The Wisdom of the Coherent:

Improving Correspondence with Coherence-Weighted Aggregation

Robert N. Collins[a], David R. Mandel[a], Chris W. Karvetski[b], Charley M. Wu[c], and

Jonathan D. Nelson[d,e]

[a]Intelligence, Influence and Collaboration Section,

Toronto Research Centre, Defence Research and Development Canada

[b]KaDSci (Knowledge and Decision Science)

[e]University of Tübingen

[d]Max Planck Institute for Human Development

[e]University of Surrey

Author Note

Robert N. Collins: https://orcid.org/0000-0002-1714-7215

David R. Mandel: https://orcid.org/0000-0003-1036-2286

Chris W. Karvetski: https://orcid.org/0000-0001-5205-7066

Charley M. Wu: https://orcid.org/0000-0002-2215-572X

Jonathan D. Nelson: https://orcid.org/0000-0002-1956-6691

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Robert N. Collins, Email: robert.collins@drdc.rddc.gc.ca

Abstract

Previous research shows that variation in coherence (i.e., degrees of respect for axioms of probability calculus), when used as a basis for performance-weighted aggregation, can improve the accuracy of probability judgments. However, many aspects of coherence-weighted aggregation remain a mystery, including both prescriptive issues (e.g. how best to use coherence measures) and theoretical issues (e.g. why coherence-weighted aggregation is effective). Using data from previous experiments employing either general-knowledge or statistical information-integration tasks, we addressed many of these issues. Of prescriptive relevance, we examined the effectiveness of coherence-weighted aggregation as a function of judgment elicitation method, group size, weighting function, and the aggressivity of the function's tuning parameter. Of descriptive relevance, we propose that coherence-weighted aggregation can improve accuracy via two distinct, task-dependent routes: a *deterministic route* in which the bases for scoring accuracy depend on conformity to coherence principles (e.g., Bayesian information integration) and a *diagnostic route* in which coherence serves as a cue to correct knowledge. The findings provide support for the efficacy of both routes, but they also highlight why coherence weighting, especially the most aggressive forms, sometimes impose costs to accuracy. We conclude by sketching a decision-theoretic approach to how the *wisdom of the coherent* within the wisdom of the crowd can be sensibly leveraged.

Keywords: coherence, correspondence, accuracy, probability judgment, aggregation

Introduction

Decision science is defined by a meta-theoretical schism between coherence and correspondence theorists. Both camps share an interest in the field's theoretical triad—namely, the descriptive, normative, and prescriptive quality on judgment and choice (Bell et al., 1988; Hammond, 2000; Kleindorfer et al., 1993; Mandel, 2000)—but they approach these issues in different ways (Dawson & Gregory, 2009; Dunwoody, 2009). Coherence theorists (Davidson, 1986; Young, 1996) focus on the internal consistency of judgments, while correspondence theorists (David, 2002; Patterson, 2003) focus on the empirical accuracy of judgments. Given their divergent interests, few studies have examined the conceptual and empirical connections between individuals' coherence and correspondence. Early studies found no consistent correlation between coherence and correspondence across multiple experiments varying in subject expertise and task difficulty (e.g., Wright & Ayton, 1987; Wright et al., 1988; Wright et al., 1994). Wright et al. (1994) concluded that individual differences in coherence might not be useful for predicting correspondence. Such findings suggest that the divergent histories of the coherence and correspondence schools of thought may not be accidental. If individual differences in coherence and correspondence are unrelated, why should it matter if the two schools pursue parallel or divergent meta-theoretical paths?

We believe this conclusion is premature. Wright and colleagues' experiments were statistically underpowered to detect significant correlations at conventional error rates (i.e., Type I = 5%, Type 2 = 20%). Additionally, the correspondence score was a mean-squared-error function, whereas the coherence score used mean errors. Thus, the correspondence measure summed opposing errors (e.g., over- and under-estimation), whereas the coherence measure canceled opposing errors (e.g., super- and sub-additivity). These differences might obscure correlations among measures of coherence and correspondence. Indeed, more recent studies have found large correlations between coherence and correspondence as determined by arithmetic accuracy (Weaver & Stewart, 2012; Weiss et al., 2009), challenging the unrelatedness hypothesis directly. Research has also found correlations between coherence (according to Bayes's theorem) and correspondence in predicting the winner the 2011 Major League Baseball series (Tsai & Kirlik, 2012). Furthermore, research shows that superforecasters—elite forecasters who scored in the top 2% of accuracy rankings in a large geopolitical forecasting tournament—perform better than other forecasters and undergraduates on probing logical coherence and internal consistency (Mellers et al., 2017).

Going beyond research that examines the correlation between coherence and correspondence, recent studies show that coherence can be leveraged to improve judgment accuracy. There are two primary mechanisms: *recalibration* methods that "coherentize" judgments (Karvetski et al. 2013; Predd et al., 2009; Mandel et al., 2018) and performance-weighted *aggregation* methods that weight aggregate estimates from a select crowd (Mannes et al., 2014)—namely, the coherent *within* the crowd. Predd et al. (2008) showed that coherence-weighted aggregation improved group forecast accuracy on topics such as sports and economics. Several studies have since generalized the utility of this method to other tasks including US presidential election forecasts (Wang et al., 2011), answers to general-knowledge and forecasting questions (Fan et al., 2019; Karvetski et al., 2013), and Bayesian judgment tasks (Karvetski et al., 2020; Mandel et al., 2018). These findings show that coherence and correspondence may, in fact, be strongly related and that knowledge of the former can be exploited to improve the latter.

The fact that individual differences in coherence sometimes predict correspondence is of prescriptive theoretical interest. Weighted aggregation methods typically require the completion of an additional task or recordkeeping of judges' past performance to derive aggregation weights (Cooke, 2015). By comparison, coherence weighting requires neither and can even be applied on a per-item, per-assessor basis. Coherence weighting simply requires a minimum of two logically-related events upon which coherence can be assessed. The number of elicitations required is comparable to other popular elicitation methods such as those requiring the construction of probability intervals (e.g., O'Hagan, 2019; Speirs-Bridge et al., 2009) or the estimation of other assessors' answers (Palley & Soll, 2019; Prelec et al., 2017). Whereas Surowiecki (2004) proposed that aggregators can improve accuracy by exploiting *the wisdom of crowds* and whereas past research has shown that "linear opinion pools" using unweighted arithmetic averages can improve judgment accuracy (Clemen & Winkler, 1999), we propose that further accuracy gains can be achieved by leveraging *the wisdom of the coherent* within the crowd.

One aim of the present work is to compare alternative coherence-weighted aggregation methods used in previous studies (e.g., Fan et al., 2019; Karvetski et al., 2013; Karvetski et al., 2020; Mandel et al., 2018; Predd et al., 2008; Wang et al., 2011) across multiple tasks and experiments. We wanted to determine whether the best coherence-weighting method—as determined by empirical accuracy—is stable or task dependent (Zellner et al., 2021). A second aim is to analyze whether individuals' out-of-sample coherence predicts correspondence; namely, whether coherence on one or more tasks predicted accuracy on another task. A third aim is to extend coherence-weighted aggregation methods across more complex sets of probabilities than have been previously examined. Earlier studies used variants of the linear additivity constraints (e.g., Predd et al., 2008; Wang et al., 2011). We extend both recalibration and aggregation methods to judgments bound by nonlinear constraints. Last but not least, we aim to explore why (and under what circumstances) coherence weighting can be effective for different tasks.

To pursue these aims, we reanalyzed data from two experiments reported in Karvetski et al. (2013) and four experiments reported in Wu et al. (2017). The experiments feature distinct tasks: Karvetski et al. (2013) had participants answer general-knowledge questions and assign probabilities that factual claims are correct, whereas Wu et al. (2017) had participants assess probabilities on the basis of statistical background information, characteristic of static Bayesian tasks (Mandel, 2014). Before analyzing these datasets, we expand on how probabilistic coherence is formally defined, calculated, and used in performance-weighted aggregation.

Probabilistic Coherence

When people estimate the probabilities of related events, their estimates will often be incoherent (Mandel, 2008; Karvetski et al., 2013; Predd et al., 2008; Tversky & Koehler, 1994), in that they violate some number of axioms of probability calculus. Notably, Kolmogorov (1933; see also De Finetti, 1937) described three probability axioms: non-negativity, unitarity, and additivity. Non-negativity states that probabilities must not take on negative values. Unitarity (also sometimes called complementarity) states that the summed probability of elementary events must equal 1. Additivity states that any countable sequence of mutually exclusive events $E_1, E_2, \ldots, E_n$ must satisfy the condition:

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) \qquad [1]$$

Thus, we can define probabilistic coherence as the extent to which sets of probabilistic judgments respect Kolmogorov's (or other) probability constraints.

In the present research, we measure incoherence using the *coherence approximation principle* (CAP; Osherson & Vardi, 2006; Predd et al. 2008). The CAP measures the Euclidean distance between an elicited probability set and the closest set of coherent probabilities, returning a recalibrated (or coherentized) set of probabilities. The distance between these two sets is the incoherence metric, ι. The recalibrated set of probabilities is provably unique for sets of probabilities bound by linear constraints such as Kolmogorov's axioms (De Finetti, 1990; Karvetski et al., 2013). More generally, the process is guaranteed to result in monotonic improvements in accuracy when applied to convex sets of probabilities (Predd et al., 2009). A set of probabilities *S* is convex if there are no arbitrary points *a* or *b* contained within *S* such that a line drawn between points *a* and *b* would contain points that do not belong to the set. If this property is not satisfied, the set *S* is non-convex and monotonic improvements are not guaranteed (i.e. an elicited but incoherent estimate may have better accuracy than its coherent equivalent). Whether a set of probabilities is convex or not depends on the geometry of the coherentized probability space as defined by its constraints.

To facilitate understanding of the CAP, Collins (2021) developed the *Interactive Coherence Approximation Principle Tool* (https://collinsrn.shinyapps.io/CAP-TOOL/) using *R Shiny* web app platform (Chang et al., 2020). The tool provides a visual, interactive application of the CAP to a set of two mutually exclusive probabilities, $P(A)$ and $P(B)$, and their summed union probability**,** $P(A \cup B)$, as used in the study by Karvetski et al. (2013). Figure 1 depicts one example of the tool, in which the elicited judgments, $y = \{P(A), P(B), P(A \cup B)\} = \{1, 1, 1\}$, exhibit subadditivity, violating the union constraint $P(A) + P(B) = P(A \cup B)$. The CAP algorithm searches the coherent probability space (defined by the grey polygonal plane) to identify the nearest set of coherent probabilities, returning the recalibrated judgments, $p = \{.5, .5, 1\}$ and coherence measure ι = 0.71 according to:

$$\iota = \sqrt{(P(A) - y_1)^2 + (P(B) - y_2)^2 + (P(A \cup B) - y_2)^2} \qquad [2]$$

More precisely, the CAP is a constrained optimization problem focused on minimizing the Euclidean distance between the elicited and coherent probability sets. The incoherence metric, ι, is therefore an objective measure of individuals' probabilistic coherence, which serves as the input for coherence-weighted aggregation judgments. In the coherence-weighted aggregation functions that we tested in the present research, the derived aggregation weight is always a monotonically decreasing function of ι such that high ι produces low aggregation weights.

As a final note, though recalibration is not the focus of our research, we analyzed the participants' recalibrated rather than elicited probabilities for two reasons. First, the process for recalibration and calculating ι are one in the same. Since coherence-based recalibration improves accuracy (De Finetti, 1990; Karvetski et al., 2013; Mandel et al., 2018; Predd et al., 2009), we see no compelling rationale for applying coherence-weighted aggregation on raw judgments. Second, by first recalibrating judgments to remove error due to incoherence, we can isolate the independent performance-weighting potential of coherence weighting separately from the reduction of incoherence due to recalibration.

Coherence Weighting Functions

There is no consensus on the "best" function for converting ι into aggregation weights. Table 1 lists some of the functions used in prior research (Fan et al., 2019; Karvetski et al., 2013; Wang et al., 2011). Each has different properties. The exponential function (Wang et al., 2011) translates ι into a corresponding aggregation weight, ω, using an exponential function with Euler's number, e, as its base. The function is insensitive to the size of the aggregation pool and the variability of ι within the aggregate pool. The standard linear difference function (Karvetski et al., 2013) calculates an aggregation weight based on the relative difference between an individuals' ι and the maximum ι within the aggregate pool. The function is sensitive to the variability but not the size of the aggregate pool. Furthermore, because it assigns the most incoherent individual a weight of 0 it is undefined at group size 1. Finally, the ranked method (Fan et al., 2019) calculates an aggregation weight that is the multiplicative inverse of its rank. In the case of a tie, the average of their rank, i.e. three participants tied for first would receive a rank of $\frac{1+2+3}{3} = 2$. The function is sensitive to pool size but not to variability in its distribution.

Two functions, the exponential function (Wang et al., 2011) and the standard linear difference function (Karvetski et al., 2013), were presented in their original form with a practitioner-controlled coherence-weight parameter. This tuning parameter was referred to as λ in Wang et al. (2011) and β in Karvetski et al. (2013). They are conceptually and mathematically similar, exponentiating the translation of ι into a corresponding weight. In the present article, we adopt the symbol convention β used by Karvetski et al. (2013) for both functions. We also parameterized the rank function (Fan et al., 2019) by raising the function to exponent β. Regardless of function, β = 0 resolves to equivalent ω = 1 for every participant (i.e., equal-weighted aggregation). As β increases, the penalty for coherence violations increases and the weight given to coherent responders increases. Sufficiently large values of β will reduce the contribution weight of all but the most coherent individual(s) to 0. Simply put, β tunes the aggressivity of the weighting function.

