors and consequently will typically produce a more accurate judgment than the judgment of the average individual (Surowiecki, 2005; Clemen, 1989; Soll & Larrick, 2009; Davis-Stober et al., 2014). The predictive accuracy of the resultant judgments 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 the weight assigned to each individual's judgment is derived from that individual's performance on a set of seed questions with known outcomes. Individuals whose performance on the seed questions are below a theoretical threshold are assigned weights of zero, and their probability estimates are removed from the crowd. Such an approach is common in the judgment aggregation literature (Armstrong, 2001; Clemen, 1989; Winkler, 1989). More recently, it has been shown that excluding almost all the individuals and just averaging the judgments of the top five individuals produces very good judgments in a number of domains (Mannes et al., 2014). For this reason, we will utilise the Top 5 algorithm in our study.

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

1.2 Single-Question Judgments Algorithms

In recent years, a number of approaches have been developed to aggregate judgments in domains where individuals' past performance on related questions is unknown, the so called single-question problem (Palley & Soll, 2019; Prelec et al., 2017; Satopää et al., 2016; Kurvers et al., 2019; McCoy & Prelec, 2017). Some of these algorithms can only make binary (yes/no) judgments so cannot easily be applied to the problem sets that we consider (Wilkening et al., 2021). However, other single-question algorithms can predict the *probability* that an event will occur (Palley & Soll, 2019; Martinie et al., 2020). For example, Palley and Soll (2019) developed a theoretical framework to address the problem that when simple averaging is used and multiple individuals base their judgments on similar information, shared informa-

tion will become over-weighted in the individuals' reports and unique information that is available to only a small subset of individuals will become under-weighted. The authors argued that the information shared between individuals can be estimated using the individuals' meta-predictions about the average judgments of others. They further proposed an aggregation algorithm, called the minimal pivoting procedure, that uses the difference between the average judgment and the average meta-judgment to adjust the average judgment. Palley and Soll demonstrated theoretically that, under idealised circumstances, this minimal pivoting procedure would completely remove the over-weighting 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 judgment 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 problems.

More recently, Martinie et al. (2020) developed the Meta-Probability Weighting (MPW) algorithm as an alternative way of using the individuals' meta-predictions about the average judgments of others to improve the average judgment. The key insight of this algorithm is that the difference between the individual's own judgment and their meta-prediction of the average judgment of 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 judgment would be any different from the judgment of others. Thus, the difference between their judgment and their meta-judgment of the average judgment 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-judgment of the average judgment of others might be very different from their own judgment. Consequently, the larger the difference between an individual's probability estiimate and meta-prediction, the more likely they are to be an expert. The MPW algorithm consequently weights individuals by the difference between their probability estimates 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 the records of the individuals' past performance are unavailable. To date, few comparisons have been made between single-question approaches and aggregation approaches that use 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 an individual's past performance, albeit in unrelated domains. A central aim of this paper is therefore to examine the accuracy of single-question approaches relative to weighting by past 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 a different domain.

1.3 Summary of our Approach

Over three experiments, we evaluated the accuracy of different judgment-aggregation approaches. Specifically, we compared the accuracy of weighting by individuals' past performance on questions from the same domain (i.e., within-domain weighting) to weighting individuals by their performance on questions from other domains (i.e., cross-domain weighting) using both the Top 5 algorithm (Mannes et al., 2014) and 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 MP 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 describes the degree to which performance in one domain is likely to be correlated with the combined performance across one or more other domains. We then conduct three experiments. In these experiments, cross-domain weighting is sometimes highly effective but sometimes not. We are able to identify two reasons why cross-domain weighting is sometimes ineffective. Specifically, since cross-domain weighting is unlikely to be more effective than within-domain weighting, if within-domain weighting does not improve upon simple averaging, then cross-domain weight will likely be similarly ineffective. Additionally, even when within-domain weighting is effective, cross-domain weighting is likely to be ineffective if the best performers in one domain are mediocre to poor performers in the other domain. In other words, as explained by our model, cross-domain weighting is likely to be effective only if performance in both the test domain and in the cross-domain reflects general expertise (i.e. ability that is not domain-specific).

2 Theoretical Model

In this section, we develop a theoretical model that illustrates how an individual'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, referred to as the 'cross' domains. We will show that, providing certain caveats are met, an individual's average performance in the cross domains predicts their performance in the test domain. Building on this result, in the next section, we will use the average performance of individuals in the cross domains, to determine the weightings of the individuals in the test domain, a method we will refer to as 'cross-domain weighting'.

Let $d_1, \ldots, d_s$ be a set of s knowledge domains. An individual's performance, x_i, in domain d_i is modelled as a combination of (1) their specific expertise for that domain, e_i, which represents factors such as domain-specific knowledge, (2) their general expertise, g, which represents factors other than domain-specific expertise, such as diligence, engagement, and general intelligence, and (3) a noise term, n_i, which models the expected variation in performance on repeated tests in the same domain by the same individual. Although we assume g to be constant across all domains by virtue of the fact that it is not domain-specific, we assume that it may not affect performance equally in all domains. For example, in a domain where performance can be increased by exerting extra effort, we would expect diligent individuals to tend to score higher than less diligent individuals. Thus, for this domain we would expect g to affect performance more than in a domain where exerting extra effort has less of an effect on performance. We capture this possibility by multiplying g by a factor γ_i which denotes the degree to which g affects performance in domain i. Putting this all together, we model performance in domain i, x_i, as follows:

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

For each domain, we define the zero point for x_i as the average performance. We define the individual's domain-specific expertise, the individual's general expertise, and the noise term as latent variables that are normally distributed around zero with variances $\sigma^2_{e_i}$, σ^2_g, and $\sigma^2_{n_i}$ respectively, (i.e., $e_i \sim \mathcal{N}(0, \sigma^2_{e_i})$, $g \sim \mathcal{N}(0, \sigma^2_g)$, and $n_i \sim \mathcal{N}(0, \sigma^2_{n_i})$ for all i). We assume that at least some variation exists in general expertise (i.e., $\sigma^2_g > 0$). We note that an individual's domain-specific expertise can, and mostly likely will, vary between domains. We define two domains i and j as 'unrelated' if their domain-specific expertise in domain i is independent of their domain-specific expertise in domain j, such that $\mathbb{E}(e_i.e_j) = 0$. Otherwise, we describe the two domains as being 'related'. We also assume that general expertise is independent of domain-specific expertise such that $\mathbb{E}(g.e_i) = 0$ for all i, and 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.

To be clear, we are using the terms 'related' and 'unrelated' to refer to theoretical constructs that have meaning only in the context of this model. Furthermore, it is not possible to experimentally determine if two domains are related or unrelated because we would expect an individual's performance in one domain to be correlated with their performance in a second domain even if the two domains were unrelated, by virtue of the individual's general expertise.

Although the assumption that individuals have general (as opposed to domain-specific) expertise might seem surprising, there is support for it in the literature. In particular, previous research on geopolitical forecasting has shown that some individuals are much better at forecasting geopolitical events than others (Mellers et al., 2015). Such forecasters, termed 'superforecasters' (Tetlock & Gardner, 2015), outperformed U.S. intelligence analysts by 30%. Crucially, the skill of the superforecasters is unlikely to be due solely to domain-specific expertise as the range of topics covered in these events was very large (Tetlock & Gardner, 2015). Rather, people who performed highly in one domain tended to perform highly in other domains, indicating that they possess a superior general forecasting ability (Mellers et al., 2017). So, what contributes to this superior general forecasting ability? There appear to be a number of factors (Mellers et al., 2015; Friedman et al., 2018). First, superforecasters were more open-minded and were more intelligent than the average forecaster. Second, they were more motivated and updated their predictions more often (i.e. they were more diligent). Third, they acquired task-specific probability skills and expressed their beliefs in a more nuanced way. Fourth, they were always members of a team and had more interaction with their teammates than 'regular' forecasters. In our model, we use the term general expertise, g, to capture the concept that some people generally perform better than others, regardless of the domain. We are agnostic as to what gives rise to it, believing that it is probably due to a combination of factors, as described above (Mellers et al., 2015; Friedman et al., 2018).

Using this model, we can determine the extent to which an individual's performance in one domain is expected to be linearly correlated with their performance averaged across a number of other domains. For simplicity, let us assume that all the domains are unrelated, thus $\mathbb{E}(e_i.e_j) = 0$ for all $i \neq j$. Let d_θ represent the domain of interest and d_{Ω_s} represent all other domains in the set of s domains (i.e., $d_{\Omega_s} := \{d_1, \ldots, d_s\} \setminus \{d_\theta\}$). We define x_θ as the individual's performance in the domain of interest and $\overline{x}_{\Omega_s}$ as the individual's performance averaged across all other s domains (i.e. across Ω_s). Thus,

$$\overline{x}_{\Omega_s} := \frac{1}{s-1}\sum_{i\neq\theta}(\gamma_i g + e_i + n_i). \tag{2}$$

The correlation between x_θ and x_{Ω_s} is given by

$$\rho_{\theta\Omega_s} = \frac{\mathbb{E}(x_\theta.x_{\Omega_s}) - \mathbb{E}(x_\theta)\mathbb{E}(x_{\Omega_s})}{\sqrt{\mathbb{E}(x_\theta^2) - (\mathbb{E}(x_\theta))^2}\sqrt{\mathbb{E}(x_{\Omega_s}^2) - (\mathbb{E}(x_{\Omega_s}))^2}}. \tag{3}$$

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

$$\rho_{\theta,\Omega_s} = \frac{\mathbb{E}(x_\theta.x_{\Omega_s})}{\sqrt{\mathbb{E}(x_\theta^2)}\sqrt{\mathbb{E}(x_{\Omega_s}^2)}}. \tag{4}$$

