PREPRINT CURRENTLY IN PRESS FOR PUBLICATION IN DECISION

The Wisdom of the Coherent:

Improving Correspondence with Coherence-Weighted Aggregation

Robert N. Collins [1], David R. Mandel [1], Chris W. Karvetski [2],

Charley M. Wu [3], and Jonathan D. Nelson [4, 5]

[1] Intelligence, Influence, and Collaboration Section, Defence Research and Development Canada

[2] Good Judgment Inc

[3] University of Tübingen

[4] Max Planck Institute for Human Development

[5] 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

This research is partially funded by the 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.

We have no known conflict of interest to disclose.

Address correspondence concerning this article 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 six experiments in two earlier studies (N = 58, N = 2,858) 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 bias 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 *causal 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 biased forms, sometimes imposes costs to accuracy. We conclude by sketching a decision-theoretic approach to how aggregators can sensibly leverage the *wisdom of the coherent* within the crowd.

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

1. Introduction

A theoretical divide between coherence and correspondence theorists has long defined decision science. Both camps share an interest in the descriptive, normative, and prescriptive quality of judgment (Bell et al., 1988; Hammond, 2000; Kleindorfer et al., 1993; Mandel, 2000). However, coherence and correspondence theorists approach these issues differently (Dawson & Gregory, 2009; Dunwoody, 2009). Coherence theorists (Davidson, 1986; Young, 1996) focus on the internal consistency of judgments; correspondence theorists (David, 2002; Patterson, 2003) focus on the empirical accuracy of judgments. Given their differing interests, few studies have examined the theoretical and empirical connections between individual differences in judgment coherence and correspondence.

Indeed, early studies found no consistent correlation between coherence and correspondence in several judgment tasks for differing levels of expertise (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 predict correspondence. Such findings suggest that the divergent histories of the coherence and correspondence literature 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 theoretical paths?

We believe this conclusion is premature. Wright and colleagues' experiments were statistically underpowered to detect correlations at conventional error rates (i.e., Type 1 = 5%, Type 2 = 20%). Furthermore, the correspondence score was a mean-squared-error function, while the coherence score used mean errors. Thus, the correspondence measure summed opposing errors (e.g., over- and under-estimation), while the coherence measure canceled opposing errors (e.g., super- and sub-additivity). These differences obscure correlations among measures of coherence and correspondence. More recent studies have indeed found correlations between coherence and correspondence (Weaver & Stewart, 2012; Weiss et al., 2009), directly challenging the hypothesis that the two are unrelated. Research has found correlations between

coherence and correspondence in predicting the winner of the 2011 Major League Baseball series (Tsai & Kirlik, 2012). Furthermore, superforecasters—elite forecasters who scored in the top 2% of accuracy rankings in a large geopolitical forecasting tournament—perform better than other forecasters on measures of logical coherence (Mellers et al., 2017).

1.1. Coherence-Based Recalibration and Aggregation

Going beyond research that examines correlations, research shows decision-makers can exploit coherence to improve correspondence. 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 use coherence as an aggregation weight (Mannes et al., 2014; see Collins et al., in press for review). Predd et al. (2008) showed that *coherence-weighted aggregation* improved group forecast accuracy on sports and economic forecasts. Studies have since generalized the method to US presidential election forecasts (Wang et al., 2011), 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). The findings show that coherence and correspondence may, in fact, be strongly related. More importantly, decision-makers can exploit knowledge of the former to improve the latter.

The fact that individual differences in coherence can predict correspondence is of prescriptive theoretical interest. Performance-weighted aggregation methods typically require that judges complete an additional task or that decision-makers keep records of the judges' past performance (Cooke, 2015). By comparison, coherence weighting requires neither; practitioners can apply the strategy as long as the elicitation contains a minimum of two logically related judgments (Predd et al., 2009). The number of elicitations is comparable to popular elicitation methods such as the construction of probability intervals (e.g., Mandel et al., 2020; 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*—namely, the unweighted average of groups of

judges—we propose that aggregators can achieve better performance by leveraging *the wisdom of the coherent* within the crowd.