Regardless of the functional form or tuning of ω, coherence-weighted aggregation proceeds identically. For each probabilistic judgment *y* and by *n* individuals for each *i*th judgment, we define an aggregated judgment as:

$$y_i = \frac{\sum_{j=1}^{n} y_{ij} \times \omega_{ij}}{\sum_{j=1}^{n} \omega_{ij}} \qquad [3]$$

In the present research, we consider three possible values for β ∈ {1, 10, 100}. The decision to use discrete values rather than continuous was pragmatic rather than theoretically driven and, technically, any value $\beta > 0$ is a valid choice. However, in practice, very large values (e.g., $\beta = 1000$) can produce undefined and rounding errors due to the extremely low ω values produced[1]. An important property to note is that setting $\beta = 0$ results in $\omega = 1$ for all participants. This produces the equivalent of an equal-weighted arithmetic average regardless of which function is used. Consequently, rather than simulating $\beta = 0$ for each of our functions, we use the equal-weighted arithmetic average for comparison:

$$y_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij} \qquad [4]$$

Performance Metric

We measured accuracy using the mean absolute error (*MAE*):

$$MAE = \tfrac{1}{k}\textstyle\sum_{i=1}^{k} |y_i - x_i|, \qquad [5]$$

where y_i is the judgment, x_i is the true value, k is the number of judgments made by the participant, and *i* represents a particular judgment. Although there is no "correct" scoring rule, Wilmott and Matsura (2005) propose that *MAE* has several beneficial properties. In our case, use of *MAE* also adds new information to the reanalysis of Karvetski et al. (2013), which was originally scored using the Brier score (i.e., mean squared error).

Importantly, both our measure of coherence, ι, and correspondence, *MAE*, summate errors in opposing directions. That is, subadditivity and superadditivity are summated rather than cancelled out for our ι measure. Likewise, under-estimation and over-estimation are summated

[1] For example, at $\beta = 1000$, the second most coherent individual would be assigned a weight $\omega = 9.33 \times 10^{-302}$, assuming no ties among the ranks. A rank of just 3 would produce a number below the lowest possible value that can be represented by many programs and programmatic languages, e.g., *R*'s lower limit of 5×10^{-324}.

for our *MAE* measure.

General-knowledge Domain

Karvetski et al. (2013) investigated the effect of coherence-based recalibration and aggregation on truth ratings of general-knowledge questions. In two experiments, 58 undergraduate psychology student participants (30 in Experiment 1, 28 in Experiment 2) rated the probability that four logically-related statements were true. For each of 60 topics, participants rated the probability that: (1) statement A was true, $P(A)$; (2) statement A was false, $P(A^c)$; (3) statement B was true (where $A \cap B \in \varnothing$), $P(B)$; and (4) either statement A or statement B was true, $P(A \cup B)$. Karvetski et al. (2013) tested several variants of coherentization and methods for calculating ι, the most effective of which was based on the additivity constraint $P(A) + P(B) = P(A \cup B)$.[2] We use the same constraints for recalibration and aggregation in our reanalyses. In Experiment 1, related judgments were spaced such that participants rated one randomly selected member of each topic set, cycling through all other topics before rating the next statement in the topic set, repeating this process four times until all 240 judgments were made. In Experiment 2 related statements were grouped such that participants concurrently rated the four statements for a given topic consecutively, with topic order and item order within any given set randomized. Given that all else was identical between Experiments 1 and 2, we refer to these as the spaced and grouped conditions, respectively.

Method

All data and *R* scripts described below may be found in the supplemental materials and

[2] The 'four-way' coherentization scheme, which included $P(A^c)$, reduced the effectiveness of coherence-weighted aggregation. Karvetski et al. (2013) suggested that individuals expressing epistemic uncertainty (i.e., responding .50) are incidentally coherent with respect to the complementarity constraint between $P(A)$ and $P(A^c)$, obscuring the correspondence 'signal'.

on the Open Science Framework (Collins et al., 2021).

Incoherence

Coherentized probabilities and ι were calculated using the quadprog package in *R Studio* (Turlach, 2019) running the *R* programming language. Participant judgments were submitted to the quadratic programming algorithm using the CAP, outlined in Figure 1 and Equation 3.

In addition to calculating the ι for each question, which we call the *calibration* ι, we also calculated the mean, out-of-sample ι across all other questions, which we call the *validation* ι.[3] That is, for Question 1, we used the average ι for each participant's answers to Questions 2-60 as the input weight; for Question 2, the average ι of each participant's answers to Questions 1 and Questions 3-60; and so on. This measure is intended to test the potential utility of out-of-sample, or past, average ι for use in coherence-weighted aggregation.

Performance Metrics

The *MAE* across judgments [$P(A)$, $P(B)$, $P(A \cup B)$] was calculated separately for each of the 60 questions and serves as the primary dependent variable of interest. We also examine the proportion of bootstrapped samples (out of 1000, described in greater detail below) in which the coherence-weighted aggregation strategies performed better than equal-weighted strategies. That is, the proportion of times $MAE_{CW} < MAE_{EW}$, where CW stands for coherence-weighted and EW stands for equal-weighted, respectively. We call this metric the *proportion improved* (*PI*).

Aggregation

Recall that an aim of our research is to examine which implementations of coherence weighting work best. To compare the efficacy of our various strategies, we used a bootstrap

[3] Though bearing some conceptual similarities, it is important to note that these definitions are distinct from the nomenclature and definitions of 'calibration' and 'validation' used in the machine-learning literature.

sampling with replacement technique that analyzed different combinations of weighting function, parameterization, and group size. For each bootstrapped sample, we added a randomly selected participant to the aggregate pool one at a time until the maximum group size, k = 100, was achieved. Each time we added a participant to the aggregate pool, we calculated aggregated *MAE* across the 60 questions for each combination of ι weights (*calibration* vs. *validation*), weighting functions (equal-weighted, exponential, standard linear difference, rank) and levels of coherence weight β (1, 10, 100), corresponding to a 2 × 3 × 4 design. We repeated this process 1,000 times for each condition (grouped vs. spaced). This approach is conceptually similar to a *multiverse* analysis (Steegen et al., 2016; Harder, 2020), in the sense that comparisons of different ι weights, weighting functions, levels of β, and group size decisions can be made while holding all else equal. However, rather than simply offering transparency in our analysis decisions, we use the multiverse approach to develop recommendations of best practices using coherence-weighted aggregation. Combined with the high precision afforded by the number of bootstraps, even a small absolute difference between methods, parameters, or group sizes will be statistically significant. The final analysis included 3,800,000 unique data points.[4] We collapsed across the 1,000 bootstrap samples, producing 3,800 *MAE* values (1,900 per condition) as well as 3,600 *PI* values (1,800 per condition).

Results

Calibration

Figures 2 and 3 show the results of the coherence-weighted aggregation for *MAE* and *PI,* respectively, as a function of k and β. The x-axis for group size k is logarithmically scaled (base

[4] Complete breakdown as follows: 2 conditions (Spaced vs. Grouped) × 100 group sizes (1-100) × 19 aggregation strategies (consisting of 3 coherence weighed functions [Exponential vs. Standard Linear Difference vs. Rank] × 3 coherency weight values [1, 10, 100] + equal-weighted function) × 1000 bootstraps.

10). Aggregation performance at $k = 1$ is equal to the sample *MAE* regardless of aggregation strategy. A corollary of this statement is that $PI = 0$ at $k = 1$ because $MAE_{CW} = MAE_{EW}$. Further, because the correct probability of a verifiable fact is either 0 or 1, participant responses cannot bracket the correct probability. Consequently, equal-weight aggregation has no effect on accuracy (Larrick & Soll, 2006).

Each coherence-weighted aggregation method reduced *MAE* compared to equal-weighted aggregation when $k > 2$. Note several important results. First, increasing k beyond 2 reduced *MAE* in most cases. However, the returns diminished as k increased, such that the addition of new assessors had relatively little effect for $k > 10$. Nevertheless, accuracy continued to improve past these thresholds, albeit at a much lower rate. A major exception to this was the standard linear difference function at $\beta = 1$, for which performance worsened for $k > 4$. Second, increasing β consistently improved *MAE* for each of our methods. These improvements were larger in the spaced condition than in the grouped condition. Increasing β was also associated with diminishing returns regardless of condition, with the largest improvements occurring between $\beta = 1$ and $\beta = 10$. Third, comparing our different methods, the standard linear difference function performed best in the spaced condition at $\beta = 1$ and where $k < 5$. In all other cases, the *MAE* for the rank function was equal to or better than the standard linear difference and exponential function. Notably, all *MAE* functions converged as both k and β increased such that performance is nearly identical for $k > 4$ and $\beta = 100$.

For the *PI* measure, each coherence-weighted aggregation method massively improved upon the equal-weighted *MAE* in both experimental conditions. In the spaced condition, $PI > .82$ at $k = 2$ for each method, converging at $PI = 1$ at $k = 11$ for every method. In the grouped condition, performance was worse, starting at $PI > .59$ at $k = 2$ for each method, converging at *PI*

= 1 at k = 27 for all methods. In contrast to increasing k, increasing β decreased *PI.* This was more pronounced in the grouped condition than in the spaced condition. When comparing the weighting functions, the standard linear difference function performed worse at lower group-size values, while the exponential and rank functions perform similarly.

Validation

Figures 4 and 5 show the results of the coherence-weighted aggregation for *MAE* and *PI,* respectively, as a function of k and β. A key finding is that the validation ι weight, i.e., one's average propensity for coherence across out-of-sample or past problems rather than a direct task-specific measure of coherence, was significantly less effective than the calibration ι, both in terms of *MAE* and in terms of *PI.* For the spaced condition, there were three important findings. First, as with calibration ι, increasing β improved *MAE* for each function. Second, increasing k had nuanced effects on *MAE* in the spaced condition, particularly for $\beta \geq 10$. *MAE* improved up to k = 5, had inconsistent effects for $5 < k < 20$, and steadily improved again for $k \geq 20$. Third, as with the calibration ι, the standard linear difference function performed best at β = 1 and where k < 5. For all other combinations of k and at β, the rank function performed as good as or better than either the exponential or standard linear difference function. Finally, weighting by validation ι produced almost no consistent or reliable improvements in the grouped condition.

For the *PI* measure, coherence weighting was effective in both conditions. This implies that coherence weighting consistently improved *MAE* in the grouped condition as well, but that the magnitude of improvement was very small. Beyond this, there were three important findings. First, if β = 1, increasing k had either no effect or a negative effect on *PI* for each function and value k. Second, increasing k again had a similarly nuanced effect. For lower values of at β = 1, increasing k tended to improve *PI.* However, as at β increased, each function conformed to a similar shape as observed with *MAE*: first improving, then worsening slightly, and then improving again. Third, no function is clearly superior, although *PI* for the exponential function was highest in most cases.

Discussion

The aggregation results confirm coherence-weighted aggregation is an effective tool in the general-knowledge domain but with caveats. Regarding our primary aim to identify the best performing coherence-weighted aggregation strategies, there are several important observations. Increasing group size *k* improves aggregated accuracy but with diminishing returns. In this experiment, where participants answered three related questions on 60 different topics, the greatest variability among aggregation methods was seen between $1 < k < 10$. Coherence-weighted methods also presented a slight risk-reward trade-off: increasing β improved accuracy but decreased *PI* slightly. The reason is that very large values of β often resulted in convergence on a coherent and certain set of probabilities, which was most often correct. However, if the coherent individual in the aggregate pool is incorrect, this will result in wildly divergent answers with large penalties to accuracy.

Although there is no unambiguously superior weighting function—each converged at high *k* and β—we believe there are reasons to prefer the exponential function. Compared to the standard linear difference function, the exponential function is unambiguously superior in terms of *PI* and did not worsen when *k* was increased. Compared to the rank function, the exponential function was less aggressive (i.e., it did not penalize incoherence as much) and also less effective at $\beta = 1$. It is conceivable that this more modest weighting strategy may be desirable and effective in some cases, thus the added flexibility of lower coherence weight is an attractive feature. We also suggest the translation of ι into corresponding weights ω is more intuitive, predictable, and useful for the exponential function than for the other functions.

Regarding our secondary aim to determine whether individual differences in average levels of coherence are useful predictors of correspondence, we found limited supporting evidence. Compared to calibration ι, validation ι was much more error-prone. Weighting according to validation ι only produced large numerical improvements in the spaced condition, demonstrating important constraints on the advantages of coherence weighting reported in Karvetski et al. (2013). Moreover, within the spaced condition, changes in both group size and weighting had inconsistent effects if validation ι was used. Nevertheless, each of our functions improved over the equal-weighted average in most cases, suggesting that while there is little to be gained—in both absolute and relative terms—there is also little to be lost. Importantly, small gains in empirical accuracy have the potential to be highly consequential.

Finally, the results shed light on why coherence weighting is effective in the general-knowledge domain. Unsurprisingly, like Karvetski et al. (2013), we saw that coherence weighting was most effective in the spaced condition. This is counterintuitive though, as the significant gaps between related judgments make it difficult to remain coherent: participants attempting to satisfy the coherence constraints would have to remember the truth rating they provided at least 60 questions ago! An explanation for this effect relies on the fact that correct responses are, by definition, coherent (Hammond, 2000). When coherence is difficult to maintain, only those knowledgeable experts who know the correct response are consistently able to maintain it. In other words, spacing responses does not affect the rate of true positives (i.e., individuals who are correct and coherent), but it does reduce the rate of false positives (i.e., individuals who are coherent but not correct). The importance of this balance between true and false positives is further exemplified in a Bayesian judgment experiment (Karvetski et al., 2020), where coherence-weighted aggregation performed best when the aggregation pool had a small number of primarily coherent judges. The results suggest that coherence-weighted aggregation does not exploit a direct relationship between the probabilistic numeracy required to maintain coherence and the knowledge required for correspondence. Rather, the results suggest that individual differences in coherence can be used to *diagnose* potential experts in the general-knowledge domain. Moreover, this diagnostic process benefits from spacing logically-related judgment queries farther apart so that coherence does not misdiagnose “mere” awareness of the logical constraints, which are unlikely to be reliable indicators of accurate world knowledge.