We consider each component of this equation separately and begin by considering the numerator,

$$\mathbb{E}(x_\theta.x_{\Omega_s}) = \frac{\gamma_\theta g^2}{(s-1)}\sum_{i\neq\theta}\gamma_i. \tag{5}$$

We define

$$\overline{\gamma}_{\Omega_s} = \frac{1}{(s-1)}\sum_{i\neq\theta}\gamma_i, \tag{6}$$

where $\overline{\gamma}_{\Omega_s}$ represents the average γ across the cross domains, Ω_s. We also note that $\sigma_g^2 = \mathbb{E}(g^2) - \mathbb{E}(g)^2$. Because $\mathbb{E}(g) = 0$, it follows that $\sigma_g^2 = \mathbb{E}(g^2)$. Substituting these equalities into equation 5 we find:

$$\mathbb{E}(x_\theta.x_{\Omega_s}) = \gamma_\theta\overline{\gamma}_{\Omega_s}\sigma_g^2. \tag{7}$$

We now turn our attention to the first denominator of equation 4:

$$\mathbb{E}(x_\theta^2) = \mathbb{E}(\gamma_\theta^2 g^2 + e_\theta^2 + n_\theta^2). \tag{8}$$

From above, we known $\mathbb{E}(g^2) = \sigma_g^2$, $\mathbb{E}(e_\theta^2) = \sigma_{e_\theta}^2$ and $\mathbb{E}(n_\theta^2) = \sigma_{n_\theta}^2$. Substituting these equalities into equation 8 we find:

$$\mathbb{E}(x_\theta^2) = \gamma_\theta^2\sigma_g^2 + \sigma_{e_\theta}^2 + \sigma_{n_\theta}^2. \tag{9}$$

We now turn our attention to the second denominator of equation 4:

$$\mathbb{E}(x_{\Omega_s}^2) = \mathbb{E}\left(\frac{1}{(s-1)^2}\left(\sum_{i\neq\theta}\sum_{k\neq\theta}\gamma_i\gamma_k g^2 + \sum_{i\neq\theta}(e_i^2 + n_i^2)\right)\right). \tag{10}$$

Using equation 6 and the above equalities, equation 10 simplifies to

$$\mathbb{E}(x_{\Omega_s}^2) = \overline{\gamma}_{\Omega_s}^2\sigma_g^2 + \frac{1}{(s-1)^2}\sum_{i\neq\theta}(\sigma_{e_i}^2 + \sigma_{n_i}^2). \tag{11}$$

We now define

$$\alpha_i := \frac{\sigma_{e_i}^2 + \sigma_{n_i}^2}{\sigma_g^2}. \tag{12}$$

α_i represents the ratio of the combined variance due to domain-specific expertise and noise relative to the variance in g. As such, it represents the degree to which variation in performance is determined by factors other than g. Substituting equation 7, equation 9 and equation 11 into equation 4 and using equation 12 to simplify the result, we find

$$\rho_{\theta,\Omega_s} = \frac{\gamma_\theta\overline{\gamma}_{\Omega_s}}{\sqrt{\gamma_\theta^2 + \alpha_\theta}\sqrt{\overline{\gamma}_{\Omega_s}^2 + \frac{1}{(s-1)^2}\sum_{i\neq\theta}\alpha_i}}. \tag{13}$$

We define

$$\overline{\alpha}_{\Omega_s} := \frac{1}{(s-1)}\sum_{i\neq\theta}\alpha_i. \tag{14}$$

where $\overline{\alpha}_{\Omega_s}$ represents α_i averaged over all cross domains (i.e. set Ω_s). Substituting equation 14 into equation 13 we find

$$\rho_{\theta,\Omega_s} = \frac{\gamma_\theta\overline{\gamma}_{\Omega_s}}{\sqrt{\gamma_\theta^2 + \alpha_\theta}\sqrt{\overline{\gamma}_{\Omega_s}^2 + \frac{\overline{\alpha}_{\Omega_s}}{(s-1)}}}. \tag{15}$$

We can now ask what happens if we were to add another cross domain to set Ω_s. The index of this additional domain would be s+1. The new set of cross domains would be designated by Ω_{s+1} and equal to $\{1,\ldots,s+1\}\setminus\{d_\theta\}$. For the sake of argument, let us assume that γ_{s+1} is equal to the mean γ averaged over set Ω_s and that α_{s+1} is equal to the mean α averaged over set Ω_s, i.e.

$$\gamma_{s+1} = \overline{\gamma}_{\Omega_s} \tag{16}$$

$$\alpha_{s+1} = \overline{\alpha}_{\Omega_s}. \tag{17}$$

From this it follows

$$\overline{\gamma}_{\Omega_{s+1}} = \overline{\gamma}_{\Omega_s} \tag{18}$$

$$\overline{\alpha}_{\Omega_{s+1}} = \overline{\alpha}_{\Omega_s}. \tag{19}$$

From equation 15, it follows that, under these conditions, adding this additional cross domain necessarily increases ρ. In other words

$$\rho_{\theta\Omega_{s+1}} > \rho_{\theta\Omega_s}. \tag{20}$$

But adding an extra cross domain does not always result in ρ increasing. Again assuming that $\overline{\gamma}_{\Omega_{s+1}} = \overline{\gamma}_{\Omega_s}$ but now allowing $\overline{\alpha}$ to vary, we see from equation 15 that $\rho_{\theta\Omega_{s+1}} = \rho_{\theta\Omega_s}$ when

$$\overline{\alpha}_{\Omega_{s+1}} = \frac{s}{s-1}\overline{\alpha}_{\Omega_s}. \tag{21}$$

Since α_i represents the ratio of the combined variance due to domain-specific expertise and noise relative to the variance in g, this shows that ρ will not increase if the variance of domain-specific expertise and noise in the new domain, $s+1$, is too large relative to the variance in g because then performance in this domain cannot be used to reliably estimate general expertise.

Let us now assume that $\overline{\alpha}_{\Omega_{s+1}} = \frac{s}{s-1}\overline{\alpha}_{\Omega_s}$. From equation 15 it is clear that $\rho_{\theta\Omega_{s+1}} > \rho_{\theta\Omega_s}$ if $\overline{\gamma}_{\Omega_{s+1}} > \overline{\gamma}_{\Omega_s}$ but $\rho_{\theta\Omega_{s+1}} < \rho_{\theta\Omega_s}$ if $\overline{\gamma}_{\Omega_{s+1}} < \overline{\gamma}_{\Omega_s}$. Because γ_i represents the degree to which general expertise, g, influences performance in domain i, this again shows that adding another cross-domain increases ρ only if performance in this additional domain is sufficiently determined by g.

Taken together, our model suggests that when within-domain performance cannot be measured but such performance is influenced by general expertise, it may be possible to identify higher performing individuals in the test domain by using performance from one or more cross-domains. Cross-domains where (i) performance is reflective of general expertise, (ii) variation in domain-specific expertise across individuals is low, and (iii) idiosyncratic noise is low are most likely to be correlated with performance in the target domain and most likely to improve probability estimates in the test domain. Adding additional domains is likely to be beneficial as long as the additional domains are sufficiently reflective of g. However, there are diminishing returns to adding additional cross-domains.

Although one can identify expertise within a domain, it is not possible to determine what fraction of this expertise is general expertise and what fraction is domain-specific expertise. This limitation needs to be kept in mind when attempting to compare the effectiveness of cross-domain weighting relative to within-domain weighting.

We now summarise a series of experiments designed to explore how cross-domain weighting algorithms perform using the above theoretical model to guide the discussion of our results. As discussed in Section 1.2, for these experiments we utilised the Top 5 algorithm (Mannes et al., 2014) and the CWM algorithm (Budescu & Chen, 2015), as these are some of the prominent algorithms in the literature. The Top 5 algorithm identifies the top five individuals and averages their judgments. The CWM algorithm weights individuals not by their past performance but by their contribution towards improving the aggregated group judgment. However, because the contribution weight as determined by the CWM algorithm is highly correlated with an individual's performance, in practice, individuals are approximately weighted by their past performance. Consequently, the CWM algorithm typically exceeds the performance of the simple average. However, because it aggregates judgments from most individuals, as opposed to aggregating judgments from the top performers, it typically does not perform as well as the Top 5 algorithm in situations where a small number of individuals have significantly greater expertise than others. We study cross-domain weighting with both methods since CWM may offer additional robustness properties in cases where the selection of the Top 5 performing individuals is a result of domain-specific expertise.

3 Experiment 1

The purpose of Experiment 1 was to determine, in a simple situation containing only two domains, how well cross-domain weighting performed relative to within-domain weighting for the Top 5 and CWM algorithms. Obviously, if the two domains were highly related, it would be unsurprising if cross-domain weighting performed well relative to within-domain weighting. Therefore, to provide a more rigorous test of the utility of cross-domain weighting, we purposely choose two domains that were sufficiently dissimilar that it was likely that domain-specific expertise in one domain would be of little benefit in the other domain: American NFL trivia and general-knowledge Science trivia. From the above model, we predicted that cross-domain weighting would be generally less effective than within-domain weighting since the correlation of performance across domains will not be perfect. 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 was hard to compare the relative predictive power of

these different forms of information. We were therefore agnostic as to whether cross-domain weighting would be superior to single-question aggregation approaches.

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 judgements 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 estimates 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 judgments in response to question (1) that were consistent with the probability estimates they provided in response to question (2).

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.