1.2. The Present Research

The present work has several aims. One is to systematically compare the coherence-weighted aggregation methods used in earlier 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 two types of tasks where expert judgment is commonly sought: factual queries and statistical prediction. We examined whether the optimal coherence-weighting methods are stable or task-dependent (Zellner et al., 2021). A second aim is to determine if individuals' coherence on one or more tasks predicts correspondence on a different task. A third aim is to extend coherence-weighted aggregation methods to judgments involving conditional probabilities. Whereas earlier studies focused on linear additivity constraints (e.g., Predd et al., 2008; Wang et al., 2011), we applied coherence-based recalibration and aggregation methods to judgments characterized by nonlinear constraints typical of conditional probability estimation. Finally, we attempt to develop a theory of why (and under what circumstances) coherence weighting can be effective.

To pursue these aims, we reanalyzed data from Karvetski et al. (2013) and 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 based on statistical background information, characteristic of statistical Bayesian tasks (Mandel, 2014). These represent two broad kinds of tasks where expert judgment is often called upon: factual queries and conditional statistical predictions. We show that coherence-weighting is effective for both task domains. However, we discover that elicitation methods that make coherence trivial, by difficulty or by accident, reduce its efficacy. Consequently, those tasked with optimizing judgments must consider these issues when determining where and how to elicit judgments and apply coherence weighing. Finally, we show that coherence on one task has limited ability to predict

correspondence on other tasks.

The remainder of the article is laid out as follows: In Section 2, we formally define probabilistic coherence, how to quantify incoherence, how to apply this metric in performance-weighted aggregation, and how we define and measure correspondence. In Section 3, we apply coherence-weighted aggregation to a general-knowledge task (Karvetski et al., 2013). In Section 4, we apply coherence-weighted aggregation to a statistical-evidence task (Wu et al., 2017). Finally, in Section 5, we conclude with a General Discussion, including the reasons coherence-weighting is effective, how it may be profitably exploited, and things to consider when choosing to utilize the method.

2. 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), violating the axioms of probability calculus. Kolmogorov (1933; see also De Finetti, 1937) describes three relevant 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 one. Additivity states that any countable sequence of mutually exclusive events $E_1, E_2, \ldots, E_n$ must satisfy the condition:

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

We can define probabilistic coherence as the extent to which sets of probabilistic judgments respect Kolmogorov's probability constraints. Consequently, we can measure incoherence as the degree of such violations.

2.1. Quantifying Incoherence: The CAP

Researchers measure incoherence using the *coherence approximation principle* (CAP;

Osherson & Vardi, 2006; Predd et al. 2008). The CAP is an optimization algorithm that takes elicited probability sets and returns: (1) a recalibrated (or coherentized) set of probabilities, and (2) a measure of the Euclidean distance between the elicited and recalibrated probabilities, ι. For judgments bound by linear constraints (i.e., addition and subtraction), the recalibrated set of probabilities is provably unique (De Finetti, 1990; Karvetski et al., 2013) and guarantees monotonic improvements in accuracy (Predd et al., 2009). More precisely, the CAP is a constrained optimization problem focused on minimizing the Euclidean distance formula between the elicited and coherentized probabilities, expressed as:

$$\iota = \sqrt{\sum_{i=1}^{n} (p_i - y_i)^2} \qquad [2]$$

for n judgments, where p is the original elicited probability judgment and y is its coherentized equivalent, subject to the relevant constraints. The incoherence metric, ι, is, therefore, an objective measure of individuals' probabilistic coherence. This measure 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. The process for recalibration and calculating ι is 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 reason not to apply coherence-weighted aggregation to the (improved) recalibrated judgments. Second, by first recalibrating judgments to remove errors due to incoherence, we isolate improvements in aggregated judgment correspondence due exclusively to weighting by individual differences in coherence, separate from the improvements converging on a coherent judgment.

2.2. 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 converts increasing values of ι into a monotonically decreasing weight ω with differing properties. The exponential function (Wang et al., 2011) converts ι using an exponential function with Euler's number, *e*, as its base. The function is a simple mapping rule, as the conversion is insensitive to the distribution of ι in the opinion pool. The standardized linear difference function (Karvetski et al., 2013) calculates the relative difference between an individuals' ι and the maximum ι within the opinion pool. The function is sensitive to the distribution but not the size of the opinion pool; regardless of the pool size, weights are calculated relative to the most coherent judge. Because the 'worst' judge always receives a weight of ω = 0, it is undefined at a group size of 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 its variability; i.e., the second-best judge will receive a rank of 2 regardless of how much worse they are relative to rank 1.