Statistical-Evidence Domain

Wu et al. (2017) investigated how information presentation formats influenced the efficacy of information search queries and the accuracy of unaggregated probability judgments. Here, we examined the effect of coherence-weighted aggregation on these judgments. Across four experiments, 2,858 participants completed the *Turtle Island* task. The fictional scenarioin the task concerned an island populated by two turtle species (i.e., Bayosians and Freqosians). The species are visually identical and can only be identified by the differential expression of two genes: the *DE* gene or the *LM* gene. Each gene has two forms (*D* or *E* for *DE*; *L* or *M* for *LM*). The rate of expression for each form of each gene differs between species. Participants received statistical evidence regarding turtles and their genes and judged prior, marginal, likelihood, and posterior probabilities regarding the species and possible gene test results. Wu et al. (2017) subsequently measured the accuracy of these probabilities. Critically, using Bayes's theorem (or intuitive reasoning that is otherwise consistent with Bayes’s theorem) it is possible to calculate these probabilities precisely. Therefore, unlike Karvetski et al. (2013), both coherence and correspondence depended on the ability to compute probabilities in a manner consistent with

Bayes’s theorem.

In a between-subjects manipulation, Wu et al. (2017) considered 14 different information formats, including numeric probability formats, numeric frequency formats, and various graphical formats presented as a combination of priors and likelihoods or marginals and posteriors. We are interested primarily in this latter manipulation. Half of the 14 information formats provided participants background evidence in the form of Bayesian *prior and likelihood* (PL) probabilities; the other half received statistical background evidence in the form of Bayesian *marginal and posterior* (MP) probabilities. The 7 information formats within each condition were otherwise identical. For example, a “numeric” versus “bar graph” manipulation was present within the PL condition and within the MP condition. Consequently, we collapsed across these sub-conditions and analyzed accuracy at the level of our primary grouping variable (i.e., PL vs. MP).

Method

The *R* scripts and aggregated data described below may be found in the supplemental materials and on the Open Science Framework (Collins et al., 2021)[5].

Incoherence

Whereas Karvetski et al. (2013) examined sets of probabilities bound by Kolmogorov’s axioms, the estimates elicited in the Turtle Island task concerned sets of probabilities bound by Bayes’s theorem. To the best of our knowledge, no practitioner has applied the CAP to such a problem. In particular, the nonlinear constraints of Bayes's theorem pose a challenge for the coherentization procedure. The theorem describes the probability of an event’s occurrence by

[5] For the participant-level raw and coherentized data sets, please contact co-author Charley M. Wu (charleymswu@gmail.com).

combining unconditional and conditional probabilistic knowledge related to the event. Bayes's theorem is formally expressed as follows:

$$P(x|y) = \frac{P(x) \times P(x|\mathrm{y})}{P(y)} \qquad [6]$$

Assuming *x* and *y* refer to distinct events or classes, the formula consists of the following components: the prior, *P*(*x*), an unconditional probability describing the chance that *x* will occur; the likelihood, *P*(*x*|*y*), a conditional probability describing the chance that *x* will occur given that *y* occurred; the marginal, *P*(*y*), an unconditional probability describing the chance that *y* will occur; and the posterior, *P*(*y*|*x*), a conditional probability describing the chance that *y* will occur given that *x* occurred.

In the Turtle Island experiments, the labeling of the underlying probabilities (e.g., whether Bayosian or Freqosian was the more populous species) was randomly assigned to each participant. For the purpose of analysis, we canonized these probabilities as follows: (a) *P*(*B*) is the prior probability that a turtle was a Bayosian turtle; (b) *P*(*E*) is the marginal probability that a turtle had the *E* form of the *DE* gene and, similarly, *P*(*L*) is the marginal probability that a turtle possessed the *L* form of the *LM* gene; (c) *P*(*E*|*B*) and *P*(*E*|*F*) are the likelihood probabilities that a Bayosian and Freqosian turtle expressed the *E* form of the *DE* gene, respectively, and similarly, *P*(*L*|*B*) and *P*(*L*|*F*) are the likelihood probabilities that a Bayosian or Freqosian turtle expressed the *L* form of the *LM* gene, respectively; and (d) *P*(*B*|*E*) and *P*(*B*|*D*) are the posterior probabilities that a turtle with the *E* or *D* form of the *DE* gene was a Bayosian turtle, respectively, and similarly, *P*(*B*|*L*) and *P*(*B*|*M*) are the posterior probabilities that a turtle with the *L* or *M* form of the *LM* gene was Bayosian, respectively. The probability of binary complements was constrained and submitted simultaneously such that *P*(*F*), the prior probability that a turtle was Freqosian, was equal to 1 – *P*(*B*). This was achieved by using a visual analog slider for the participant to input probability values; if, e.g., $P(B)$ was set to 20%, then $P(F)$ would automatically be set to 80%, and both probabilities were visually apparent.

To make the problem computationally manageable, and to provide ample opportunities to assess the relationship between coherence and correspondence, we analyzed the accuracy and coherence of participants' unconditional and conditional probability estimates separately. The type of probability differed between conditions and between probability components. So too did the calculation of ι and *MAE*. The full list of probability sets and their coherence constraints can be found in Appendix A. Here, e.g., consider the estimate of marginal probabilities in the PL condition. The derivation of the marginal probability $P(E)$ given a combination of priors and likelihoods can be expressed as:

$$P(E) = P(B) \times P(E|B) + P(F) \times P(E|F) \quad [7]$$

and participants must also estimate $P(L)$, which can be expressed as:

$$P(L) = P(B) \times P(L|B) + P(F) \times P(L|F) \quad [8]$$

We applied the CAP to these derivations using the "NlcOptim" package in R (Chen & Yin, 2019). For the estimation of marginal probability estimates in the PL condition, let vector $y = [y_1, y_2, y_3, y_4, y_5, y_6, y_7]$ be the elicited vector of probability estimates and vector $p = [P(B), P(E), P(L), P(E|B), P(E|F), P(L|B), P(L|F)]$ be the proximate coherentized probability estimates. According to the CAP, we can calculate the ι and coherentized probabilities by minimizing the square root of the squared deviation between these probability sets (as in Equation 2), subject to the following inequality constraints:

$$0 \leq \{P(B), P(E), P(L), P(E|B), P(E|F), P(L|\mathrm{B}), P(L|F)\} \leq 1, \quad [9]$$

as well as the following equality constraints (constructed by rearranging equations 7 and 8):

$$0 = P(E) - \{P(B) \times P(E|B) + [1 - P(B)] \times P(E|F)\} \quad [10]$$

$$0 = P(L) - \{P(B) \times P(L|B) + [1 - P(B)] \times P(L|F)\} \qquad [11]$$

The optimization procedure also requires user-set *tolerance* parameters. Tolerances control the stopping rules for the optimization procedure. The *constraint* tolerance defines the minimum constraint violation below which the coherentized estimate satisfies the constraint and is considered valid. This is a necessary but insufficient criterion for stopping the optimization procedure. The *objective* tolerance defines the minimum improvement in ι required for the optimization to select a new iteration solution over a previous iteration solution, and theoretically sets a soft lower bound for ι (i.e. the optimization will continue until the ι is equal to or lower than the tolerance). The *X* tolerance defines the minimum shift in probability values required for the optimization procedure to consider a new set of values. The procedure ceases if and only if all three tolerance criteria are satisfied. Each of our tolerances was set to .01, the lowest possible value a participant could submit and also the lowest increment for participant answers.

Due to a combination of the precision afforded to participant responses, rounding to an identical level of precision following coherentization, and the potential non-uniqueness of coherentization solutions, it is possible for ostensibly coherent probability sets to receive a score of $\iota = .01$. To compensate for this rounding-induced incoherence, we treated all values of $\iota \leq .01$ as if they were $\iota = 0$.

For the validation ι, we used the ι calculated for the opposite component: for the unconditional probability estimates, we weighted participant responses using the conditional probability estimate ι, and vice versa.

Performance Metrics

We calculated the *MAE* for the unconditional probability estimates and the conditional probability estimates separately before averaging the two. The *MAE* reflected the error across all probabilities relevant to the coherentized set, including both the given estimates and the derived estimates. As with the general-knowledge domain, we also calculated *PI*, the proportion of times $MAE_{CW} < MAE_{EW}$ for all bootstrapped simulations. Finally, because the only difference between the four experiments in Wu et al. (2017) are the values of the canonized environmental probabilities, we collapsed across the four experiments for our analyses.

Aggregation

We compared the effectiveness of our aggregation methods using a method identical to that described in the prior study with one exception. Because the Turtle Island experiments had different probabilities for the priors, likelihoods, marginal and posterior probabilities, we performed 1,000 bootstrapped samples for each of the four experiments, collapsing afterward. After collapsing across bootstraps and experiments, and excluding redundant combinations of conditions, the final analyses consisted of 15,200,000 unique data points.[6] We collapsed across the 1,000 bootstrap samples and 4 experiments to 3,800 *MAE* values (1,900 per condition) as well as 3,600 *PI* values (1,800 per condition).

As a final note, unlike the general-knowledge tasks bound by Kolmogorov's axioms, sets of Bayesian probabilities are non-linear and do not form a convex set. Therefore, recalibration is not guaranteed to improve accuracy and the linear average of these probabilities is unlikely to be coherent. This introduces a component of error in the aggregated responses that was not present in our analyses of the Karvetski et al. (2013) dataset.

Results

[6] Complete breakdown as follows: 2 conditions (Prior and Likelihood vs. Marginal and Posterior) × 100 group sizes (1-100) × 19 aggregation strategies (consisting of 3 coherence weighed functions [Exponential vs. Standard Linear Difference vs. Rank] × 3 coherency weight values [1, 10, 100] + equal-weighted function) × 1000 bootstraps × 4 experiments (1-4)

Calibration

Figures 6 and 7 show the results of the coherence-weighted aggregation for *MAE* and *PI*, respectively, as a function of *k* and β. Again, the *x*-axis for group size *k* was logarithmically scaled (base 10). Note some preliminary points: First, as in the prior study, aggregation performance at group size $k = 1$ was equal to the sample *MAE*, providing a benchmark. Second, participants' responses can bracket the correct answer and, therefore, equal-weighted aggregation was effective at reducing error (Larrick & Soll, 2006). Finally, recall that the standard linear is difference function is undefined at $k = 1$.

Each coherence-weighted aggregation method was effective at reducing *MAE* compared to the equal-weighted average, with some noteworthy exceptions. First, increasing *k* tended to improve *MAE* in most cases, but these improvements were associated with diminishing returns. Again, the largest numerical improvements occurred between $1 < k < 10$, regardless of function or β value. However, unlike the general-knowledge domain, there was a critical point for each of our functions beyond which increasing *k* worsened accuracy slightly. In particular, there is a noticeable worsening of accuracy when comparing $k = 25$ with $k = 100$ in the PL condition, though this trend does not revert the aggregated accuracy to that of equal-weighted aggregation. Second, increasing β improved *MAE* up to a certain point, after which accuracy worsened. This worsening was more pronounced for larger group sizes and with the rank function. Unlike aggregated accuracy in the general-knowledge domain, the coherence-weighted functions did not converge.

Regarding the *PI* measure, each of our methods improved upon the equal-weighted *MAE* most of the time. Note other important results. First, increasing *k* improved *PI* to a critical point, after which *PI* would plateau or even decrease slightly depending on the function and β. Second, increasing β led to a decrease in *PI* for each function and *k* value. Third, comparing our different strategies, the exponential function, once again, performed best most of the time. The results of the *MAE* and *PI* analyses indicate that, given these data and the weighting strategies we considered, the best results are obtained with a group size in the low double-digits and a moderately to highly aggressive (i.e., $\beta \geq 10$) weighting strategy.

To better understand why performance appeared to worsen between $k = 25$ and $k = 100$ in the PL condition where $\beta = 100$ (and to a lesser extent between $k = 50$ and $k = 100$ in the marginal and posterior condition), we examined the distribution of bootstrapped MAE values (Figure 8). As the histogram shows, increasing *k* from 25 to 100 produces a slight increase in cases where *MAE* is between .00 and .05. However, there a slightly larger increase in the number of cases where *MAE* is between .20 and .25, as well as increases in cases with even higher *MAE*. An inspection of the raw data reveals that four participants across three of the experiments answered .50 for every probability estimate. This set of responses and has *MAE* ≈ .25 in each case. However, it is evident that, although coherent, these responses reflect either a straight-lining or "fifty-fifty blip" (e.g., Fischhoff & Bruine de Bruin, 1999) response tendency. Regardless of their basis, these incidentally coherent cases are what cause the secondary peak seen in the histogram. That is, in some cases, the aggressive coherence-weighting strategy "selected" these coherent but inaccurate individuals outright. The larger the group size, the greater the chance that they would be selected. In fact, at maximum group size, it is certain that they are included.

Validation

Figures 9 and 10 show the results of the coherence-weighted aggregation for *MAE* and *PI*, respectively. As with the general-knowledge domain, the validation ι weighting strategy was much less effective than the calibration ι weighting strategy, both in terms of *MAE* and *PI*. Similarly, improvements were also subject to diminishing returns. Nevertheless, coherence weighting improved *MAE* in most instances, both in absolute terms and relative to the equal-weighted average. Otherwise, the results were broadly similar to the calibration ι, including the slight worsening of *MAE* for large group sizes at $\beta = 100$. The results indicate that the best performance was achieved with a combination of moderate-to-large group size combined with a moderate to severely aggressive combination of weighting function and tuning parameter.

Discussion

The results demonstrate that coherence weighting as a method of performance-weighted aggregation can efficaciously be applied to the statistical-evidence domain using nonlinear optimization constraints. In applying the CAP to a set of probabilities bound by nonlinear constraints, we generated an ι that was an effective basis for coherence-weighted aggregation. This shows that the CAP can be generalized to problem sets more complex than those explored in previous studies. We discuss the process by which coherence-weighted aggregation yields benefit in the statistical-evidence domain in the General Discussion.

Regarding what worked best, as with the general-knowledge domain, the exponential function often produced the best results (in terms of lower *MAE*) and behaved more predictably in response to changes in *k* and β. This function was best paired with large group sizes and moderate weighting ($\beta = 10$) or moderate group sizes ($10 \lesssim k \lesssim 25$) and severe weighting ($\beta = 100$). Incremental changes were minimally positive and sometimes negative for $k > 10$, reflecting strongly diminishing returns.

The finding that increasing group size can worsen accuracy is particularly interesting. Intuitively, the addition of more information to the aggregate pool should not worsen accuracy.