Following the lead of Budescu and Chen 2015, we assessed performance using a transformed Brier score. We did this because we find the transformed Brier score to be more intuitive than the standard Brier score because the transformed Brier score ranges from 0 to

100 and *higher* values represent higher performance. Conversely, the standard Brier score ranges from 0 to 1 where, counter-intuitively, *lower* values represent higher performance. Like the original Brier score, the transformed Brier score is strictly proper. The transformed Brier score for the jth participant is given by:

$$S_j := 100 - 100 \sum_{k=1}^{K_j} \frac{(O_k - P_k)^2}{K_j}, \tag{22}$$

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 judgments by participant j (i.e., $k \in \{1, \ldots, K_j\}$), and P_k is the probability assigned to the outcome being 'true' by the jth participant. 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 prediction over all events and 75 being the score that is generated from an uninformed predictions of $P_k = 0.5$ in all questions. As such, a transformed Brier score of 75 or less indicates that the person or algorithm lacks any meaningful expertise in that domain.

3.2 Aggregation Algorithms

The predictions of each of the five aggregation algorithms (two of which came in both a within-domain and cross-domain version) were obtained as follows:

1. For the simple average, the probability estimate for each event was calculated by taking an unweighted average of the probability estimate of all individuals for that event.

2. For the Top 5 algorithm, the top five individuals were identified based on their performance on past questions and the probability estimate for each event was calculated by taking an unweighted average of their probability estimates for that event. For within-domain weighting, we measured each individual's past performance on all questions within that domain, except for the question for which we were seeking a probability estimate (i.e. the 'test' question). 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 past performance of each individual. The top five individuals were then identified and their probability estimates averaged to provide a probability estimate for the test question. This process was repeated for each question within the NFL domain, meaning that, potentially, for each question, a different set of five individuals was used. For cross-domain weighting, we measured each individual's past

performance on all questions that were not in the domain of interest. Using this data, the top five individuals were identified and their average probability estimate was used as the probability estimate for each test question, within the domain of interest. This meant that the same set of five individuals was used for all questions within a given test domain.

3. For the CWM algorithm, we use the contribution metric proposed by Budescu & Chen (2015) to provide a measure of each individual's expertise relative to other individuals in the crowd, as this contribution metric has been shown to be more effective than metrics based on individual measurements of past performance, such as mean Brier scores, at identifying and extracting individual expertise (Budescu & Chen, 2015; Chen et al., 2016). Formally, the contribution of the jth individual is equal to the difference between the transformed Brier score of the crowd's average probability estimate, $\overline{S}$, and the transformed Brier score of the crowd's average probability estimate without that individual, $\overline{S}_{\neg j}$, averaged over all N_j probability estimates made by that individual in the training set:

$$C_j := \sum_{i=1}^{N_j} \frac{\overline{S} - \overline{S}_{\neg j}}{N_j} \tag{23}$$

   As before, each individual's performance was either assessed within the domain of interest (for the purposes of within-domain weighting) or on questions outside the domain of interest (for the purposes of cross-domain weighting).

4. For the MPW algorithm, a weight was constructed for each individual using that individual's probabilistic estimate that the event was true, $P_{i,k}$, and their meta-prediction of the average probability estimated by other people, $M^P_{i,k}$. Let $j = \{1, \ldots, N_k\}$ be the set of individuals who answered problem k. An individual's 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 an individual's i's probability estimate and their meta-prediction and the denominator is the sum of this absolute difference over all N_k individuals. By construction, the weights assigned to the individuals add up to 1. The probability estimate generated by the MPW algorithm for

event k was the weighted average of each individual's probabilistic estimate:

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

5. For the MP algorithm, the probability estimate for each event was the average probability estimate of the individuals plus a correction term that was equal to the difference between the average probability estimate and the average meta-prediction (Palley & Soll, 2019). For example, if the average probability estimate 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 probability estimate 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 probability predictions were always bounded between 0 and 1.

3.3 Statistics

We use the transformed Brier score to assess the performance of each aggregation approach. Our main comparisons of interest are between the two variants of cross-domain weighting (cross-domain Top 5 and cross-domain CWM) and five other aggregation approaches: the simple average, the MPW algorithm, the Minimal Pivoting algorithm, within-domain Top 5, and within-domain CWM.

For completeness and ease of readability, we use one-sided paired-sample Bayesian t-tests to compare the mean score of each approach to the other six approaches calculated. These t-tests are calculated (i) across all three experiments, (ii) for each experiment separately, and (iii) on each domain of each experiment separately. We provide all paired tests in Appendix A and discuss the comparisons that relate to the provided figures in the main text.

For each statistical comparison in this paper, we report a Bayes factor (BF_{+0}) calculated using a one-sided 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 approach A does not produce more accurate probability estimates than approach B) or the alternative hypothesis (i.e. that approach A produces more accurate probability estimates than approach B) is better supported by the data. We interpret these Bayes factors in accordance with the recommendations of Kass & Raftery (1995), and use a Bayes factor of 40 as a cutoff when discussing whether there is strong support for the H1 hypothesis that approach A produces a more accurate

probability estimate than approach B.

3.4 Results and Discussion

Figure 1 shows the mean performance of each algorithm separately for the NFL Trivia and Science Trivia domains. We consider first the NFL domain. As discussed earlier, a transformed Brier score of 75 indicates a lack of expertise. We see that the simple average exceeds this threshold, but not by much, indicating that most participants lacked expertise. This was confirmed by Figure 2 which shows that only two participants had a transformed Brier score greater than 80. Consequently, it was not possible to identify multiple experts in this domain as almost no one in this domain had any significant expertise. Consequently, the within-domain Top 5 algorithm, which relies on identifying five experts, performed poorly, failing to outperform the simple average, as shown in Figure 1. In fact, if anything, it seems to have performed a little worse than the simple average. As discussed above, the advantage of the simple average is that by averaging over a large number of participants, it can reduce noise and achieve a better signal. As the Top 5 algorithm averages over only 5 participants, it largely forgoes this advantage. The Top 5 algorithm will therefore outperform the simple average only if it can overcome this disadvantage by identifying five individuals who are significantly better at the task than the others. For the NFL domain, it was not able to do this, even when using within-domain weighting.

So why was performance so poor in the NFL domain? If performance is determined, at least in part, by general expertise, shouldn't there be a reasonable level of performance in the NFL domain, even if most of the participants lacked any domain-specific knowledge? Or do these results indicate that all participants had little general expertise? The answer to this question is informed by Equation 1 which shows that the degree to which general expertise affects performance in domain i is determined by the constant γ_i. The fact that performance was poor in the NFL domain indicates that both e_i and $\gamma_i g$ were likely small. Since γ_i and g are independent, the fact that performance was poor in the NFL domain does not necessarily mean that general expertise, g, was small for all participants.

Turning our attention to the Science domain, we see that the simple average performs better than it did in the NFL domain. Furthermore, we see that the within-domain Top 5 algorithm is more predictive than the simple average ($BF_{+0} > 1000$), showing that in the Science domain it is possible to identify some participants who perform considerably better than others. As expected, the within-domain Top 5 algorithm performed better than the cross-domain Top 5 algorithm, though the statistical support for this relationship is weak ($BF_{+0} = 5.50$). Looking at Figure 2 we can immediately see why cross-domain weighting has

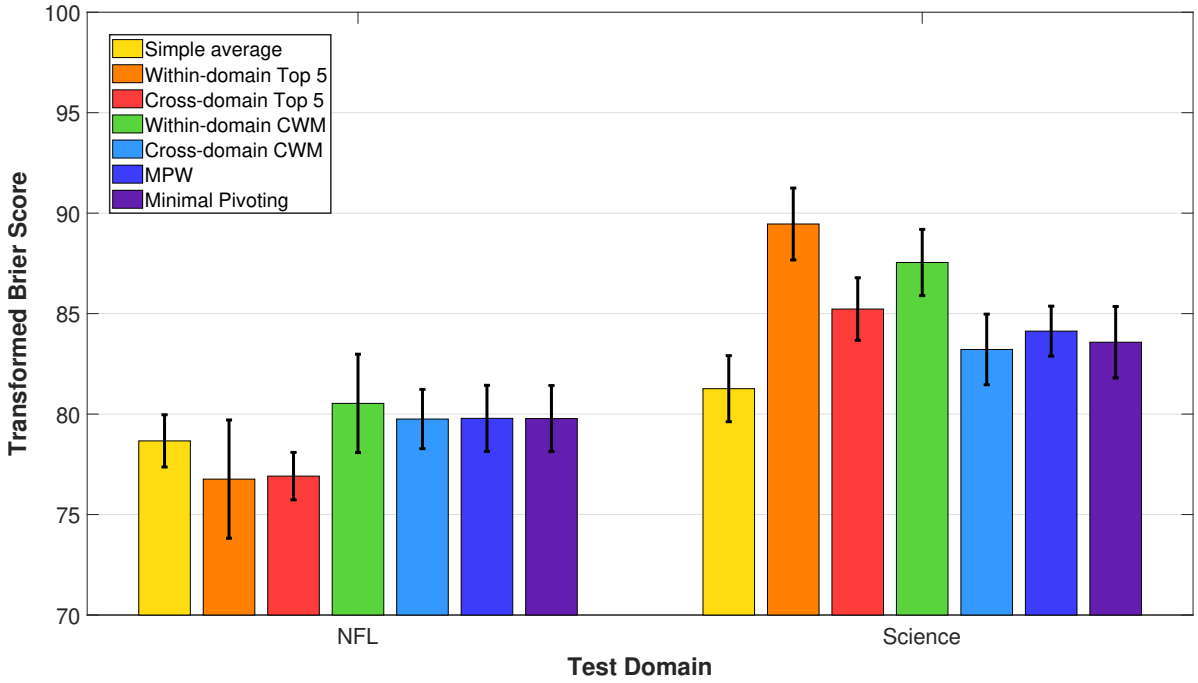

Figure 1: Results from Experiment 1 showing the mean score for each algorithm on the NFL Trivia (left) and Science Trivia domain (right). Error bars show the standard error. Within each domain, the algorithms are ordered from left to right, from the Simple average to Minimal Pivoting algorithm, in the same order as given in the legend.

slightly reduced performance. The top performer in the NFL domain performed particularly poorly in the Science domain. Indeed, taken together, the top five performers in the NFL domain, did not perform particularly well in the Science domain, with the result that when the Top 5 algorithm used performance in the NFL domain to select the Top 5 performing individuals for the Science domain, it did not perform particularly well.