Two functions, the exponential function (Wang et al., 2011) and the standardized linear difference function (Karvetski et al., 2013), include a practitioner-controlled parameter. This tuning parameter used the symbol λ in Wang et al. (2011) and β in Karvetski et al. (2013). Both 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), raising the function to exponent β. Regardless of function, β = 0 resolves to ω = 1 for every participant (i.e., equal-weighted aggregation). As β increases, the penalty for coherence violations increases, and the relative bias toward coherent responders increases. Sufficiently large values of β will reduce the contribution weight of all but the most coherent individual(s) to

zero. Simply put, β tunes the bias of the weighting function.

Regardless of the translation function, coherence-weighted aggregation proceeds identically. For each coherent probabilistic judgment y and by n individuals for each ith judgment, we define an aggregated judgment as:

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

In the present research, we consider three values for $\beta \in \{1, 10, 100\}$. The decision to use discrete values rather than continuous values was pragmatic rather than theoretically driven. Technically, any value $\beta > 0$ is a valid choice, with increasingly larger values behaving like a step function that assigns $\omega = 1$ to the most coherent judge and $\omega = 0$ to all others. Regardless, an important property to note is that setting $\beta = 0$ results in $\omega = 1$ for all participants. Thus, [3] simplifies to the equal-weighted average given by the equation

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

2.3. Correspondence Metric: MAE

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

$$MAE = \frac{1}{k} \sum_{i=1}^{k} |y_i - x_i|, \quad [5]$$

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 root mean square errors or Brier scores are popular probabilistic judgment error metrics, it is important to note there is no "correct" scoring rule. Rather, the scoring rule should be determined by the needs of the decision-maker, namely, whether they wish to penalize large mistakes more than small mistakes. Furthermore, Willmott and Matsuura (2005) propose that *MAE* is the superior measure for comparing average model performance. In our case, the use of *MAE* also adds information to the reanalysis of Karvetski et al. (2013), originally scored using the Brier score (i.e., mean squared error).

3. 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, fifty-eight 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 the 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, $P(A \cup B)$. For example: A - Neil Armstrong was the first man to set foot on the moon; B - Buzz Aldrin was the first man to set foot on the moon; A^c - Neil Armstrong was NOT the first man to set foot on the moon; and $A \cup B$ - Either Neil Armstrong or Buzz Aldrin was the first man to set foot on the moon. Karvetski et al. (2013) evaluated several variants of coherentization and methods for calculating ι, the most effective of which was based on the additivity constraint $P(A) + P(B) = P(A \cup B)$, which we use here also. In Experiment 1, participants rated one randomly selected judgment in each topic, cycling through all other topics before rating the next statement in the topic set. In Experiment 2, participants rated the statements for a given topic consecutively. We refer to these two experiments as the spaced and grouped conditions, respectively.

3.1. Method

All data and *R* scripts described below are available on the Open Science Framework (OSF; Collins et al., 2021).

3.1.1. Incoherence Metrics

We coherentized probabilities and calculated ι via the CAP [2] using the quadprog package in *R Studio* (Turlach, 2019) running the *R* programming language. We calculated two different ι measures for use in the aggregation model. The first, *endogenous* ι, is the measured incoherence of a judge on the target judgment. The second, *exogenous* ι, is the average

incoherence for all judgements other than the target judgment. 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. We used this metric to evaluate whether ι has potential use in disposition- and history-based performance-weighting methodologies that weight judges according to past performance or expertise (Cooke et al., 1988; Budescu & Chen, 2015; Mellers et al., 2015) per one of the primary aims of the study.