This is counterbalanced, however, by the fact that coherence weighting restricts the pool of information to a smaller set of coherent individuals. Our findings show that this process can sometimes result in the selection of a coherent but inaccurate assessor. These individuals represent *false positives*: they are coherent not because of mathematical rigor or internal consistency (as with *true positives*) but rather due to response biases that incidentally produce coherent responses, such the fifty-fifty blip (Bruine de Bruin et al., 2002; Fischhoff & Bruine de Bruin, 1999). Coherence-weighted aggregation on its own cannot discriminate between the false-positive and true-positive cases. Therefore, aggressive coherence-weighting strategies run the risk of selecting false-positive assessors—or averaging them with true-positive assessors—which can significantly worsen accuracy. Thus, optimal group size will depend on the optimal ratio of true-positive to false-positive assessors. We return to this issue in the General Discussion.

Regarding our aim of testing the usefulness of validation coherence weighting, we find that validation ι provided an effective basis for coherence-weighted aggregation. Promisingly, weighting by validation ι improved *MAE* most of the time in most cases in both the PL and MP conditions. Again, combinations of the exponential function with medium group size and severe weighting or large group size and moderate weighting performed best. This is intuitive, given that the knowledge required to produce one coherent Bayesian estimate is intricately related to the knowledge required to produce another coherent Bayesian estimate.

General Discussion

The present investigation offers important findings on prescriptive issues concerning coherence-weighted aggregation. It also offers insight for basic theoretical issues concerning the conditions under which coherence-weighted aggregation is likely to be effective. First, we show that with the right parameter selection, coherence-weighted aggregation can be effective in two highly dissimilar task domains: problems involving the recruitment of general knowledge and problems involving the integration of statistical evidence. Second, we further show that choices of ideal group size and parameter tuning are task dependent. Third, we find that the exponential function has many desirable properties and fewer drawbacks relative to competing weighting functions. Fourth, we show that where coherence is difficult to maintain (spaced general-knowledge questions) or is strongly related to accuracy on the task (statistical-evidence domain), out-of-sample validation ι holds some promise of predicting correspondence on other tasks. Fifth, we show that the CAP (Osherson & Vardi, 2006; Predd et al., 2008) could be extended to sets of nonlinear probabilities. Finally, our results indicate that there are at least two potential signals of correspondence that practitioners can exploit with coherence-weighted aggregation. Specifically, coherence-weighted aggregation can be effective because coherence carries either a *diagnostic* or *determinative* signal. The nature of the signal being exploited has consequences for the optimal coherence-weighting method. Given the implications this has for exploiting the wisdom of the coherent, we turn to this issue next.

The Wisdom of the Coherent: Why Does Coherence Weighting Work?

There are several ways in which coherence-weighted accuracy can improve judgment. An obvious reason is that, by coherence-weighting judgments, aggregation may coalesce around a coherent and plausible set of probabilities, eliminating a component of aggregated estimation error that might otherwise be attributable to incoherence (e.g., Karvetski et al., 2013; Mandel et al., 2018). As noted in the Introduction, we were not principally interested in this mechanism of action, which is why we first coherentized the judgments in both studies. Rather, we have proposed that there are two critical types of task-dependent signals that can be capitalized on to exploit the “Wisdom of the Coherent.”

The first of these signals we have described as a determinative route to accuracy improvement. In such cases where this route applies, the judgment task requires the application of coherence-based rules of information to reach the correct answer. For instance, in the experiments by Wu et al. (2017) and other Bayesian inference tasks (Mandel, 2014), participants are given statistical information (e.g., base rates and diagnostic probabilities) that is sufficient to produce the precise and accurate estimates provided the information is integrated using Bayes's theorem. One might say that knowledge of Bayes's theorem suffices to yield both coherent and probabilistically accurate judgments. However, this obscures the fact that coherence is the basis for scoring accuracy in such cases.

The second signal we have described as following a diagnostic route to accuracy improvement. Take, for instance, a general-knowledge task such as that used in Karvetski et al. (2013): if one is asked to judge the probability that Neil Armstrong was the first person to set foot on the moon, and if one is sure that Armstrong was the first person to set foot on the moon, then that individual will likely be equally certain that anyone other than Armstrong was not the first person to set foot on the moon. This individual is likely to respect the complementarity rule, by assigning $P(A) = 1$ to the "Armstrong" (A) hypothesis and $P(\neg A) = 0$ to the "anyone other than Armstrong" ($\neg A$) hypothesis. Unlike the determinative case, the assessor may not even understand the coherence principles applied. The exploitable signal here heavily leverages the assumption that precise and certain answers are accurate, and therefore may not benefit from the permission of slight coherence violations as we have suggested for the determinative case.

From a decision-theoretic standpoint, it is important for practitioners seeking to improve judgment accuracy through coherence-weighted aggregation to consider which of these routes might be exploitable given the nature of the judgment task. The two routes were dissociable in the present research, but they are not mutually exclusive, in general. Forecasting is a good example of where both types of bases for exploiting coherence may be present. For instance, accurate forecasting may involve having sufficient knowledge of factors that are shaping the outcome as well as the statistical know-how to coherently combine this information. Nevertheless, the signal being exploited has implications for how best to extract the wisdom of the coherent among the crowd. This includes the ideal elicitation process before aggregation, the utility of validation ι (in contradistinction to calibration ι), and the optimal aggregator function (including the decision about optimal group size and weighting function). Next, we examine these issues in greater detail.

Exploiting the Wisdom of the Coherent

Elicitation Method

The judgment context in which practitioners find themselves can also affect decisions about the elicitation process. Take, for instance, the decision to structure elicitations in ways that are aimed at mitigating incoherence or inaccuracy. These include using pair-wise estimates (Por & Budescu, 2017), evaluation frames (Williams & Mandel, 2007), increasing the proximity of related items (Karvetski et al., 2013), using information presentation formats that make complementary categories salient (Wu et al., 2017), and other consider-the-opposite strategies such as dialectical bootstrapping (Herzog & Hertwig, 2014). There are also elicitation procedures that involve assessing confidence intervals in order to improve best estimates (Hemming et al., 2018; cf. Mandel et al., 2020). In the determinative case, it is best to make the elicitation process as easy as possible: any elicitation method that serves to improve coherence should also improve accuracy. Those who understand the probabilistic axioms can capitalize on clarity and achieve low ι, whereas those who do not understand the relevant probabilistic calculus are unlikely to be affected—for better or worse—by the difficulty of the elicitation process.

In the diagnostic case, however, it may be useful to let individual differences in incoherence “flourish” so that they can be profitably exploited through coherence-weighted aggregation. In Karvetski et al. (2013) and the present re-analyses of their datasets, it has been shown that maximally spacing logically-related judgments so that coherence principles, such as additivity, are less mentally accessible is ideal. This elicitation strategy allows the wisdom of the coherent to be more effectively harnessed. If one were only considering elicitation on its own (i.e., without an accompanying aggregation methodology), the idea of spacing judgments to increase incoherence might seem to be a perverse and counterintuitive strategy. However, when considered alongside potential recalibration and aggregation methods, its advantages become clear. The reasoning is that the relationship between coherence and correspondence could be incidental or diagnostic. By grouping logically-related judgments, it is more likely that an assessor is coherent because they understand and accept the relevant principle. If they understand and accept additivity, for instance, they will try to provide additive judgments even if they do not know the correct answer. However, when items are spaced apart, it is less likely that assessors will be aware of the logical constraints on the set of probabilities. In this case, coherent assessors are more likely to be coherent because they know the correct answer and are thus able to provide a certain or near-certain response that respects additivity. Accordingly, weighting by coherence will separate the wheat from the chaff.

Interestingly, several methods hold promise for capitalizing on either signal of correspondence. For instance, one study showed that allowing judges to opt-in (or out) of judgments improved aggregated accuracy (Bennet et al., 2018). The researchers attributed this to the fact that metacognitive assessments of one's own expertise can be leveraged for accuracy.

Likewise, elicitations that determine what answers are “surprisingly popular” (Prelec et al., 2017), or that gather information about the gap between an assessor’s judgment and the assessor’s judgment of the average peer judgment (Palley & Soll, 2019) hold promise for isolating knowledgeable contributors to an aggregation pool. For the diagnostic signal, much of the benefit we observed comes by virtue of coherence via certainty (e.g. probabilities of exactly 0 and 1), exploiting a similar principle. For the determinative signal, we might expect judges with no knowledge of Bayes’s theorem to simply opt out, and where opting out is not an option, elicitation procedures should attempt to distinguish coherence due to understanding (or at least implementing) coherence principles versus coherence that arrives incidentally via response biases such as straight-lining or responding .5 to convey one’s utter epistemic uncertainty. As we observed in our re-analyses of the Turtle Island datasets, failure to identify such cases and the conditions under which such cases are more likely could impose costs to accuracy. Information about response consistency (reliability) could also be leveraged to improve accuracy through repeated elicitation (Miller & Steyvers, 2017). For the diagnostic signal, reliability is a proxy for certainty and expertise. For the determinative signal, reliability is a proxy for the correct application of relevant probabilistic axioms and theories.

Importantly, the methods described here can also be combined with other techniques aimed at improving accuracy of aggregated estimates. For example, practitioners may combine coherence-weighted aggregation with other performance-weighted methods (Bolger & Rowe, 2015; Budescu & Chen, 2015; Himmelstein et al., 2021), use competitive elicitation methods (Lichtendahl et al., 2013), enhance the salience of private versus public information (Larrick et al., 2012), choose smaller, wiser crowds (Soll et al., 2010), trim opinion pools to account for under- and over-confidence (Yaniv, 1997; Jose et al., 2014), up-weight assessors who update esitmates frequently in small increments (Atanasov et al., 2020), or extremize the judgments (Baron et al., 2014; Hanea et al., 2021; Satopää & Ungar, 2015). The latter is particularly appealing in general-knowledge tasks where the "outcomes" are true or false, with associated maximally extreme probabilities of 1 and 0.

The Utility of Validation ι

The distinction between ι that diagnoses accuracy versus ι that determines accuracy has the benefit of explaining the validation ι results intuitively. Although validation ι has the potential to be useful in either case, our results indicate some differences between the two. To the extent that coherence determines accuracy on a task, it is not surprising that ι on that task would be a useful signal for correspondence on a closely related task. That is, if one has the knowledge and ability to derive coherent and accurate Bayesian probabilities on one Bayesian task (e.g., one involving the estimation of likelihoods), it will likely serve as a useful indicator of their ability to derive coherent and accurate probabilities on another Bayesian task (e.g., judging posterior probabilities). By comparison, where coherence diagnoses accuracy, it is not surprising that ι on one topic does not strongly predict correspondence on another topic. Rather, its utility may depend on two factors: (1) a clear and uncontaminated diagnostic signal (i.e., the spaced condition), and (2) how closely related the topic for which ι was measured is to the task at hand. That is, does knowledge in the calibration domain predict knowledge in the validation domain? Note that this implies that a topic-specific conceptualization of validation ι could be particularly effective, which is an issue that future research could profitably to investigate.

Choosing the 'Best' Function

An aim of the present research was to compare and contrast the various functions used in previous studies (Fan et al., 2019; Karvetski et al, 2013; Wang et al., 2011). The present findings clarify, perhaps unsurprisingly, that there is no single "best" method. Weighting functions, the tuning parameter β, and group size *k* interacted in complex ways that affected *MAE* and *PI*. Nevertheless, lessons from these findings can be drawn. First, note that the weighting function and β value work in tandem to determine how quickly (i.e., aggressively) an assessor's aggregation weight approaches 0 as ι increases. In much the same way that β = 100 is more aggressive than β = 10, the rank function is more aggressive than the standard linear difference function (which, in turn, is more aggressive than the exponential function). As we have seen in our re-analysis of Wu et al. (2017), the most aggressive weighting strategies are not always optimal. In fact, across both re-analyses we conducted, more aggressive strategies often resulted in lower *PI*. In this way, the exponential function can be seen as more flexible than either the rank or standard linear difference function given that its responsiveness to changes in β allow it to achieve a more modest weighting strategy if required by the practitioner.

Providing further evidence for this recommendation, we observed sub-optimal behavior in both the rank function and standard linear difference function in varying cases. In the statistical-evidence domain, *MAE* and *PI* started to worsen at lower values of β and *k* for the rank function than for either competing function. It is also possible for the function to award relatively high weights to very incoherent answers, potentially affecting the aggregate. For example, imagine a set of ι values, {0.00, 0.80, 0.82, 0.83}, submitted to each function. The first assessor is perfectly coherent, but the next three assessors are all quite bad. Both the exponential and standard linear difference functions would assign each of the individuals with ι ≥ .80 a justifiably low weight. By comparison, because the rank function only considers the relative rank order of ι, the second assessor with ι = 0.80 would still receive appreciable weighting during aggregation. Furthermore, the average of probabilities bound by non-linear constraints is not guaranteed to be coherent. Thus, the sometimes-deficient performance of the rank function is unsurprising.

Finally, regarding the standard linear difference function, across both domains and at low levels of β the function would achieve peak accuracy at low group sizes but worsen thereafter. This is because the function will always assign a weight of 0 to the most incoherent individual, reducing the effective size of the aggregate pool's judgments by 1. This is empirically obvious as the standard linear difference function is undefined at $k = 1$ and does not change as a function of β at $k = 2$. That the most incoherent judgment is ignored has a dramatic effect, particularly at low group sizes. By comparison, the exponential function had neither of these drawbacks, lacking the quirks described above that make the competing functions unpredictable.

Regarding group size, we find that a larger crowd is generally better, although gains in accuracy were associated with diminishing returns, similar in nature to other studies of aggregation (Han & Budescu, 2019; Mandel et al., 2018). Beyond a small-to-moderate group size of between 2 and 10 judges, the addition of a single extra participant did not improve accuracy in either domain. This is in line with research showing small-to-medium-sized groups are ideal during aggregation (Han & Budescu, 2019; Mannes et al., 2014; Navajas et al., 2018). That said, the results of Turtle Island show that large groups increased the risk of selecting at least one "false positive"—an individual who was incidentally coherent. These individuals have the potential to produce highly inaccurate estimates when coherence-weighted aggregation is applied. Indeed, this was precisely the reason that Karvetski et al. (2013) did not use the complementarity rule as a basis for coherentizing—they found that too many fifty-fifty blip judgments undermined the value of coherence weighting. Considering the risk factor and the cost of recruiting additional judges, we believe the present results are in line with conventional wisdom suggesting well-selected medium-sized groups often perform better than either small or large groups. To eliminate the deleterious effects of incidentally coherent assessors who show a fifty-fifty blip bias, practitioners could exploit methods for mitigating such bias (e.g., Bruine de Bruin et al., 2002). Simple procedures such as scanning for straight-liners and omitting their judgments could be used to improve aggregation methods.