The Top 5 algorithm works by identifying the top five performers and averaging their probability estimates. The CWM algorithm also weights the higher performers more. However, unlike the Top 5 algorithm, it averages together more than just the Top 5 performers. As such, its performance typically lies somewhere between the performance of the Top 5 algorithm and the simple average, which is what we see in the Science domain for both the within-domain and cross-domain weighting.

Turning our attention now to the single-question algorithms, we see that they perform less well than the within-domain CWM and Top 5 algorithms, but at about the same level as the cross-domain CWM and Top 5 algorithms. Statistical tests for these comparisons are available in Appendix A.

Returning our attention to Figure 2, we are struck by the number of individuals who

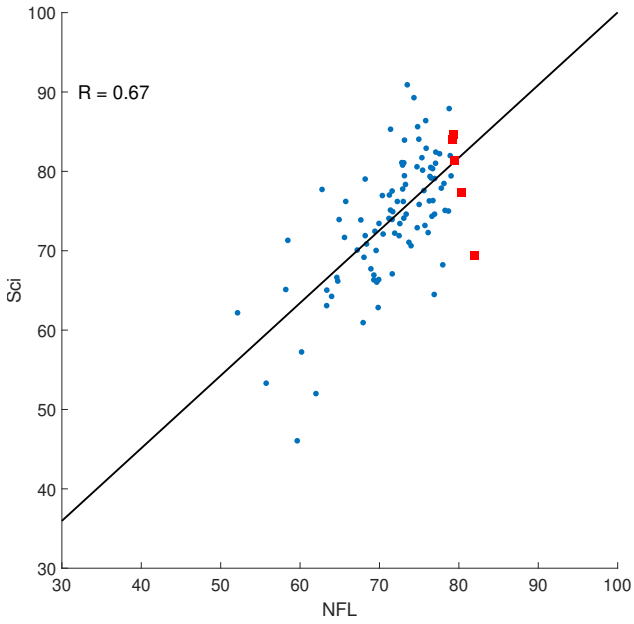

Figure 2: The correlation in performance, as measured by the transformed Brier score, in the NFL domain and in the Science domain in Experiment 1. Each circle/square represents a single participant. The filled, red squares signify the participants selected by the Top 5 algorithm when making predictions in the Science domain based on the performance of participants in the NFL domain

scored less than 75. If an individual give a probability estimate of 50% for every question (i.e. displays no confidence as to what the outcome is), they would score 75. To score less than 75 they need to report probabilities closer to 0% or 100% (i.e. display some confidence about what the outcome is) but often be wrong. For example, if an individual were to randomly respond with probabilities of either 0% or 100% and be incorrect half the time, they would score 50. (To score less than 50 they would need to be incorrect more than half the time!) From Figure 2, we see that a large number of our participants scored in the range 50-75 indicating that they were overconfident in that they reported probabilities that deviated from 50% but then often got the answer wrong.

4 Experiment 2

So far, we have not found any strong evidence that cross-domain weighting is effective. We believe that the main reason for this was that we could not identify expertise in the NFL domain, as evidenced by the poor performance of the within-domain Top 5 algorithm. As the cross-domain algorithm is unlikely to perform better than the within-domain algorithm, it was not surprising that the cross-domain Top 5 algorithm performed poorly in the NFL

domain as well. It also meant that it was not possible to use expertise in the NFL domain to identify expertise in the Science domain, as evidenced by the cross-domain Top 5 algorithm performing poorly relative to the within-domain Top 5 algorithm in the Science domain.

The purpose of Experiment 2 was to determine if, in different circumstances, cross-domain weighting would be effective. Specifically, we tested whether, if we were to use different domains, cross-domain weighting would perform better. From our model, we would expect that the performance of cross-domain weighting would improve if we increased the number of cross-domains, providing performance in each cross-domain reflected general expertise. Therefore, in addition to the Science domain, we selected two other domains where we believed this condition would likely hold.

As discussed earlier, if these two additional domains were highly related both to each other and to the Science domain, it would be unsurprising if cross-domain weighting performed well relative to within-domain weighting. Therefore, to provide a rigorous test of the utility of cross-domain weighting, we purposely choose these two additional domains to not be obviously related to each other or to the Science domain.

The two domains we chose were Art Evaluation (Art) and Emotional Intelligence (EI). For the Art domain, participants were shown a series of photographs of artworks and needed to evaluate whether each one was worth more or less than US$1000. For the EI domain, expertise was measured using questions adapted from the Situational Test of Emotional Understanding and Situational Test of Emotion Management (MacCann & Roberts, 2008). We chose Art and EI questions because they are likely to tap into a different set of skills and knowledge compared both to each other and to Science Trivia (e.g., Cattell, 1963; Mayer et al., 2001; Lam & Kirby, 2002).

In the previous experiment, there were 50 questions in each domain, resulting in 100 questions in total. Now that we had three domains, we were concerned that we might overwhelm people with the number of questions that we were asking. We therefore reduced the number of questions in each domain to 40, for a total of 120 questions.

4.1 Methods

We adopted the same methodology as the previous experiment but 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 1) whether the market price of the original version of that artwork would exceed US$10,000, (2) the probability that this statement was "true" (from 0–100), and (3) what they thought the average probability estimated by other peo-

ple on question (2) would be (from 0-100). Participants were asked to provide votes and probability estimates that were consistent (i.e., providing a probability estimate greater than 0.5 when predicting 'True' and a probability less than 0.5 when predicting 'False'), and we subsequently excluded any estimates that were inconsistent from our analyses.

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 US$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 US$4.00 for completing the experiment. Participants from any of our previous experiments were prevented from participating in this experiment. We collected responses for 100 participants, and we then excluded participants who reported cheating during the experiment as we did for the previous experiments. Twelve participants were excluded and analyses were conducted on the remaining 88 people.

4.2 Results and Discussion

Figure 3 shows the mean performance of each algorithm separately in each of the three domains. Collapsing across all domains, the within-domain Top 5 algorithm performed better than the simple average indicating that expertise could be reliably identified using within-domain data. Gratifyingly, the cross-domain Top 5 algorithm also performed well, indicating that expertise in each domain could be identified using data from the other two domains. The reason why this was possible is shown in Figure 4. This plot shows performance in each test domain (y-axis) as a function of the combined performance on the other two domains (x-axis). We see that participants who had the best combined performance on the other two domains always did well in the test domain.

In all three domains, the within-domain CWM algorithm performed similarly to the within-domain Top 5 algorithm. As with the other two experiments, the two single-question algorithms did not perform much better than the simple average.

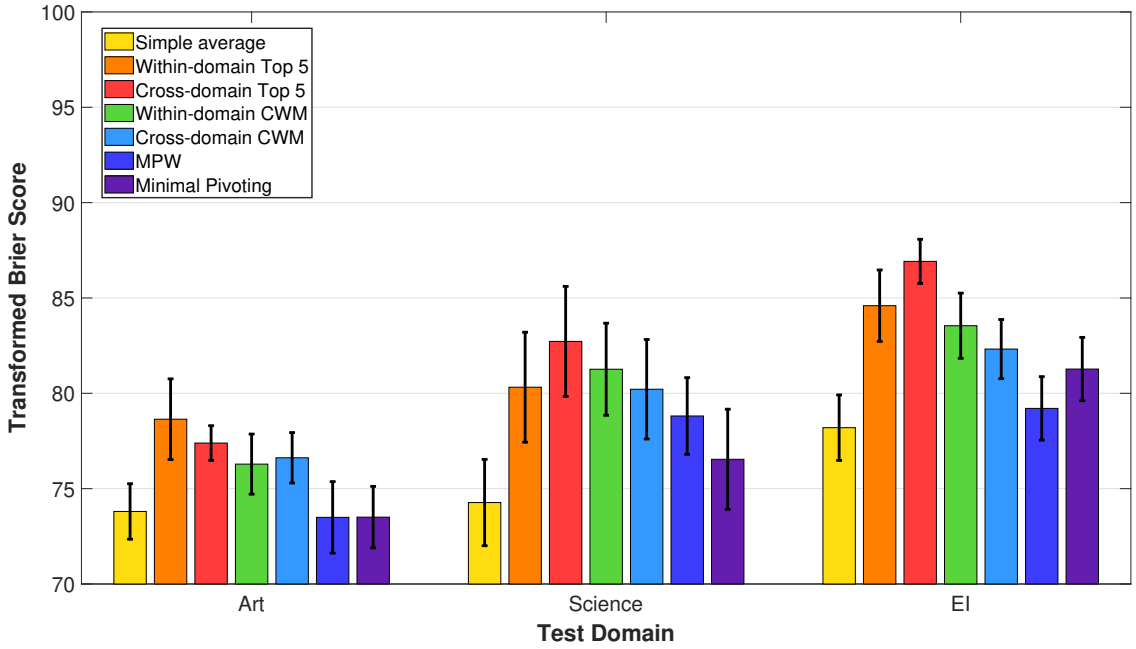

Figure 3: The mean score for each model in Experiment 2 on the Art questions (left), Science Trivia questions (centre), and Emotional Intelligence questions (right). Error bars show the standard error. Within each domain, the algorithms are ordered from left to right, from the Simple average to Minimal Pivoting algorithm, in the same order as given in the legend.