3.1.2. Correspondence Metrics

We calculated the *MAE* [5] across judgments {$P(A)$, $P(B)$, $P(A \cup B)$} for each of the 60 questions. This error serves as the primary dependent variable of interest. The difference between the *MAE* for coherence-weighted strategies and *MAE* for equal-weighted strategies will therefore represent the mean improvement in *MAE* across the 1,000 bootstrapped trials. We also examine the proportion of bootstrapped samples (out of 1,000, 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. Although ties are exceedingly rare, we treat them as losses due to the computational burden posed by coherence-weighting. We call this measure of performance *proportion improved* (*PI*).

3.1.3. Simulation and Aggregation Method

One of our research aims was to determine the implementations of coherence weighting that work best. To compare the efficacy of our various strategies, we used bootstrap sampling with replacement technique. Using bootstrap sampling with replacement allows us to examine the effect of aggregation at arbitrary group sizes not limited by the original sample size (30 in Experiment 1, 28 in Experiment 2). We analyzed different combinations of weighting function, parameterization, and group size. For each experimental condition (spaced vs. grouped), and for

each bootstrap, we added a randomly selected participant to the opinion pool one at a time until we achieved the maximum group size, k = 100. Each time we added a participant to the opinion pool, we calculated aggregated *MAE* across the 60 questions for each of our different aggregation strategies. This corresponds to a 2 (Condition: spaced, grouped) × 100 (Group Size: k = 1-100) × 3 (Weighting Function: exponential, standardized linear difference, rank) × 3 (Parameter: β = {1, 10, 100}) × 2 (Item Sample: endogenous, exogenous) design. We also include an equal-weighted function as a baseline comparison; however, this was not crossed with the parameter and item sample manipulation. We completed 1,000 bootstrap simulations for each condition. This approach is conceptually like a *multiverse* analysis (Steegen et al., 2016; Harder, 2020), in that we can compare different ι weights, weighting functions, levels of β, and group size while holding all else equal. We use this multiverse approach to develop decision-theoretic recommendations for ideal coherence-weighting strategies. The final analysis included 3,800,000 data points. We collapsed across the 1,000 bootstrap samples, producing 3,800 unique *MAE* values (1,900 per condition) as well as 3,600 unique *PI* values (1,800 per condition).

3.1.4. Analysis Plan

Due to the number of bootstraps, the standard errors (*SE*) were extremely low in all cases. In over 99% of simulations *SE* < .001. The confidence intervals were invisible on figures at conventional resolutions. The low *SE* means that virtually any absolute differences between methods, parameters, and group sizes will be statistically significant. Combined with the sheer number of comparisons, and the fact we had no *a priori* theories about specific comparisons of interest, we relied on visual inferences from plotted figures. Nevertheless, we provide full statistical and distributional data for specific formal comparisons in the data files on the Open Science Framework (OSF). To determine the significance of our *PI* measure, we compared this metric to the 99% confidence interval for a binomial distribution with 1,000 events (*lower*

bound = .459, upper bound = .541). A *PI* higher than the upper bound indicates greater than chance levels of performance improvements. Conversely, a *PI* lower than the lower bound indicates greater than chance levels of performance decrements. We suggest that this is the appropriate benchmark, as even small, reliable gains in aggregated accuracy have the potential to be highly consequential. The *PI* metric is like the probability of superiority (Grissom & Kim, 2005; Mandel et al., 2018; Ruscio & Mullen, 2012; Vargha & Delaney, 2000), and provides a measure of the reliability and consistency of accuracy improvements by employing coherence weighting.

3.2. Results

Figures 1-4 show the results of the coherence-weighted aggregation for *MAE* and *PI* for each of endogenous ι and exogenous ι, respectively, as a function of *k* and β. We scaled group size *k* on the x-axis logarithmically (base 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, the error of the equal-weighted judgment is strictly equivalent to the average error of the judges (Larrick & Soll, 2006). In other words, equal-weighted aggregation is ineffective at reducing *MAE* for any value of *k*. This contrasts with *MSE*, which decreases as *k* increases due to noise reduction (see Karvetski et al., 2013). Thus, *MAE* allows us to focus on improving the signal and reducing the distance from the resolution values.