Toward a Decision-Theoretic Framework for Optimizing Probability Judgment

The foregoing discussion provides a basis for sketching a decision-theoretic framework for optimizing probability judgments. Central to this framework is the view that decisions about optimization strategies must focus broadly on *ensembles* of relevant factors, which include, but which are not restricted to, the selection of: (a) format for information presentation, (b) methods for eliciting judgments, (c) methods for recalibrating or otherwise transforming judgments either before or after aggregation, and (d) methods for aggregating judgments (Karvetski et al., 2020). The alternative—narrowly considering the effect of one of these factors on its own—is unlikely to reveal important lessons for judgment accuracy optimization that are robust and generalizable. From a methodological perspective, a focus on ensembles entails a greater degree of complexity in experimental variables. We need to better understand how the pillars of optimization—information representation, judgment elicitation, and post-judgment recalibration and aggregation—interact amongst themselves and how those interactions are moderated by task characteristics (Zellner et al., 2021).

In the present research, we shed light on the relation between the two central criteria undergirding assessments of human performance in judgment and decision-making environments: coherence and correspondence. Our results show that there is indeed a relation between the two criteria and further suggests that variation in the strength of that relation can be anticipated within a decision-theoretic framework. By analyzing task characteristics and differentiating those tasks in which individual differences in coherence are likely to *diagnose* correspondence (i.e., accuracy) from those in which individual differences in coherence must have a *determinative* effect on correspondence, we can begin to identify which optimization ensembles for improving the accuracy of probability judgment will be effective and what the most likely impediments to their success will be.

Funding and Statement of Interest

This research is partially funded by Canadian Safety and Security Program project CSSP-2018-TI-2394 under the direction of David R. Mandel. Charley M. Wu is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC 2064/1 – 390727645.

The authors report no conflict of interest.

References

Atanasov, P., Witkowski, J, Ungar, L., Mellers, B., & Tetlock, P. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes, 160,* 19-35. https://doi.org/10.1016/j.obhdp.2020.02.001

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, *11*(2), 133–145. https://doi.org/10.1287/deca.2014.0293

Bell, D., Raiffa, H., & Tversky, A. (1988). Descriptive, normative, and prescriptive interactions in decision making. In D. Bell, H. Raiffa & A. Tversky (Eds.), *Decision making: Descriptive, normative, and prescriptive interactions* (pp. 9-30). Cambridge University Press.

Bennett, S. T., Benjamin, A. S., Mistry, P. K., & Steyvers, M. (2018). Making a wiser crowd: Benefits of individual metacognitive control on crowd performance. *Computational Brain & Behavior*, *1*(1), 90–99. https://doi.org/10.1007/s42113-018-0006-4

Bolger, F., & Rowe, G. (2015). The aggregation of expert judgment: Do good things come to those who weight? *Risk Analysis*, *35*(1), 5–11. https://doi.org/10.1111/risa.12272

Bruine de Bruin, W., Fischbeck, P. S., Stiber, N. A., & Fischhoff, B. (2002). What number is “fifty-fifty”?: Redistributing excessive 50% responses in elicited probabilities. *Risk Analysis, 22*(4), 713–723. https://doi.org/10.1111/0272-4332.00063.

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

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2020). Shiny: Web application framework for R. *R package version 1.4.0.2.* https://CRAN-R-project.org/package=shiny.

Clemen, R. T., & Winkler, R. L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*(2), 187–203. https://doi.org/10.1111/j.1539-6924.1999.tb00399.x

Collins, R. N. (2021, May 1). *Interactive Coherence Approximation Principle Tool*. Shiny Apps. https://collinsrn.shinyapps.io/CAP-TOOL/

Collins, R. N., Mandel, D. R., Karvetski, C. W., Nelson, J. D., & Wu, C. M. (2021, April 27). *The wisdom of the coherent: Improving correspondence with coherence-weighted aggregation*. Retrieved from osf.io/46fdm

Cooke, R. M. (2015). The aggregation of expert judgment: Do good things come to those who weight? *Risk Analysis*, *35*(1), 12–15. https://doi.org/10.1111/risa.12353

David, M. (2002). The correspondence theory of truth. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (retrieved from https://plato.stanford.edu/entries/truth-correspondence/). The Metaphysics Research Lab, Stanford University.

Davidson, D. (1986). A coherence theory of truth and knowledge. In E. LePore (Ed.), *Truth and Interpretation. Perspectives on the Philosophy of Donald Davidson* (pp. 307–319). Blackwell.

Dawson, N. V., & Gregory, F. (2009). Correspondence and coherence in science: A brief historical perspective. *Judgment and Decision Making*, *4*(2), 8.

De Finetti, B. (1990). *Theory of probability: a critical introductory treatment*. John Wiley & Sons.

Fan, Y., Budescu, D. V., Mandel, D., & Himmelstein, M. (2019). Improving accuracy by coherence weighting of direct and ratio probability judgments. *Decision Analysis*, *16*(3), 197–217. https://doi.org/10.1287/deca.2018.0388

Fischhoff, B., & Bruine de Bruin, W. (1999). Fifty-fifty = 50? *Journal of Behavioral Decision Making, 12*(2), 149–167. https://doi.org/10.1002/(SICI)1099-0771(199906)12:2<149::AID-BDM314>3.0.CO;2-J.

Hammond, K. R. (2000). Coherence and correspondence theories in judgment and decision making. In Connolly, T., Hammond, K., & Arkes, H. (Eds.) *Judgment and Decision Making: An Interdisciplinary Reader.* Second Ed. Cambridge University Press. pp. 53-65.

Han, Y., & Budescu, D. (2019). A universal method for evaluating the quality of aggregators. *Judgment and Decision Making*, *14*(4), 395–411.

Hanea, A. D. Wilkinson, D., McBride, M., Lyon, A., van Ravenzwaaij, D., Singleton Thorn, F., Gray, C., Mandel, D. R., Willcox, A., Gould, E., Smith, E., Mody, F., Bush, M., Fidler, F., Fraser, H., & Wintle, B. (2021). *Mathematically aggregating experts' predictions of possible futures*. Preprint available at MetaArXiv: https://doi.org/10.31222/osf.io/rxmh7

Hemming, V., Walshe, T. V., Hanea, A. M., Fidler, F., & Burgman, M. A. (2018). Eliciting improved quantitative judgements using the IDEA protocol: A case study in natural resource management. *PloS One*, *13*(6), e0198468. https://doi.org/10.1371/journal.pone.0198468

Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, *18*(10), 504–506. https://doi.org/10.1016/j.tics.2014.06.009

Himmelstein, M., Atanasov, P., & Budescu, D. V. (2021). Forecasting forecaster accuracy: Contributions of past performance and individual differences. *Judgment and Decision Making*, *16*(2), 40.

Jose, V. R. R., Grushka-Cockayne, Y., & Lichtendahl Jr., K. C. (2013). Trimmed opinion pools and the crowd's calibration problem. *Management Science, 60*(2), 463-475.

Karvetski, C. W., Mandel, D. R., & Irwin, D. (2020). Improving probability judgment in intelligence analysis: From structured analysis to statistical aggregation. *Risk Analysis*, *40*(5). https://doi.org/10.1111/risa.13443

Karvetski, C. W., Olson, K. C., Mandel, D. R., & Twardy, C. R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis*, *10*(4), 305–326. https://doi.org/10.1287/deca.2013.0279

Kleindorfer, P., Kunreuther, H., & Schoemaker, P. (1993). *Decision sciences: An integrative perspective*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139173537

Kolmogorov, A. (1956). *Foundations of the theory of probability*. New York: Chelsea Publishing Company.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127. https://doi.org/10.1287/mnsc.1050.0459

Lichtendahl, K. C., Grushka-Cockayne, Y., & Pfeifer, P. E. (2013). The wisdom of competitive crowds. *Operations Research*, *61*(6), 1383–1398. https://doi.org/10.1287/opre.2013.1213

Mandel, D. R. (2000). On the meaning and function of normative analysis: Conceptual blur in the rationality debate? *Behavioral and Brain Sciences, 23*(5),686-687.

Mandel, D. R. (2008). Violations of coherence in subjective probability: A representational and assessment processes account. *Cognition*, *106*(1), 130–156. https://doi.org/10.1016/j.cognition.2007.01.001

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01144

Mandel, D. R., Collins, R. N., Risko, E. F., & Fugelsang, J. A. (2020). Effect of confidence interval construction on judgment accuracy. *Judgment and Decision Making, 15*(5), 783-797.

Mandel, D. R., Karvetski, C. W., & Dhami, M. K. (2018). Boosting intelligence analysts' judgment accuracy: What works, what fails? *Judgment and Decision Making*, *13*(6), 607–621.

Mannes, A. E., Larrick, R. P., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers of social psychology. Social judgment and decision making* (pp. 227–242). Psychology Press.

Mannes, A. E., Larrick, R. P., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers of social psychology. Social judgment and decision making* (pp. 227–242). Psychology Press.

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

Miller, B., & Steyvers, M. (2017). Leveraging response consistency within individuals to improve group accuracy for rank-ordering problems. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Osherson, D., & Vardi, M. Y. (2006). Aggregating disparate estimates of chance. *Games and Economic Behavior*, *56*(1), 148–173. https://doi.org/10.1016/j.geb.2006.04.001

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

Patterson, D. (2003). What is a correspondence theory of truth? *Synthese*, *137*(3), 421–444. https://doi.org/10.1023/B:SYNT.0000004905.68653.b3

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature, 541*, 532-535.

Por, H., & Budescu, D. V. (2017). Eliciting subjective probabilities through pair-wise comparisons. *Journal of Behavioral Decision Making*, *30*(2), 181-196. https://doi.org/10.1002/bdm.1929

Predd, J. B., Osherson, D. N., Kulkarni, S. R., & Poor, H. V. (2008). Aggregating probabilistic forecasts from incoherent and abstaining experts. *Decision Analysis*, *5*(4), 177–189. https://doi.org/10.1287/deca.1080.0119

Satopää, V., & Ungar, L. (2015). Combining and extremizing real-valued forecasts. *http://arxiv.org/abs/1506.06405*

Soll, J. B., Larrick, R. P., & Mannes, A. E. (2010). When it comes to wisdom, smaller crowds are wiser, in M. C. Campbell, J. Inman, & R. Pieters (Eds)., *Advances in Consumer Research Volume 37* (pp. 94 – 97). Association for Consumer Research.

Speirs-Bridge, A., Fidler, F., McBride, M., Flander, L., Cumming, G., & Burgman, M. (2010). Reducing overconfidence in the interval judgments of experts. *Risk Analysis, 30*, 512–523. https://doi.org/10.1111/j.1539-6924.2009.01337.x

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

Tsai, J., & Kirlik, A. (2012). Coherence and correspondence competence: Implications for elicitation and aggregation of probabilistic forecasts of world events. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 313–317. https://doi.org/10.1177/1071181312561073

Tversky, A., & Koehler, D. J. (1994) Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*(4), 547-567. https://doi.org/10.1037/0033-

295X.101.4.547

Turlach, B. A. (S original); Weingessel, A. (R Port); Moler, C. (Fortran Contributions) (2019). Quadprog: Functions to solve quadratic programming problems. *R Package version 1.5-8*. retrieved from https://cran.r-project.org/web/packages/quadprog/quadprog.pdf.

Wang, G., Kulkarni, S., Poor, H. V., & Osherson, D. N. (2011). Improving aggregated forecasts of probability. *2011 45th Annual Conference on Information Sciences and Systems*, 1–5. https://doi.org/10.1109/CISS.2011.5766208

Weiss, C. (2008). Communicating uncertainty in intelligence and other professions. *International Journal of Intelligence and CounterIntelligence*, *21*(1), 57–85. https://doi.org/10.1080/08850600701649312

Weaver, E. A., & Stewart, T. R. (2012). Dimensions of judgment: Factor analysis of individual differences. *Journal of Behavioral Decision Making*, *25*(4), 402–413. https://doi.org/10.1002/bdm.748

Williams, J. J., & Mandel, D. R. (2007). Do evaluation frames improve the quality of conditional probability judgment? In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 1653-1658), Mahwah: Erlbaum.

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error () over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*, 79–82. https://doi.org/10.3354/cr030079

Wright, G., & Ayton, P. (1986a). Subjective confidence in forecasts: A response to Fischhoff and MacGregor. *Journal of Forecasting*, *5*(2), 117–123. https://doi.org/10.1002/for.3980050205

Wright, G., & Ayton, P. (1986b). The psychology of forecasting. *Futures*, *18*(3), 420–

439. https://doi.org/10.1016/0016-3287(86)90023-6

Wright, G., & Ayton, P. (1987). Task influences on judgemental forecasting. *Scandinavian Journal of Psychology*, *28*(2), 115–127. https://doi.org/10.1111/j.1467-9450.1987.tb00746.x

Wright, G., Rowe, G., Bolger, F., & Gammack, J. (1994). Coherence, calibration, and expertise in judgmental probability forecasting. *Organizational Behavioural Human Decision Processes, 57*(1), 1-25. https://doi.org/10.1006/obhd.1994.1001

Wright, G., Saunders, C., & Ayton, P. (1988). The consistency, coherence and calibration of holistic, decomposed and recomposed judgemental probability forecasts. *Journal of Forecasting*, *7*(3), 185–199. https://doi.org/10.1002/for.3980070304

Wu, C. M., Meder, B., Filimon, F., & Nelson, J. D. (2017). Asking better questions: How presentation formats influence information search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1274–1297. https://doi.org/10.1037/xlm0000374

Chen, X. & Yin, X. (2019). NlcOptim: Solve nonlinear pptimization with nonlinear constraints. *R package version 0.6*. retrieved from https://cran.r-project.org/web/packages/NlcOptim/NlcOptim.pdf.