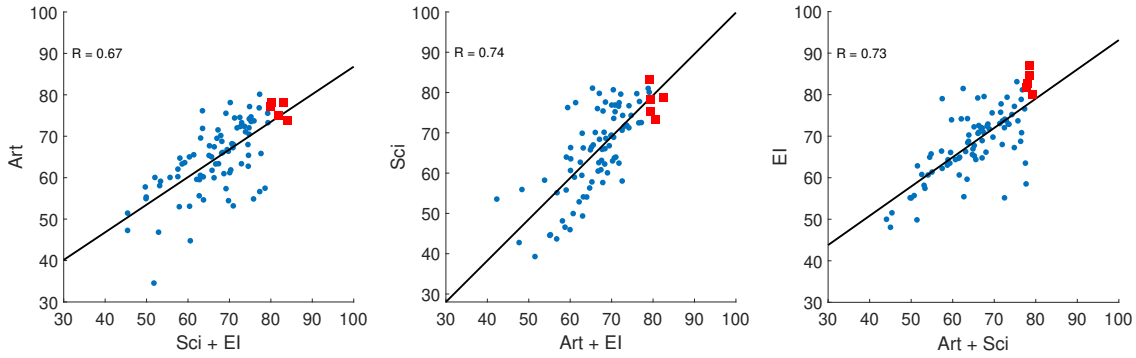

Figure 4: Each subplot shows, for Experiment 2, the correlation in performance, as measured by the transformed Brier score, between the test domain (y-axis) and the performance averaged over the other two domains (x-axis). The test domains were Art, Science and EI for the three subplots respectively. As before, each circle/square represents a single participant. The filled, red squares denote the five participants selected by the cross-domain Top 5 algorithm when it was making predictions in the subplot's test domain.

5 Experiment 3

The previous experiment presented the first convincing evidence that cross-domain weighting can be effective. Collapsing across all domains, the cross-domain Top 5 algorithm significantly outperformed both the simple average and the two single-question algorithms. We argued that cross-domain weighting worked in Experiment 2 because in each of the three domains in that experiment expertise could be reliably identified, at least using within-domain data. But were these results a fluke? The purpose of Experiment 3 was to repeat Experiment 2 to see if findings were repeatable.

5.1 Methods

The methods of this experiment were identical to those of the previous experiment, except for one difference. In the previous experiment, participants were asked (but not required) to provide votes and probability estimates that were consistent (i.e., provide a probability estimate greater than 0.5 when predicting 'True' and a probability less than 0.5 when predicting 'False'), and we subsequently excluded any probability estimates that were inconsistent from our analyses. The problem with this approach is that we might have been excluding the less diligent participants, so might be artificially inflating the effectiveness of cross-domain weighting. To avoid this potential problem, in the current experiment, participants were forced to provide votes and probability estimates that were consistent in order to proceed. This meant that no participants needed to be excluded. As before, participants were asked if they cheated and looked up any of the answers. Twenty-one participants were excluded for this reason and the analyses were conducted on the remaining 79 people.

5.2 Results and Discussion

The results are shown in Figure 5. They are very similar to the results in Experiment 2. Within each domain, expertise could be readily identified as indicate by the high performance of the within-domain Top 5 algorithm. Similarly, cross-domain weighting worked well, in all three domains. The reason for this is shown in Figure 6. As before, the filled, red squares show the five participants used by the Top 5 algorithm in that test domain. In each test domain (y-axis), the combined performance in the other two domains (x-axis) reliably predicted performance in that test domain. This is why cross-domain weighting was effective in this experiment. Taken together, these results show that the findings of Experiment 2 are repeatable and cross-domain weighting can be effective.

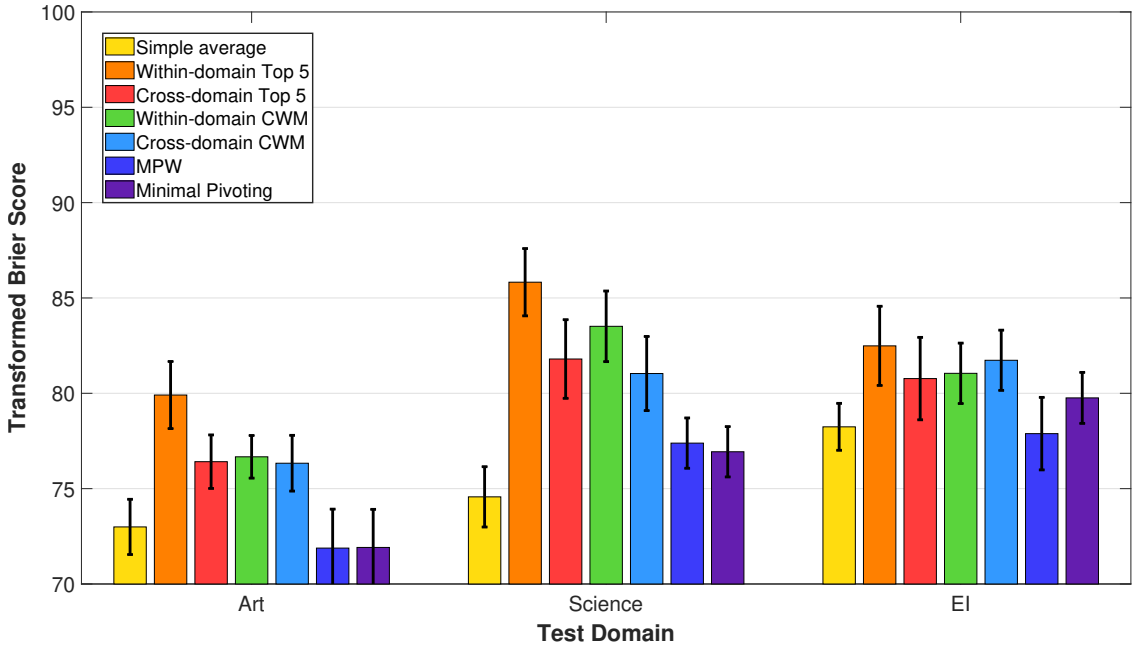

Figure 5: The mean score for each model in Experiment 3 on the Art questions (left), Science Trivia questions (centre), and Emotional Intelligence questions (right). Error bars show the standard error. Within each domain, the algorithms are ordered from left to right, from the Simple average to Minimal Pivoting algorithm, in the same order as given in the legend.

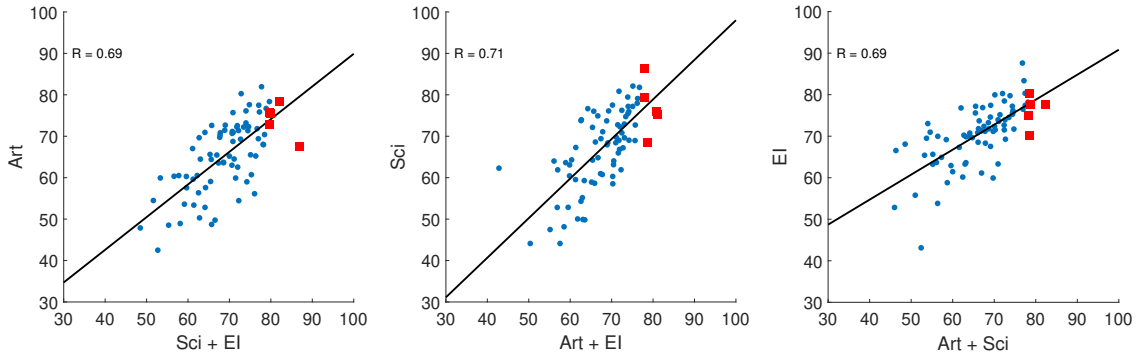

Figure 6: Each subplot shows, for Experiment 3, the correlation in performance, as measured by transformed Brier score, between each test domain (y-axis) and the performance averaged over the other two domains (x-axis). The test domains were Art, Science and EI for the three subplots respectively. As before, each circle/square represents a single participant. The filled, red squares denote the five participants selected by the cross-domain Top 5 algorithm when it was making predictions in the subplot's test domain.

6 Simulating changes in training set size and in the number of individuals

A potential limitation of our results thus far is that cross-domain weighting appears to require a larger set of training questions in order to obtain a similar level of performance as within-domain weighting. Here we address this concern by comparing the within-domain Top 5 algorithm to the cross-domain Top 5 algorithm when both algorithms are restricted to use the same number of training questions. We chose to study the Top 5 algorithm as opposed to the CWM algorithm, as the former is both simpler to understand and generally outperformed the later. Similar findings apply to the CWM algorithm although, for the reasons previously noted, its performance generally lay between that of the Top 5 algorithm and the simple average.

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 probability estimate), we divided probability estimates 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.

The pool of potential cross-domain questions is larger than the within-domain pool. To ensure that the simulation used within-domain and cross-domain pools of the same size, the following approach was used: on each bootstrap iteration, a subset of questions equal to the number of training questions in the within-domain pool was randomly selected without replacement from the cross-domain pool of training questions. Responses to the remaining questions in the cross-domain pool were removed for that iteration. 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 Experiment 1, and in the range of {10, 20, 30, 39} for Experiments 2 and 3 because the datasets for the latter two experiments were smaller.

On each of the 1,000 iterations, we calculated the performance of the within-domain Top 5 algorithm and the cross-domain Top 5 algorithm, 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 algorithm's mean score for that set size. As a reference, we also calculated the mean score of the simple average on the original sample in each dataset.

As can be seen in Figure 7, there was very little difference in score between cross-domain weighting (blue line) and within-domain weighting (red line) in all three experiments, regardless of the number of training questions used. Thus, while some of our original cross-domain estimates used more training questions to estimate each individual's contribution, our results here suggest that the benefits of this larger training sample were small.