3.2.1. Endogenous ι

3.2.1.1. MAE. The *MAE* results for endogenous **ι** are shown in Figure 1. 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 improvements afforded by an additional judge diminished as k increased, such that the addition of new assessors had a negligible effect for $k > 10$. Nevertheless, accuracy continued to improve past these thresholds, albeit at a much lower rate per additional judge. An exception to this was the standardized 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. The largest improvements occurred between $\beta = 1$ and $\beta = 10$. Third, comparing our different methods, the standardized 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 standardized linear difference and exponential function. Notably, all *MAE* functions converged as both k and β increased such that performance is nearly identical for all conditions where $k > 4$ and $\beta = 100$.

3.2.1.2. PI. The *PI* results for endogenous ι are shown in Figure 2. Each coherence-weighted aggregation method dramatically improved upon the equal-weighted *MAE* most times 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, and 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 standardized linear difference function performed worse at lower group-size values, while the exponential and rank functions perform similarly.

3.2.2. Exogenous ι

3.2.2.1. MAE. Figure 3 shows *MAE* results when using exogenous ι. For the spaced condition, there were three important findings. First, as with endogenous 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 endogenous ι, the standardized linear difference function performed best at $\beta = 1$ and where $k < 5$. For all other combinations of *k* and at β, the rank function performed as well as or better than either the exponential or standardized linear difference function. Finally, weighting by exogenous ι produced almost no consistent or reliable improvements in the grouped condition.

3.2.2.2. PI. Figures 4 shows *PI* results when using exogenous ι. For the *PI* measure, coherence weighting was typically effective in both conditions. The exceptions were $k < 4$ for some functions and parameter levels. This confirms that coherence weighting consistently improved *MAE* in the grouped condition, albeit only slightly. Regardless, the magnitude of improvement was small. Beyond this, there were three important findings. First, if $\beta = 1$, 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 $\beta = 1$, increasing *k* tended to improve *PI*. However, as 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 superior, although *PI* for the exponential function was highest in most cases.

3.3. Discussion

The aggregation results confirm that coherence-weighted aggregation is an effective tool in the general-knowledge domain, reducing *MAE* by as much as 21.4% in the best-case scenario relative to equal-weighted equivalents (using endogenous weight, and the rank function at $\beta = 100$ and $k = 100$; see Figure 1). However, there are caveats to this conclusion. 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 (Figures 1-4). We saw the greatest variability among aggregation methods between $1 < k < 10$, with performance converging at larger group sizes. Coherence-weighting presented a slight risk-reward trade-off: increasing β improved accuracy (Figures 1 & 3) but decreased *PI* slightly (Figures 2 & 4). The reason is that large values of β often converged on 'coherent and certain' probabilities, like an extremization procedure. This was most often correct; however, in the rarer case that an individual is equally coherent and certain but incorrect, divergent answers with large penalties to accuracy will occur.

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. First, compared to the standardized linear difference function, the exponential function is unambiguously superior in terms of *PI* (Figures 2 & 4) and did not worsen *MAE* when we increased *k* (Figures 1 & 3), in part because the standardized linear difference function always zeros out the least coherent forecaster and thus the exponential function has the advantage of keeping this extra forecaster within the aggregation. Second, compared to the rank function, however, the exponential function was less biased (i.e., it did not penalize incoherence as much) and less effective at $\beta = 1$. It is conceivable that this more modest weighting strategy may be desirable and effective in certain contexts. Thus, the added flexibility of lower coherence weight is potentially useful. 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.

Our second aim was to determine whether individual differences in average levels of coherence are useful predictors of correspondence. We found limited supporting evidence. Compared to endogenous ι, exogenous ι was much more error-prone. Weighting according to exogenous ι only produced numerical improvements in the spaced condition (Figures 3 & 4). This demonstrates important constraints on the advantages of coherence weighting reported in Karvetski et al. (2013). Moreover, even within the spaced condition, changes in both group size

and weighting strategy had inconsistent effects when using exogenous ι. Nevertheless, each of our functions improved over the equal-weighted average in most cases (Figures 3 & 4). This suggests that aggregators have little to gain from employing the method—in both absolute and relative terms—but also little to lose. Even 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 (Figures 1-4). This is counterintuitive, as the significant gaps between related judgments make it difficult to remain coherent. Certainly, participants who diligently respect coherence constraints would have to remember the truth probability they provided at least 60 questions ago! One explanation for this effect relies on the fact that correct responses are, by definition, coherent (Hammond, 2000). When coherence is difficult to maintain, *true* experts who know the correct response will nevertheless be coherent. By contrast, the poor mental availability of prior estimates will make it difficult for non-experts to be coherent unless coherence is an incidental byproduct of suboptimal response strategies; e.g. a mid-lined .50 response to every prompt is coherent for complementary probability judgments. . 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). However, another potential source of false positives is overconfident, *false* experts who are certain that the incorrect response is accurate (Tetlock, 2009), who will also be both coherent and incorrect.