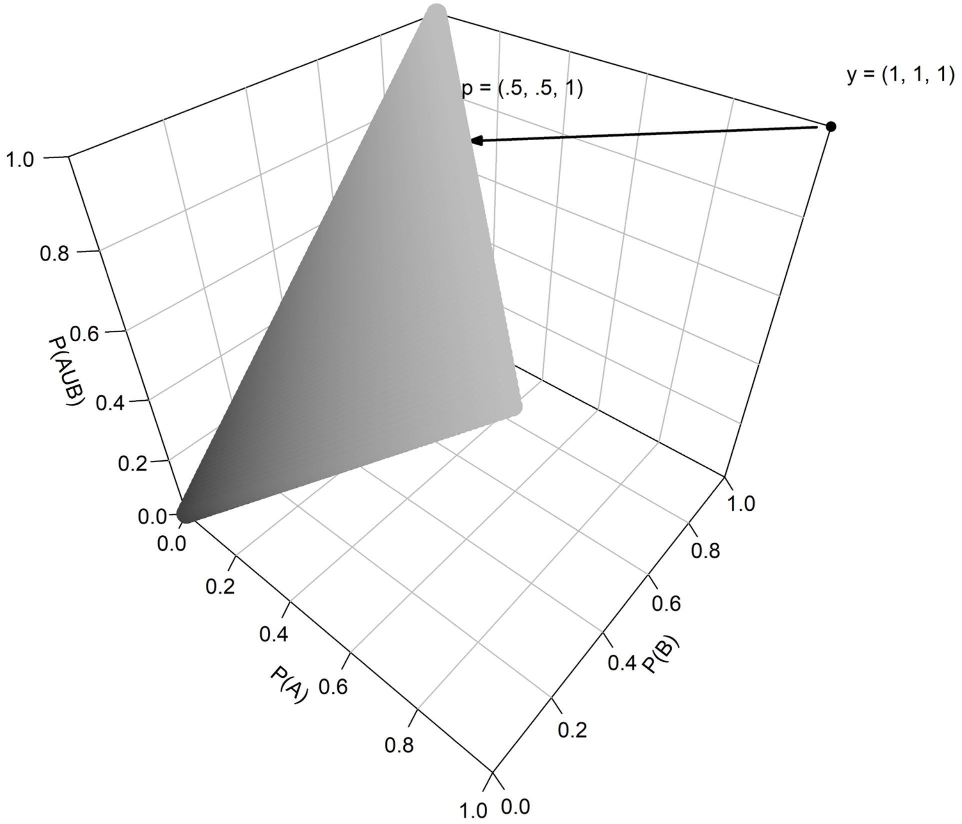
Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processed, 69*(3), 234-249.

Young, J. O. (2001). A defence of the coherence theory of truth. *Journal of Philosophical Research*, *26*(1), 89-101.

Zellner, M., Abbas, A. E., Budescu, D. V., & Galstyan, A. (2021). A survey of human judgement and quantitative forecasting methods. *Royal Society Open Science*, *8*(2), 201187. https://doi.org/10.1098/rsos.201187

Figure 1

The Coherence Approximation Principle Applied to P(*A*) + *P*(*B*) = *P*(*A* ∪ *B*)



The Coherence Approximation Principle (*Predd et al. 2008, Osherson and Vardi 2006*) *using a set of three probability estimates with the constraint P*(*A*) + *P*(*B*) = *P*(*A* ∪ *B*), *equivalent to the 3-way constraint enforced in Karvetski et al.* (*2013*). *The flat, triangular polygonal plane defines the possible combinations of coherent probabilities. It is flat and bounded by the vertices* [*(0, 0, 0), (0, 1, 1), (1, 0, 0)*]. *Interactive tool available from https://collinsrn.shinyapps.io/CAP-TOOL/*

Figure 2

Mean Absolute Error for Coherence-Weighted Aggregation in the General-knowledge Domain (Calibration Metric)

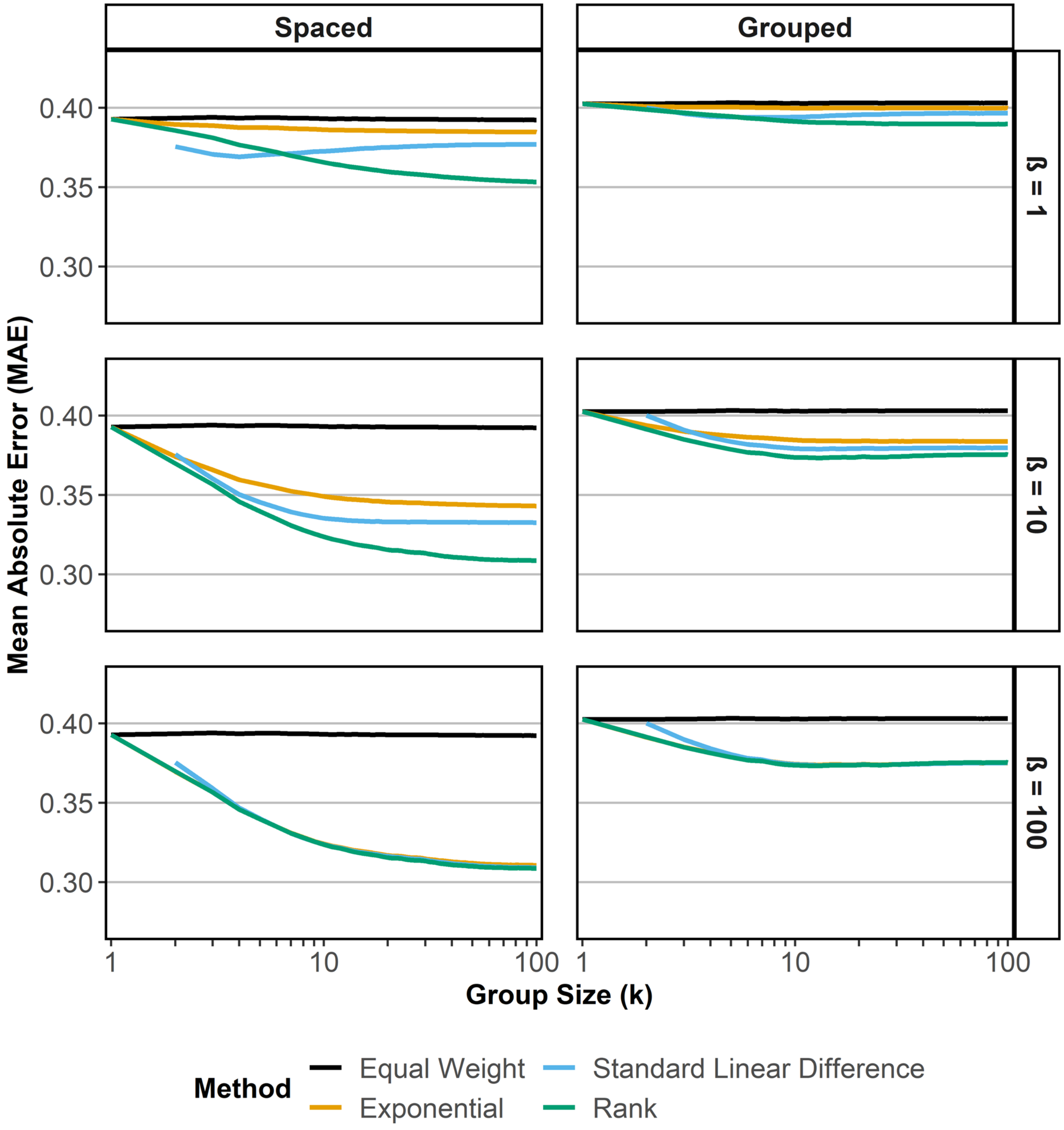


Figure 3

Proportion Improved for Coherence-Weighted Aggregation in the General-knowledge Domain (*Calibration Metric*)

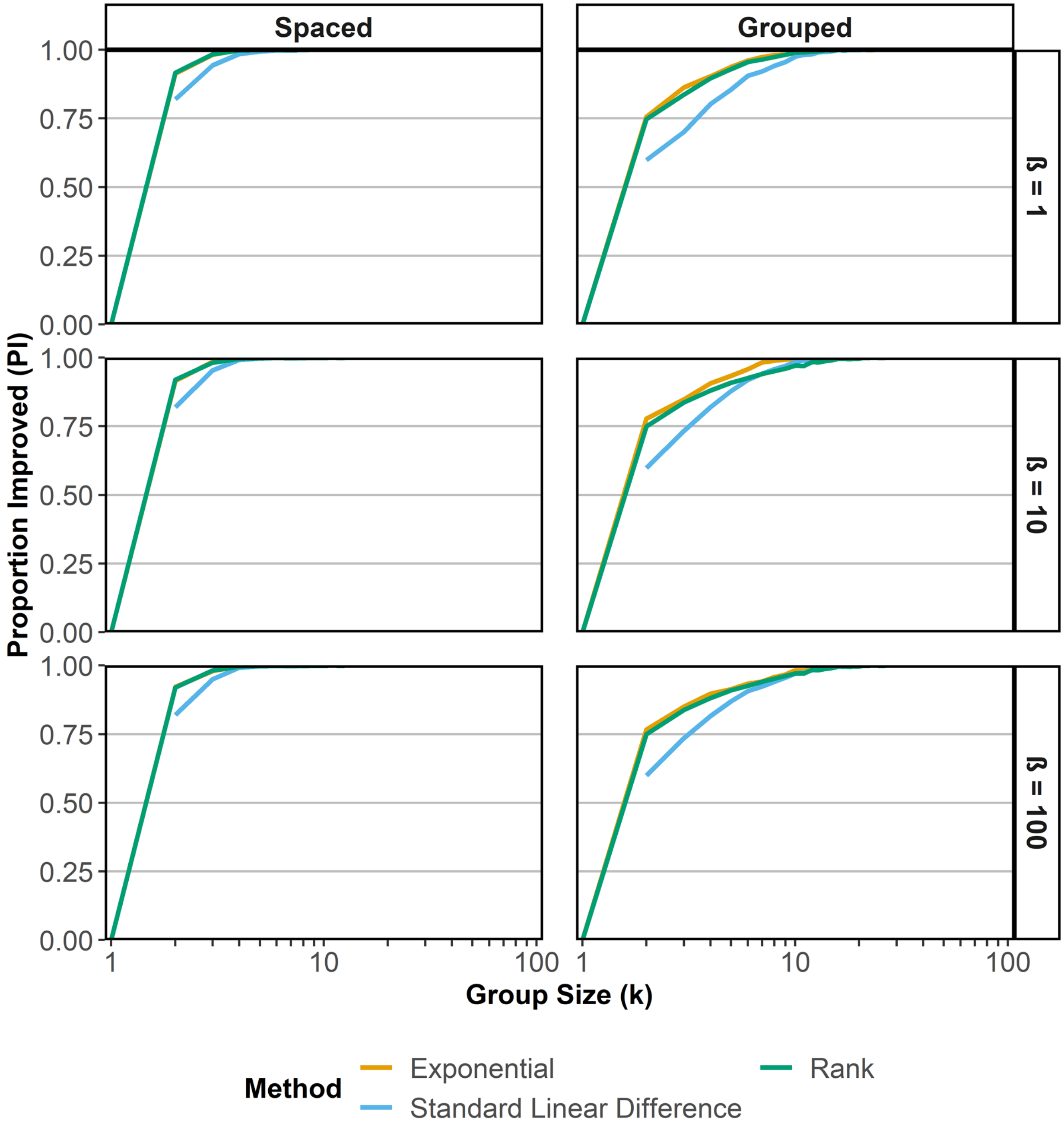


Figure 4

Mean Absolute Error for Coherence-Weighted Aggregation in the General-knowledge Domain (Validation Metric)

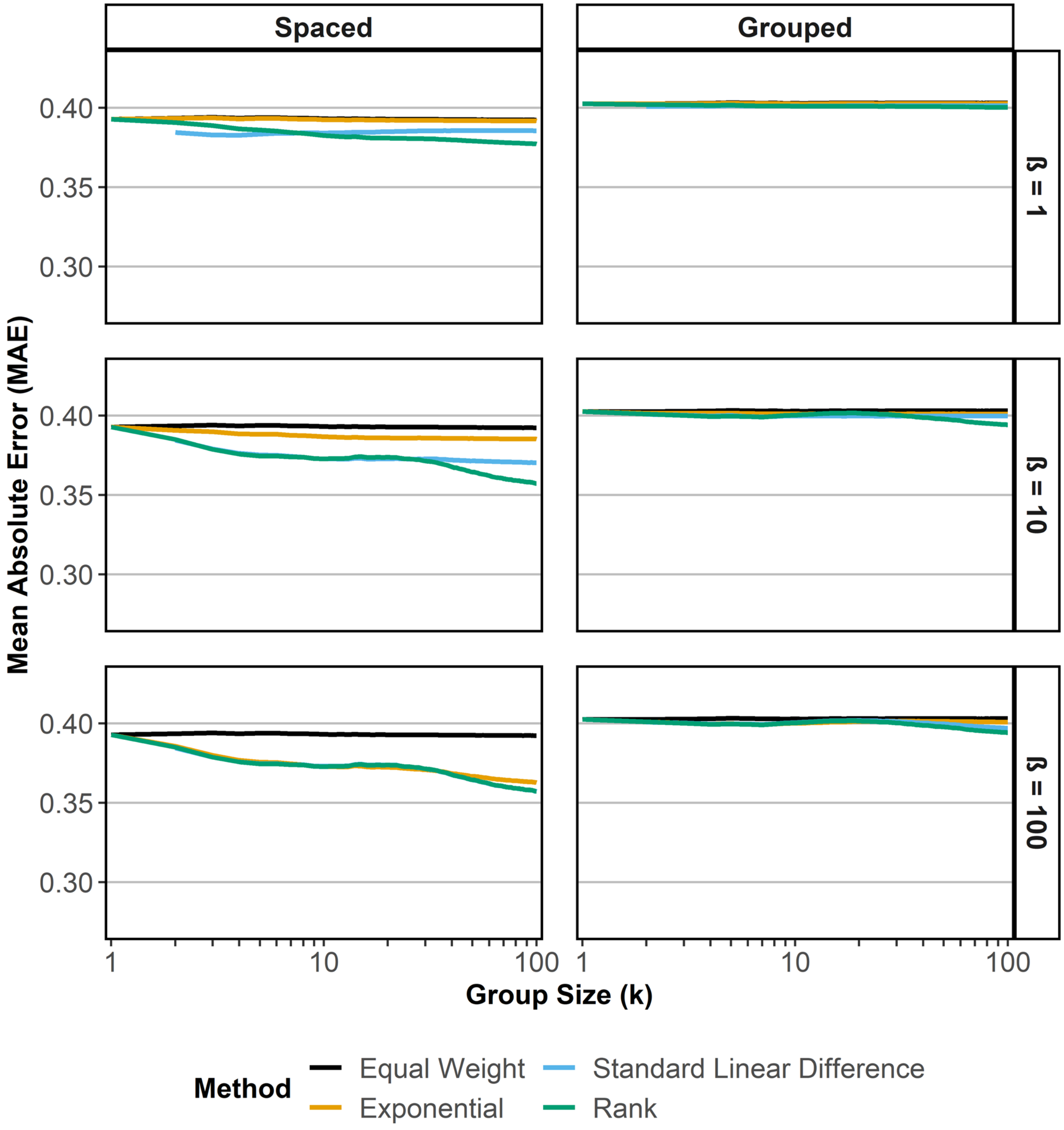


Figure 5

Proportion Improved for Coherence-Weighted Aggregation in the General-knowledge Domain (Validation Metric)