We also investigated how our results would be altered by varying the number of participants. Figure 8 shows the results of these simulations. Returning to the original dataset for each experiment, we simulated the change in mean transformed Brier score for within-domain weighting and cross-domain weighting that would have occurred had we used a different number of participants. As before, we did this using the bootstrap (Efron & Tibshirani, 1994) to resample data from each of the three experiments. For each test question, a random set of k participants was sampled from the pool of participants. We repeated this process 1,000 times for each sample size in the range of {10, 20, 40, 60, 80} for data from Experiments 1, 2 and 3.

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 participants. We then averaged each model's performance across each test event in the dataset to obtain that model's mean score for that sample size. As a reference, we also calculated the mean score of the simple average on the original sample in each dataset.

The results show that the performance of the algorithms are largely invariant to changes in the sample sizes when at least 40 participants are used in each algorithm. However, for sample sizes of 20 or fewer participants, performance rapidly decreases as the sample size decreases. Intuitively, this makes sense. The Top 5 algorithm will perform well only if at least 5 experts exist in the sample. For large samples, it is likely that at least 5 experts will exist. But for samples less than 20, it may not be possible to identify 5 experts, with the result that the performance of the Top 5 algorithm suffers.

Overall, these simulation results suggest that expertise can be estimated efficiently using questions from other domains and cross-domain weighting also appears to be highly robust to different training set sizes providing a sufficient number of participants are used. These results are consistent with those of Chen et al. (2016), who demonstrated the robustness of the standard CWM across different training set sizes.

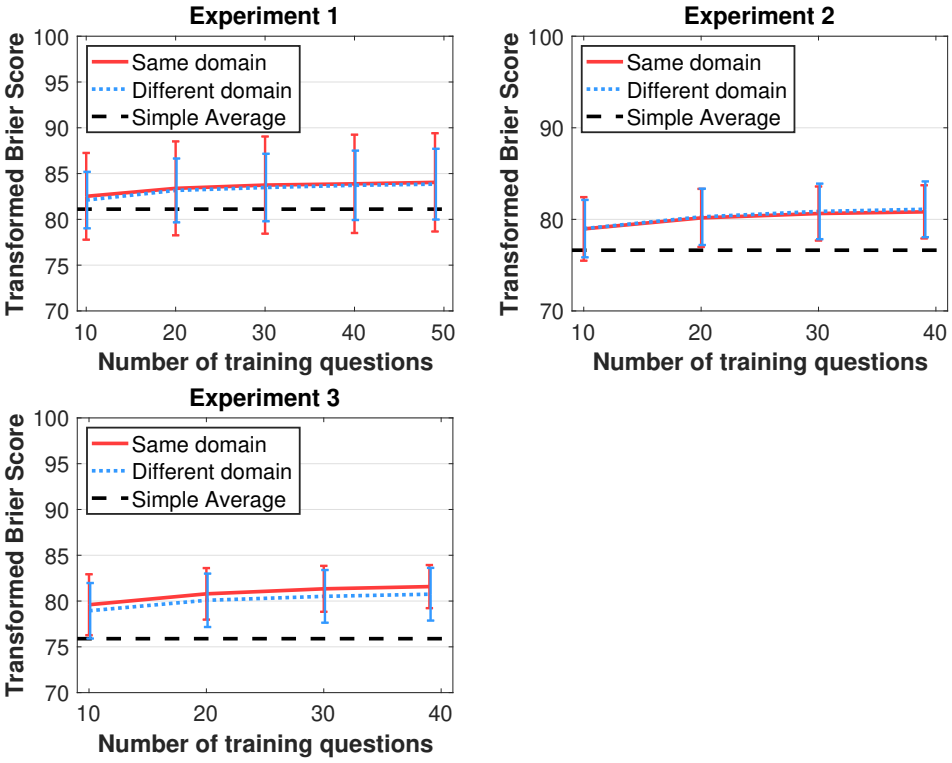

Figure 7: Simulations using data from Experiments 1 – 3 showing the mean transformed Brier score for within-domain weighting (red) compared to cross-domain weighting (blue) over different training set sizes. Error bars show the standard error. The performance of the simple average (dashed line), which does not use training data, is shown for reference.

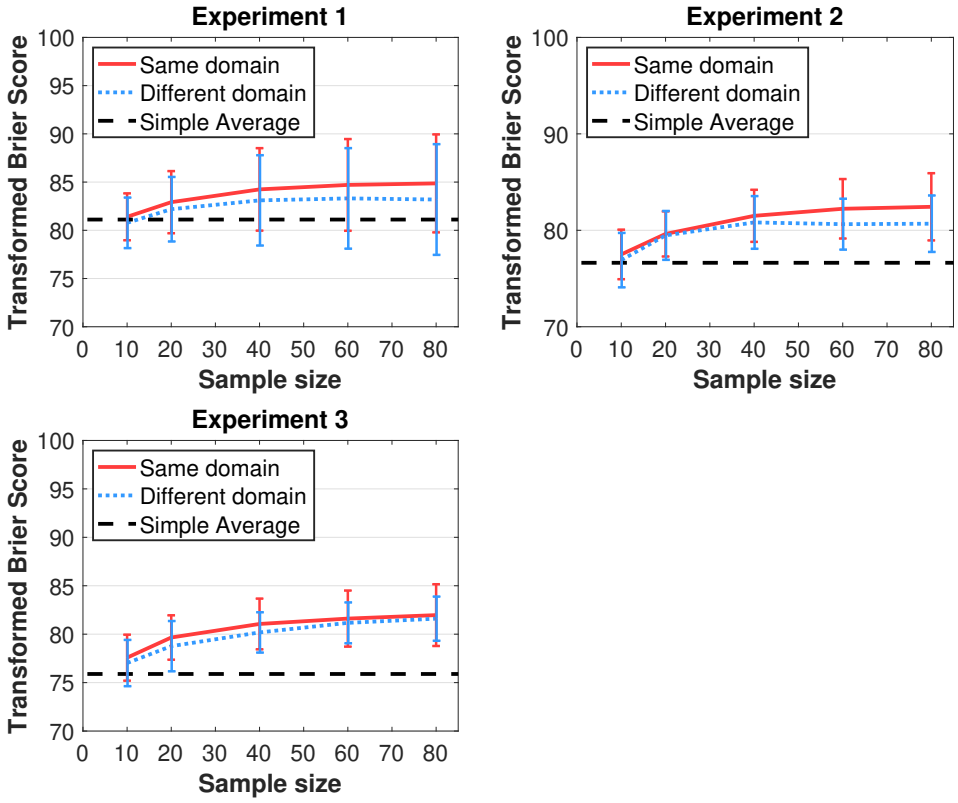

Figure 8: Simulations using data from Experiments 1 – 3 showing the mean transformed Brier score for within-domain weighting (red) compared to cross-domain weighting (blue) over different population sizes. Error bars show the standard error. The performance of the simple average (dashed line), which does not use training data, is shown for reference.

7 General Discussion

The aim of the current paper was to investigate whether an individual's expertise in a test domain could be accurately identified using their performance on decision problems in other domains. Over three experiments, we examined the performance of cross-domain weighting relative to within-domain weighting for both the Top 5 algorithm (Mannes et al., 2014) and the CWM algorithm (Budescu & Chen, 2015). We also measured the performance of two single-question algorithms: the MPW algorithm (Martinie et al., 2020) and the MP algorithm (Palley & Soll, 2019). We did this using several different question domains, including questions from NFL trivia, Science trivia, a test of Emotional Intelligence, and judgements about the price of professional and amateur artworks.

In each experiment, our initial focus was on comparing the Top 5 algorithm (Mannes et al., 2014) to the simple average. We found that for the NFL domain in Experiment 1 the per-

formance of the Top 5 algorithm did not exceed that of the simple average, even when using within-domain weighting. As discussed in the Introduction, because the noise associated with the probability estimates of different individuals tend not to be perfectly correlated, averaging over a large number of individuals tends to reduce noise and generate a more reliable signal, a phenomenon often referred to as the 'Wisdom of Crowds' (Surowiecki, 2005). Because the Top 5 algorithm averages over only five individuals, it largely forgoes this advantage, especially if the errors of those five individuals happen to be partially correlated. It therefore only betters the simple average if it can identify five individuals who are considerably better than the average and, better yet, have largely uncorrelated errors. It failed to do so in NFL domain in Experiment 1.

As discussed in Section 2 where we derived our theoretical model, we would not expect cross-domain weighting to perform well unless within-domain weighting performs well since if one cannot identify expertise in a test domain using within-domain performance data, it is unlikely that one can do better using out-of-domain performance data, which is necessarily less predictive of performance in the test domain. It was therefore no surprise that, in the NFL domain in Experiment 1, the cross-domain Top 5 algorithm performed poorly.

For the Science domain in Experiment 1, the within-domain Top 5 algorithm performed better than the simple average, demonstrating that it was able to identify high-performing individuals. However, the cross-domain Top 5 algorithm did not outperform the simple average because it was not possible to identify high performers in the Science domain using performance data from the NFL domain. As shown in Figure 2, individuals who had the best performance in the NFL domain had, at best, mediocre performance in the Science domain.

In Experiments 2 and 3, the situation was very different. In each experiment, in all three domains, the within-domain Top 5 algorithm performed well, showing that high-performing individuals could be reliably identified in all three domains. Furthermore, for each test domain, individuals who had the best combined performance across the other two domains, also performed well in that test domain. Thus, for each test domain, high performers could be identified using data from the other two domains. This allowed the cross-domain Top 5 algorithm to perform well.

Taken as a whole, these results suggest that, in practice, cross-domain weighting is likely to perform well only when high performing individuals can be identified not only in the test domain but also in each of the cross-domains. If one cannot reliably identify high performing individuals in the cross-domains, then it is unlikely that one will be able to use the performance in the cross-domains to identify high performing individuals in the test domain.