The importance of this balance between true and false positives was clear in a Bayesian judgment experiment (Karvetski et al., 2020): 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 for coherence and the knowledge required for correspondence.

Rather, individual aggregators can exploit individual differences in coherence to *diagnose* potential experts for factual queries. This diagnostic process benefits from spacing logically related judgment queries farther apart so that: (1) coherence does not misdiagnose "mere" awareness of the logical constraints, and (2) coherence does not misdiagnose "overconfidence" as expertise. Both are unlikely to be reliable indicators of accurate world knowledge.

4. Statistical-evidence Domain

Wu et al. (2017) investigated how information presentation influenced the accuracy of probability judgments. Across four experiments, 2,858 participants completed the *Turtle Island* task. The fictional scenario in the task concerned an island populated by two turtle species (i.e., Bayosians and Freqosians). The species were visually identical, identifiable only by differentiating 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 differed between species. Wu et al. (2017) provided participants in different conditions with different background information: half of the participants received the Bayesian *prior and likelihood* (PL) probabilities; the other half received the Bayesian *marginal and posterior* (MP) probabilities. Participants used this information to judge the remaining, missing probabilities. That is, participants in the PL condition used the information to judge the marginal and posterior; participants in the MP condition used the information to judge the prior and likelihood. Critically, using Bayes' theorem (or intuitive reasoning consistent with Bayes' theorem) it was possible to calculate these probabilities precisely. Therefore, coherence *and* correspondence depended on reasoning consistent with Bayes' theorem.

4.1. Method

The *R* scripts and aggregated data described below are available on the OSF (Collins et

al., 2021)[1].

4.1.1. Incoherence Metrics

Whereas Karvetski et al. (2013) examined sets of probabilities constrained to Kolmogorov's axioms, the estimates elicited in the Turtle Island task concerned sets of probabilities constrained to Bayes' theorem:

$$P(x \mid y) = \frac{P(x) \times P(y \mid x)}{P(y)} \quad [6]$$

where 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, $P(y \mid x)$, a conditional probability describing the chance that y will occur given that x occurred; the marginal, $P(y)$, an unconditional probability describing the chance that y will occur; and the posterior, $P(x \mid y)$, a conditional probability describing the chance that x will occur given that y occurred.

In the Turtle Island experiments, the labeling of the underlying probabilities was randomized. For simplicity, we canonized the 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) $P(E \mid B)$ and $P(E \mid F)$ are the likelihood probabilities that a Bayosian and Freqosian turtle expressed the E form of the DE gene, respectively, and similarly, $P(L \mid B)$ and $P(L \mid F)$ are the likelihood probabilities that a Bayosian or Freqosian turtle expressed the L form of the LM gene, respectively; and (d) $P(B \mid E)$ and $P(B \mid D)$ are the posterior probabilities that a turtle with the E or D form of the DE gene was a Bayosian turtle, respectively, and similarly, $P(B \mid L)$ and $P(B \mid M)$ are the posterior probabilities that a turtle with the L or M form of the LM gene was Bayosian, respectively. The researchers constrained the

[1] For the participant-level raw and coherentized data sets, please contact co-author Charley M. Wu (charley.wu@uni-tuebingen.de).