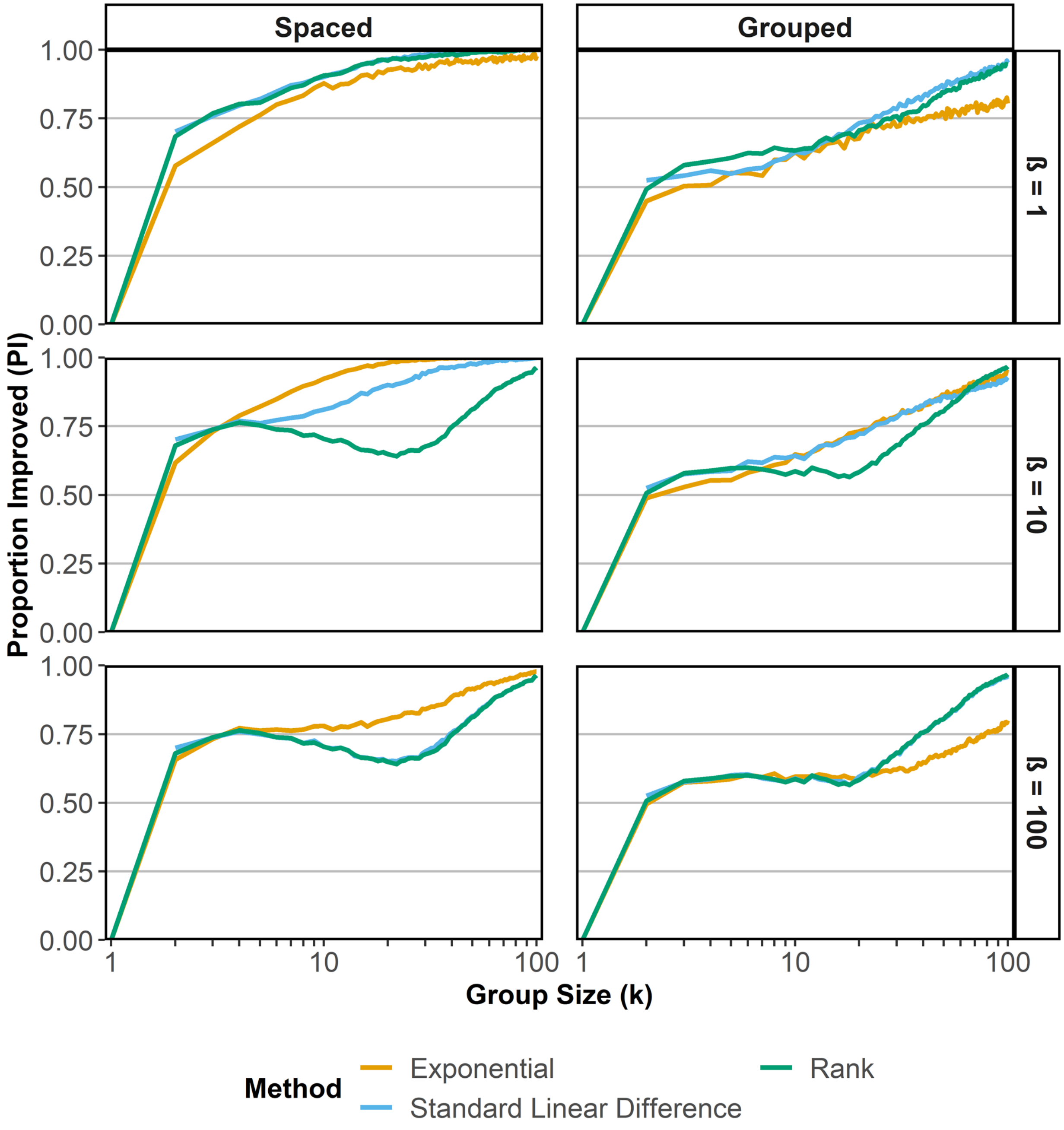


Figure 6

Mean Absolute Error for Coherence-Weighted Aggregation in the Statistical-evidence Domain

(*Calibration Metric*)

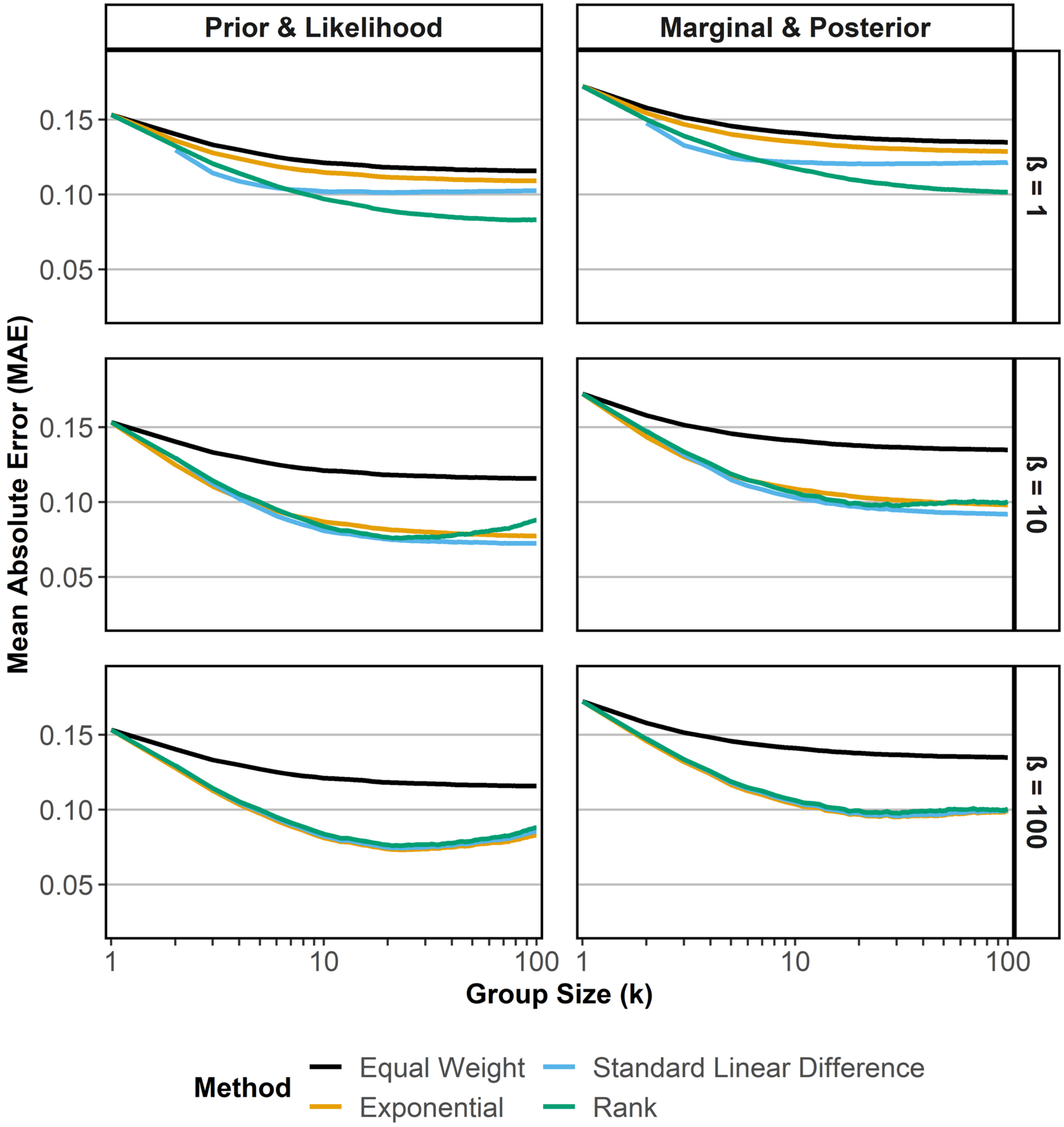


Figure 7

Proportion Improved for Coherence-Weighted Aggregation in the Statistical-evidence Domain (*Calibration Metric*)

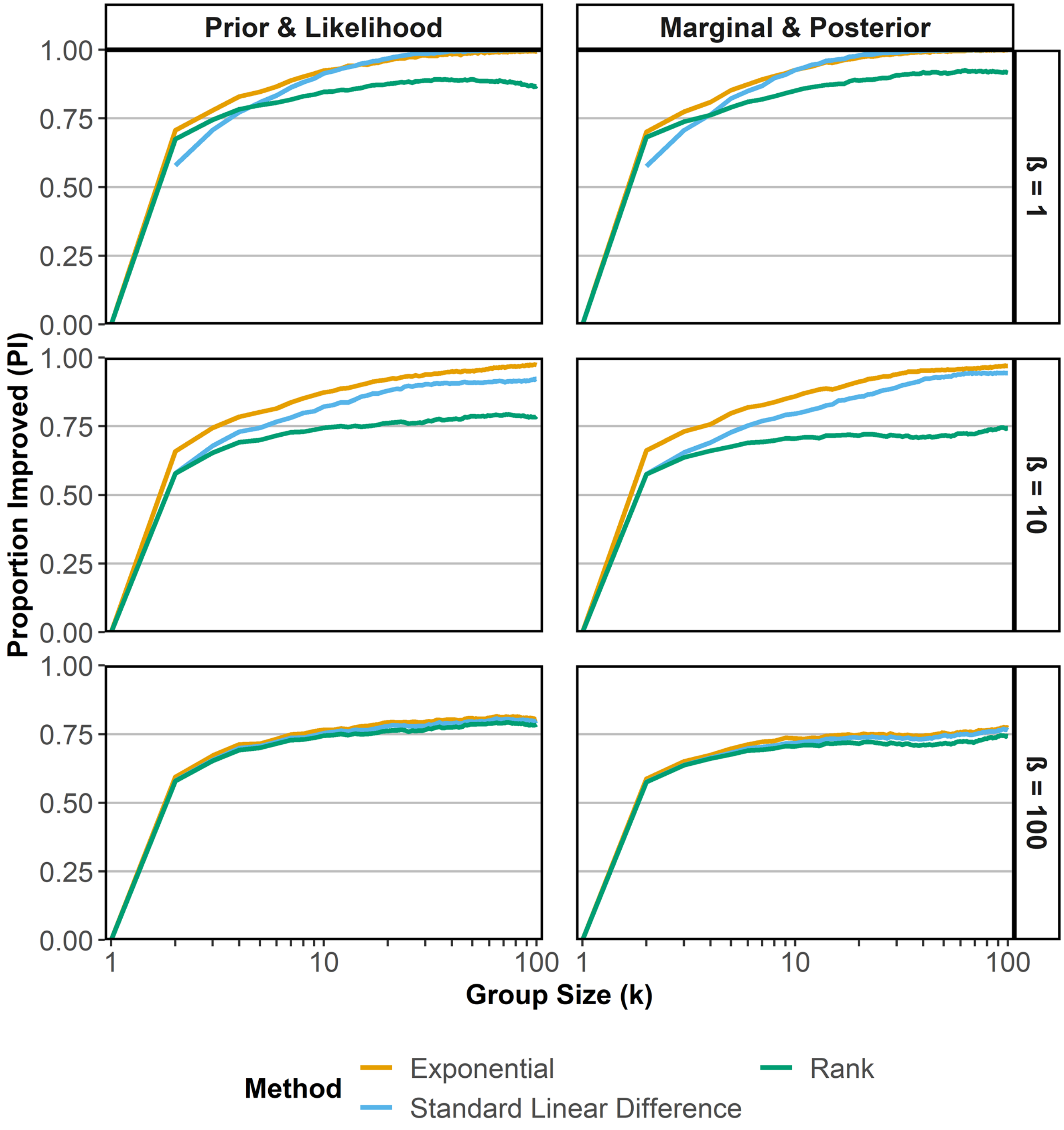
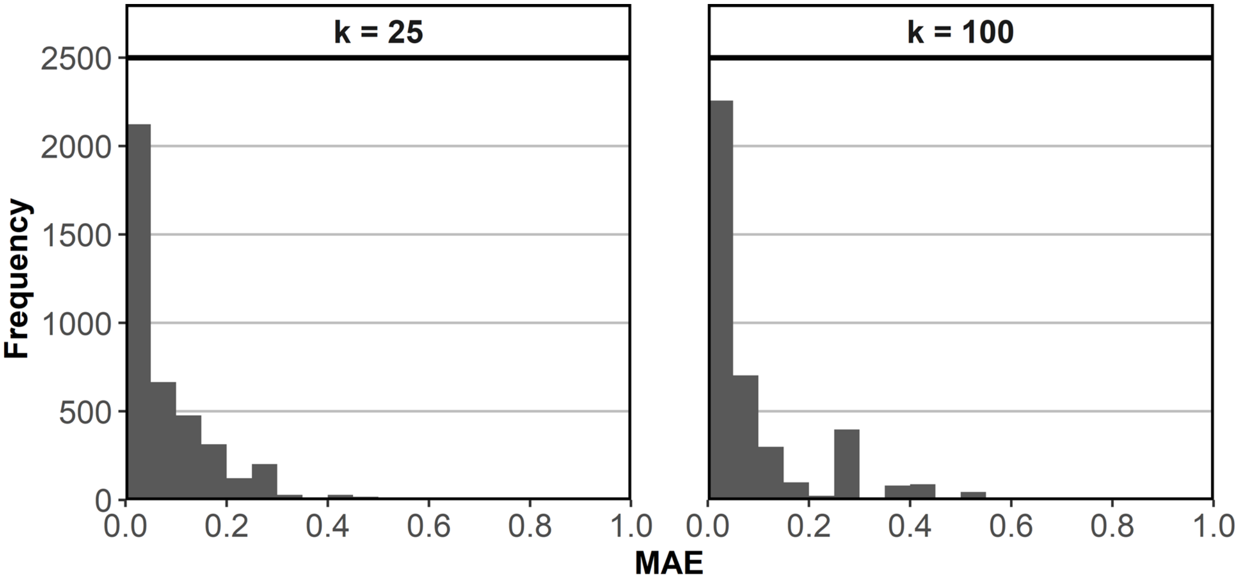