So far we have considered only the Top 5 algorithm (Mannes et al., 2014). The other

algorithm that utilised both within-domain and cross-domain weights was the CWM algorithm Budescu & Chen (2015). While this algorithm did weight the probability estimates of higher performing participants more, it considered more than just the top five participants. As such, its performance was typically between that of the Top 5 algorithm and the simple average for both the within-domain condition and the cross-domain condition. As such, our conclusions regarding when the Top 5 algorithm would and would not perform well relative to the simple average apply equally to the CWM algorithm. In particular, like the Top 5 algorithm, the cross-domain CWM algorithm performed well relative to the simple average only in Experiments 2 and 3, for the reasons discussed above.

A potential criticism of our analysis of cross-domain weighting is that in Experiments 2 and 3, the cross-domain algorithms had double the number of questions in the training set as the within-domain algorithms. As such, the cross-domain algorithms were potentially given an unfair advantage. In Section 6, we performed a series of simulations to investigate this issue. We found that, even when the same number of questions was used for both within-domain weighting and cross-domain weighting, our previous conclusions held.

In Section 6 we also investigated how the performance of the Top 5 algorithm would vary as a function of the number of participants. We found that if there were over 40 participants, increasing the number of participants further had little effect on the performance of the algorithm. Conversely, if there were 20 or fewer participants, decreasing the number of participants reduced the performance of the algorithm. Intuitively, this makes sense. For the Top 5 algorithm to perform well, there needs to be at least 5 experts in the participant pool. This is likely to occur if there are 40 participants or more but is increasingly less likely to occur if there are fewer than 20 participants.

In this paper we attempted to make the domains as dissimilar as possible. This was to ensure that we provided a proper test of cross-domain weighting. However, when using this technique practically it is likely better to select cross domains as similar to the test domain as possible. This way the algorithm can take advantage of both shared general expertise and any shared domain-specific expertise. As such, cross-domain weighting would be expected to work better than it worked in the experiments reported here.

One potential limitation of the study was that we could not prevent participants looking up the answers to the quiz questions. We asked participants not to do this and did not reward them based on their performance in the quiz, so as to avoid encouraging them to cheat. We also tried, where possible, to construct questions where it would be relatively hard to look up the answers. That said, some participants still reported cheating, so were excluded from our analysis. As it was made clear to participants that they would not be penalised for admitting to cheating, there is no reason to believe that any participants who cheated did

not acknowledge this.

In our experiments, we also measured the effectiveness of the two single-question algorithms: the MPW algorithm (Martinie et al., 2020) and the MP algorithm (Palley & Soll, 2019). In general, the improvement of these algorithms upon the simple average was, at most, small indicating that it is preferable to use algorithms that weight based on past performance. Of course, obtaining an estimate of past performance might be more difficult to do in some circumstances than in others. In particular, to estimate performance one needs to know the ground truth, which might not be known. However, we would argue that one can usually find related situations that approximate the situation of interest and where the ground truth is known. For example, in radiology it can be hard to know the ground truth with regards to mammograms because those that are deemed not at risk from cancer are not biopsied, so false negatives can be missed for long periods of time. However, in other tasks, such as diagnosing X-rays for potential neck-of-femur hip fractures false negatives are likely to be corrected much faster. In particular, for a person with a hip fracture that is wrongly diagnosed as not having one, the complications resulting from this misdiagnosis will likely become apparent within a short period of time, allowing the mistake to be corrected. Thus, one could use performance on diagnosing X-rays for possible neck-of-femur fractures as a proxy for performance on diagnosing mammograms given that performance can be accurately ascertained on the former.

Ideally, one would want to test people on more than one cross-domain. As our theoretical work shows that, providing performance in these cross domains sufficiently reflects general expertise, the correlation between performance in the test domain and performance averaged across all the cross domains increases as the number of cross-domains increases. Presumably the correlation would be higher still if the cross-domains are related to the domain of interest.

Our results provide insight into the generality of expertise across a range of decision-making domains. Previous applications of expertise-identification approaches such as the Top 5 algorithm (Mannes et al., 2014) and the CWM algorithm (Budescu & Chen, 2015) have been largely limited to estimating individuals' expertise by their performance within similar or identical domains to the questions of interest (Mellers et al., 2015; Cooke, 1991). Our results also show that cross-domain weighting may be more effective than existing single-question aggregation approaches. While previous research has shown that single-question aggregation approaches can be useful for leveraging expertise when previous performance in the same domain is unknown (Martinie et al., 2020), the current results suggest that it is usually better to estimate expertise in the test domain using measures of expertise in other domains. Ideally, these domains would be related to the domain of interest, but here we

showed that even seemingly unrelated domains can be effective, providing expertise can be reliably identified in them. The cross-domain weighting approach is therefore an attractive and effective alternative for decision makers seeking to improve judgments on novel problems for which there are no records of previous performance. We hope that our results will inspire future researchers to examine the accuracy 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.

Friedman, J. A., Baker, J. D., Meller, B. A., Tetlock, P. E., & Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly*, *62*, 410–422.

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.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The Wisdom of Select Crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299.

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). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, *21*(1), 1–14.

Mellers, B. A., Baker, J. D., Chen, E., Mandel, D. R., & Tetlock, P. E. (2017). How generalizable is good judgment? A multi-task, multi-benchmar study. *Judgment and Decision Making*, *12*(4), 369–381.

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 revising 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.

Tetlock, P. E. & Gardner, E. (2015). *Superforecasting: The Art and Science of Prediction*. Crown Publishers.

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.

Appendix A: Bayes-Factor Tables

This appendix provides a complete set of one-sided paired-sample Bayesian t-tests that compare the mean score of each approach to the other six approaches calculated. These t-tests are calculated (i) across all three experiments, (ii) for each experiment separately, and (iii) on each domain of each experiment separately.

We report a Bayes factor (BF_{+0}) calculated using a one-sided 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 approach A does not produce higher Brier scores than approach B) or the alternative hypothesis (i.e. that approach A produces higher Brier scores than approach B) is better supported by the data. We interpret these Bayes factors in accordance with the recommendations of Kass & Raftery (1995), summarised in Table A1. Note that these are the BF_{10} cutoffs and we are reporting the one-sided tests, BF_{+0}. As such, we use a BF of 40 as a threshold for strong support. For readability, we have bolded cells where the Bayes factors is above 40 in each table.

Table A1: Interpretations for Bayes Factors (BF_{10})

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

Experiment 1: All Domains									
	Simple Average	Within-Domain Top-5	Cross-Domain Top-5	Within-Domain CWM	Cross-Domain CWM	MPW	Minimal Pivoting	Mean Brier Score	N
Simple Average	x	0.03	0.05	0.02	0.03	0.03	0.02	79.97	100
Within-Domain Top-5	2.86	x	0.43	0.05	0.43	0.31	0.41	83.11	100
Cross-Domain Top-5	0.40	0.05	x	0.03	0.08	0.06	0.07	81.07	100
Within-Domain CWM	**1193.06**	0.39	2.86	x	**40.21**	27.84	**87.16**	84.04	100
Cross-Domain CWM	12.25	0.05	0.18	0.03	x	0.06	0.08	81.49	100
MPW	10.67	0.06	0.27	0.03	0.25	x	0.19	81.96	100
Minimal Pivoting	**1319.03**	0.05	0.20	0.03	0.17	0.07	x	81.68	100

Experiment 2: All Domains									
	Simple Average	Within-Domain Top-5	Cross-Domain Top-5	Within-Domain CWM	Cross-Domain CWM	MPW	Minimal Pivoting	Mean Brier Score	N
Simple Average	x	0.02	0.01	0.02	0.02	0.05	0.03	75.42	120
Within-Domain Top-5	**3633.77**	x	0.05	0.30	0.61	19.53	**112.87**	81.19	120
Cross-Domain Top-5	**4.51E+06**	0.36	x	4.46	**130.77**	**1.42E+04**	**1.54E+05**	82.34	120
Within-Domain CWM	**5361.47**	0.05	0.03	x	0.76	**314.10**	**4846.56**	80.36	120
Cross-Domain CWM	**2095.16**	0.04	0.02	0.04	x	24.52	**863.70**	79.72	120
MPW	0.43	0.03	0.01	0.02	0.03	x	0.11	77.17	120
Minimal Pivoting	4.19	0.02	0.01	0.02	0.02	0.10	x	77.11	120

Experiment 3: All Domains									
	Simple Average	Within-Domain Top-5	Cross-Domain Top-5	Within-Domain CWM	Cross-Domain CWM	MPW	Minimal Pivoting	Mean Brier Score	N
Simple Average	x	0.01	0.02	0.01	0.01	0.08	0.05	75.27	120
Within-Domain Top-5	**2.13E+08**	x	**133.36**	**215.59**	**456.64**	**6.71E+04**	**2.64E+06**	82.74	120
Cross-Domain Top-5	**332.01**	0.02	x	0.05	0.10	16.05	23.06	79.66	120
Within-Domain CWM	**2.90E+06**	0.02	0.28	x	0.64	**3623.10**	**6.92E+05**	80.41	120
Cross-Domain CWM	**3.33E+05**	0.02	0.11	0.04	x	**163.76**	**3.94E+04**	79.70	120
MPW	0.14	0.01	0.03	0.02	0.02	x	0.06	75.72	120
Minimal Pivoting	0.43	0.01	0.03	0.01	0.01	0.20	x	76.20	120

Table A2: Bayes-Factors from one-sided paired-samples Bayesian t-test of the row and column method. The H1 hypothesis is that the row method has a strictly greater Brier score than the column method. Bayes-Factors above 40 are interpreted as strong support that the row method better predicts the outcomes than the column method. The mean Brier Score for the row method is shown at the end of each row.

Combined Data From Experiments 1-3: All Domains									
	Simple Average	Within-Domain Top-5	Cross-Domain Top-5	Within-Domain CWM	Cross-Domain CWM	MPW	Minimal Pivoting	Mean Brier Score	N
Simple Average	x	0.01	0.01	0.01	0.01	0.02	0.01	76.71	340
Within-Domain Top-5	**4.94E+11**	x	0.75	0.86	**79.06**	**7.47E+05**	**3.46E+07**	82.30	340
Cross-Domain Top-5	**7.66E+08**	0.02	x	0.04	0.68	**1708.00**	**9753.57**	81.02	340
Within-Domain CWM	**1.35E+14**	0.02	0.13	x	**270.11**	**4.42E+08**	**5.19E+12**	81.46	340
Cross-Domain CWM	**4.86E+10**	0.01	0.02	0.01	x	**1004.62**	**3.81E+06**	80.23	340
MPW	0.88	0.01	0.01	0.01	0.01	x	0.05	78.07	340
Minimal Pivoting	**245.12**	0.01	0.01	0.01	0.01	0.07	x	78.13	340

Experiment 1: NFL Domain									
	Simple Average	Within-Domain Top-5	Cross-Domain Top-5	Within-Domain CWM	Cross-Domain CWM	MPW	Minimal Pivoting	Mean Brier Score	N
Simple Average	x	0.39	0.77	0.07	0.07	0.08	0.06	78.67	50
Within-Domain Top-5	0.09	x	0.15	0.04	0.06	0.06	0.06	76.76	50
Cross-Domain Top-5	0.07	0.16	x	0.06	0.05	0.06	0.06	76.91	50
Within-Domain CWM	0.56	**121.38**	0.92	x	0.27	0.31	0.31	80.53	50
Cross-Domain CWM	0.63	0.97	4.26	0.10	x	0.15	0.15	79.75	50
MPW	0.54	1.64	1.38	0.09	0.16	x	0.16	79.79	50
Minimal Pivoting	1.45	1.62	1.69	0.09	0.16	0.15	x	79.78	50

Experiment 1: Science Domain									
	Simple Average	Within-Domain Top-5	Cross-Domain Top-5	Within-Domain CWM	Cross-Domain CWM	MPW	Minimal Pivoting	Mean Brier Score	N
Simple Average	x	0.02	0.04	0.01	0.04	0.04	0.03	81.26	50
Within-Domain Top-5	**1.62E+04**	x	5.50	1.42	**264.50**	**111.29**	**183.42**	89.46	50
Cross-Domain Top-5	24.02	0.05	x	0.05	1.91	0.40	0.66	85.23	50
Within-Domain CWM	**3.30E+05**	0.06	1.94	x	**1.50E+04**	**602.54**	**7090.42**	87.55	50
Cross-Domain CWM	18.45	0.03	0.06	0.02	x	0.08	0.09	83.22	50
MPW	9.92	0.04	0.08	0.03	0.42	x	0.28	84.13	50
Minimal Pivoting	**2470.70**	0.04	0.07	0.02	0.32	0.10	x	83.58	50

Table A3: Bayes-Factors from one-sided paired-samples Bayesian t-test of the row and column method. The H1 hypothesis is that the row method has a strictly greater Brier score than the column method. Bayes-Factors above 40 are interpreted as strong support that the row method better predicts the outcomes than the column method. The mean Brier Score for the row method is shown at the end of each row.

Experiment 2: Art Domain									
	Simple Average	Within-Domain Top-5	Cross-Domain Top-5	Within-Domain CWM	Cross-Domain CWM	MPW	Minimal Pivoting	Mean Brier Score	N
Simple Average	x	0.06	0.05	0.08	0.07	0.19	0.21	73.80	40
Within-Domain Top-5	3.83	x	0.33	0.84	0.50	1.51	3.49	78.64	40
Cross-Domain Top-5	14.27	0.11	x	0.33	0.29	1.73	7.06	77.39	40
Within-Domain CWM	0.59	0.08	0.11	x	0.12	2.16	1.79	76.29	40
Cross-Domain CWM	1.05	0.09	0.11	0.26	x	10.00	12.39	76.62	40
MPW	0.16	0.07	0.06	0.06	0.05	x	0.17	73.50	40
Minimal Pivoting	0.14	0.06	0.05	0.06	0.05	0.17	x	73.51	40

Experiment 2: Emotional Intelligence Domain									
	Simple Average	Within-Domain Top-5	Cross-Domain Top-5	Within-Domain CWM	Cross-Domain CWM	MPW	Minimal Pivoting	Mean Brier Score	N
Simple Average	x	0.05	0.02	0.04	0.05	0.13	0.05	78.20	40
Within-Domain Top-5	21.87	x	0.07	0.38	0.92	8.30	2.35	84.59	40
Cross-Domain Top-5	**1.51E+04**	1.41	x	9.51	**361.61**	**1112.40**	**502.61**	86.92	40
Within-Domain CWM	**57.88**	0.10	0.05	x	1.27	6.16	7.19	83.54	40
Cross-Domain CWM	13.98	0.07	0.04	0.07	x	1.36	0.53	82.32	40
MPW	0.24	0.05	0.04	0.05	0.07	x	0.08	79.21	40
Minimal Pivoting	6.15	0.06	0.04	0.05	0.09	0.66	x	81.27	40

Experiment 2: Science Domain									
	Simple Average	Within-Domain Top-5	Cross-Domain Top-5	Within-Domain CWM	Cross-Domain CWM	MPW	Minimal Pivoting	Mean Brier Score	N
Simple Average	x	0.052	0.044	0.039	0.044	0.053	0.056	74.27	40
Within-Domain Top-5	6.09	x	0.081	0.1	0.18	0.378	1.965	80.32	40
Cross-Domain Top-5	32.124	0.63	x	0.549	2.081	4.503	17.001	82.72	40
Within-Domain CWM	**177.41**	0.375	0.085	x	1.156	3.847	**163.224**	81.26	40
Cross-Domain CWM	37.106	0.162	0.061	0.069	x	0.556	27.895	80.21	40
MPW	5.446	0.099	0.054	0.055	0.085	x	1.215	78.81	40
Minimal Pivoting	3.737	0.062	0.047	0.04	0.045	0.068	x	76.54	40

Table A4: Bayes-Factors from one-sided paired-samples Bayesian t-test of the row and column method. The H1 hypothesis is that the row method has a strictly greater Brier score than the column method. Bayes-Factors above 40 are interpreted as strong support that the row method better predicts the outcomes than the column method. The mean Brier Score for the row method is shown at the end of each row.

Experiment 3: Art Domain									
	Simple Average	Within-Domain Top-5	Cross-Domain Top-5	Within-Domain CWM	Cross-Domain CWM	MPW	Minimal Pivoting	Mean Brier Score	N
Simple Average	x	0.04	0.06	0.05	0.04	0.27	0.35	72.99	40
Within-Domain Top-5	**273.36**	x	3.04	4.65	3.75	12.68	**45.16**	79.91	40
Cross-Domain Top-5	2.34	0.06	x	0.15	0.18	1.37	1.78	76.41	40
Within-Domain CWM	15.50	0.05	0.20	x	0.24	8.63	30.08	76.67	40
Cross-Domain CWM	**50.16**	0.06	0.16	0.13	x	5.16	**86.08**	76.33	40
MPW	0.12	0.05	0.07	0.05	0.05	x	0.17	71.88	40
Minimal Pivoting	0.10	0.04	0.06	0.04	0.04	0.18	x	71.92	40

Experiment 3: Emotional Intelligence Domain									
	Simple Average	Within-Domain Top-5	Cross-Domain Top-5	Within-Domain CWM	Cross-Domain CWM	MPW	Minimal Pivoting	Mean Brier Score	N
Simple Average	x	0.06	0.08	0.06	0.05	0.19	0.08	78.24	40
Within-Domain Top-5	3.62	x	0.63	0.83	0.32	1.85	1.19	82.49	40
Cross-Domain Top-5	0.65	0.08	x	0.15	0.09	0.56	0.28	80.77	40
Within-Domain CWM	2.99	0.08	0.21	x	0.09	1.09	0.63	81.05	40
Cross-Domain CWM	5.52	0.11	0.45	0.55	x	1.77	1.21	81.73	40
MPW	0.15	0.06	0.09	0.07	0.06	x	0.07	77.89	40
Minimal Pivoting	0.60	0.07	0.12	0.08	0.07	0.90	x	79.76	40

Experiment 3: Science Domain									
	Simple Average	Within-Domain Top-5	Cross-Domain Top-5	Within-Domain CWM	Cross-Domain CWM	MPW	Minimal Pivoting	Mean Brier Score	N
Simple Average	x	0.02	0.04	0.02	0.04	0.08	0.05	74.57	40
Within-Domain Top-5	**1.36E+05**	x	15.35	**27.05**	**490.03**	**2.48E+04**	**1.65E+06**	85.83	40
Cross-Domain Top-5	**67.99**	0.05	x	0.07	0.39	7.85	**49.62**	81.80	40
Within-Domain CWM	**1.86E+04**	0.05	0.88	x	28.31	**193.16**	**1.41E+05**	83.51	40
Cross-Domain CWM	**283.13**	0.04	0.10	0.05	x	2.68	**104.28**	81.04	40
MPW	0.78	0.02	0.05	0.04	0.06	x	0.24	77.39	40
Minimal Pivoting	5.18	0.01	0.04	0.02	0.04	0.13	x	76.94	40

Table A5: Bayes-Factors from one-sided paired-samples Bayesian t-test of the row and column method. The H1 hypothesis is that the row method has a strictly greater Brier score than the column method. Bayes-Factors above 40 are interpreted as strong support that the row method better predicts the outcomes than the column method. The mean Brier Score for the row method is shown at the end of each row.