Figure 8

Histogram of Bootstrapped MAE at β = 100 in the Prior and Likelihood (PL) Condition



Note: *4000 bootstraps total*

Figure 9

Mean Absolute Error for Coherence-Weighted Aggregation in the Statistical-evidence Domain (Validation Metric)

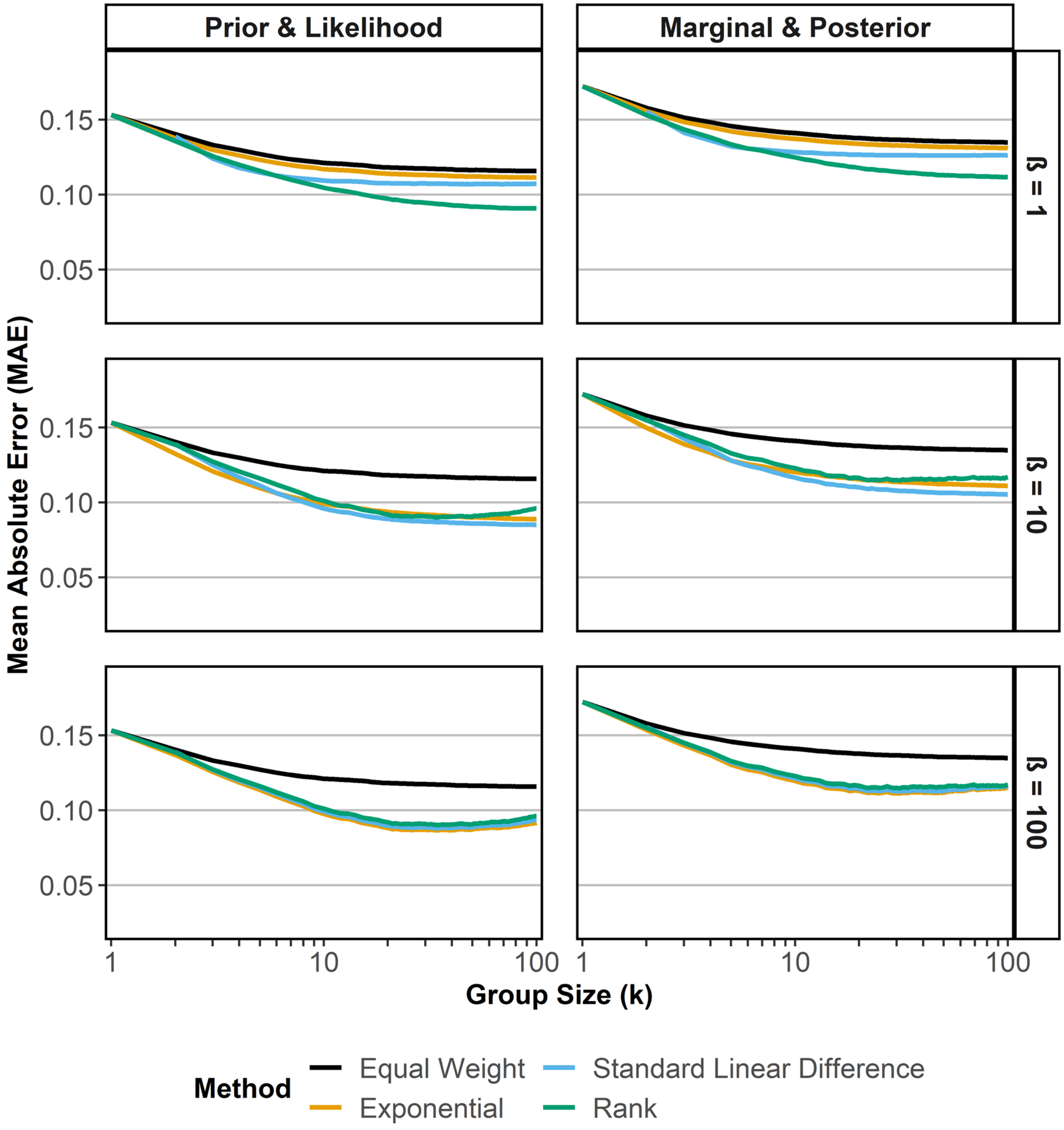


Figure 10

Proportion Improved for Coherence-Weighted Aggregation in the Statistical-evidence Domain (Validation Metric)

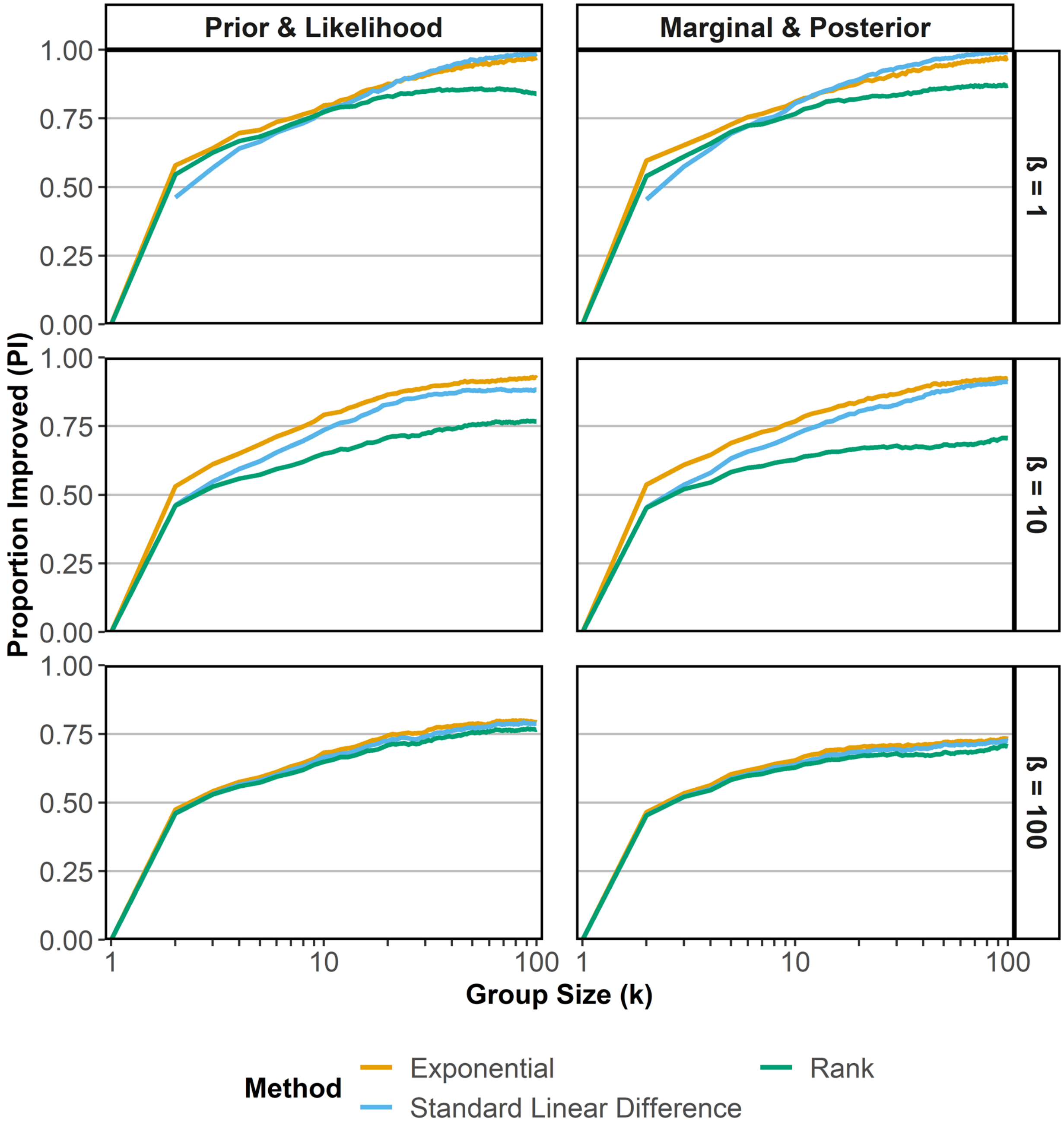


Table 1

List of Functions for Translating the Incoherence Metric ɩ into an Aggregation Weight ω

Weight Function	Formula
Exponential (Wang et al., 2011)	$\omega_i = e^{(-\iota_i \times \beta)}$
Standardized Linear Difference (Karvetski et al., 2013)	$\omega_i = \left(\frac{max\,(\iota) - \iota_i}{max\,(\iota)}\right)^{\beta}$
Rank (Fan et al., 2019)	$\omega_i = \left(\frac{1}{rank(\iota_i)}\right)^{\beta}$

APPENDIX A. *List of constraints on individuals' Bayesian probability estimates*

Cond	Given	Derive	Constraints
PL	*Prior* $P(B)$ *Likelihood* $P(E\mid B)$, $P(E\mid F)$, $P(L\mid B)$, $P(L\mid F)$	*Marginal* $P(E)$, $P(L)$	*Non-Negativity/Bounds* (*Inequality*) $0 \le \{P(B), P(E), P(L), P(E\mid B), P(E\mid F), P(L\mid B), P(L\mid F)\} \le 1$ *Non-Linear/Bayesian* (*Equality*) $0 = P(E) - \{P(B) \times P(E\mid B) + [1 - P(B)] \times P(E\mid F)\}$ $0 = P(L) - \{P(B) \times P(L\mid B) + [1 - P(B)] \times P(L\mid F)\}$
		Posterior $P(B\mid E)$, $P(B\mid D)$, $P(B\mid L)$, $P(B\mid M)$	*Non-Negativity/Bounds* (*Inequality*) $0 \le \{P(B), P(E\mid B), P(E\mid F), P(L\mid B), P(L\mid F), P(B\mid E), P(B\mid D), P(B\mid L), P(B\mid M)\} \le 1$ *Non-Linear/Bayesian* (*Equality*) $0 = P(B\mid E) - \{P(B) \times P(E\mid B)\} / \{[P(B) \times P(E\mid B) + [1 - P(B)] \times P(E\mid F)\}$ $0 = P(B\mid D) - \{P(B) \times [1 - P(E\mid B)]\} / \{[P(B) \times [1 - P(E\mid B)] + [1 - P(B)] \times [1 - P(E\mid F)]\}$ $0 = P(B\mid L) - \{P(B) \times P(L\mid B)\} / \{[P(B) \times P(L\mid B) + [1 - P(B)] \times P(L\mid F)\}$ $0 = P(B\mid M) - \{P(B) \times [1 - P(L\mid B)]\} / \{[P(B) \times [1 - P(L\mid B)] + [1 - P(B)] \times [1 - P(L\mid F)]\}$
MP	*Marginal* $P(E)$, $P(L)$ *Posterior* $P(B\mid E)$, $P(B\mid D)$, $P(B\mid L)$, $P(B\mid M)$	*Prior* $P(B)$	*Non-Negativity/Bounds* (*Inequality*) $0 \le \{P(B), P(E), P(L), P(B\mid E), P(B\mid D), P(B\mid L), P(B\mid M)\} \le 1$ *Non-Linear/Bayesian* (*Equality*) $0 = P(B) - \{P(E) \times P(B\mid E) + [1 - P(E)] \times P(B\mid D)\}$ $0 = P(B) - \{P(L) \times P(B\mid L) + [1 - P(L)] \times P(B\mid M)\}$
		Likelihood $P(E\mid B)$, $P(E\mid F)$, $P(L\mid B)$, $P(L\mid F)$	*Non-Negativity/Bounds* (*Inequality*) $0 \le \{P(E), P(L), P(E\mid B), P(E\mid F), P(L\mid B), P(L\mid F), P(B\mid E), P(B\mid D), P(B\mid L), P(B\mid M)\} \le 1$ *Non-Linear/Bayesian* (*Equality*) $0 = P(E\mid B) \times \{P(E) \times P(B\mid E) + [1 - P(E)] \times P(B\mid D)\} - P(E) \times P(B\mid E)$ $0 = P(E\mid F) \times \{P(E) \times [1 - P(B\mid E)] + [1 - P(E)] \times [1 - P(B\mid D)]\} - P(E) \times [1 - P(B\mid E)]$ $0 = P(L\mid B) \times \{P(L) \times P(B\mid L) + [1 - P(L)] \times P(B\mid M)\} - P(L) \times P(B\mid L)$ $0 = P(L\mid F) \times \{P(L) \times [1 - P(B\mid L)] + [1 - P(L)] \times [1 - P(B\mid L)]\} - P(L) \times [1 - P(B\mid L)]$