probability of binary complements such that $P(F)$, the prior probability that a turtle was Freqosian, was equal to $1 - P(B)$. Participants input probabilities using a visual analog slider; 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 the best of our knowledge, no practitioner has yet applied the CAP to conditional probability estimates with nonlinear constraints such as Bayesian probabilities. To make the problem computationally manageable, we coherentized and aggregated participants' unconditional and conditional probability estimates separately. The inequality constraints were identical between conditions and judgments, i.e., $0 \leq P \leq 1$ for all judgments. By contrast, the equality constraints differed between condition and judgment. To construct the equality constraints for the unconditional and conditional judgment tasks in each condition, we rearranged and substituted terms in Bayes' formula [6] to solve for zero using only the given information and judged probability. For example, for unconditional probability judgments in the *PL* condition, participants judged marginal probabilities $P(E)$ and $P(L)$. For both judgments, we solved for 0 using a combination of the marginal judgment and the prior $P(B)$, the likelihoods $P(E \mid B)$ and $P(E \mid F)$ or $P(L \mid B)$ and $P(L \mid F)$. We provide the full set of judgment constraints in Appendix A and the code on OSF.

We applied the CAP using the "NlcOptim" package in R (Chen & Yin, 2019). The optimization procedure required three user-set *tolerance* parameters that controlled the criterion and stopping rules for the aggregation procedure. A complete description of these tolerances and their functions can be found in the package documentation. Due to the limited precision afforded to participant responses, ostensibly coherent responses could receive a score of $\imath = .01$ due to rounding errors. To compensate for this rounding-induced incoherence, each of our tolerances was set to .01. In practice, this treated any score as coherent if both equality constraint violations and $\imath$ were less than .01. For aggregation, we treated all values of $\imath \leq .01$ as if they were $\imath = 0$. For the exogenous $\imath$, we used the $\imath$ calculated for the opposite component: for

the unconditional probability estimates, we weighted participant responses using the conditional probability estimate ι, and vice versa.

4.1.2. Correspondence Metrics

We calculated the *MAE* [5] for the unconditional probability estimates and the conditional probability estimates separately before averaging the two. The *MAE* reflected the error across both the given and judged probabilities. As with the general-knowledge domain, we also calculated *PI*. Finally, because the only difference between the four experiments in Wu et al. (2017) is the values of the canonized environmental probabilities, we collapsed across the four experiments for our analyses.

4.1.3. Simulation and Aggregation Method

We compared the effectiveness of our aggregation methods using a method identical to that described in the general-knowledge domain with one exception. Because the Turtle Island experiments had different probabilities for the priors, likelihoods, marginal and posterior probabilities, we had to aggregate each experiment separately. This corresponds to a 2 (Condition: prior & likelihood, marginal & posterior) × 4 (Experiment: 1-4) × 100 (Group Size: k = 1-100) × 3 (Weighting Function: exponential, standardized linear difference, rank) × 3 (Parameter: β = {1, 10, 100}) × 2 (Item Sample: endogenous, exogenous) design. Again, we included an equal-weighted function for a baseline comparison. Excluding redundant combinations of conditions, the final analyses consisted of 15,200,000 unique data points. We collapsed across the 1,000 bootstrap samples and 4 experiments to calculate 3,800 *MAE* values (1,900 per condition) as well as 3,600 *PI* values (1,800 per condition).

4.1.4. Analysis Plan

We approached the analysis similarly to the general-knowledge domain. Although *SE* was typically higher in this experiment, the largest was still *SE* < .002 and the majority were *SE* ≤ .001. Thus, we will again rely on visual inferences from the graphs for *MAE* comparisons. For

our *PI*, we used the 99% confidence interval corresponding to the binomial distribution for random chance of 4000 events (*lower bound* = .480, upper bound = .520).

4.2. Results

Figures 5-8 show the results of the coherence-weighted aggregation for *MAE* and *PI* for endogenous ι and exogenous ι, respectively, as a function of k and β. Again, we scaled the x-axis for group size k logarithmically (base 10). As in the prior study, aggregation performance at group size $k = 1$ was equal to the sample *MAE*. Unlike the prior study, participants' responses can bracket the correct answer, and simple equal-weighted aggregation reduced error (Larrick & Soll, 2006).

4.2.1. Endogenous ι

4.2.1.1. MAE. Figure 5 shows *MAE* results when using endogenous ι. 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. 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.

4.2.1.2. PI. Figures 6 shows *PI* results when using endogenous ι. Each of our methods improves 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 show that, given these data and the weighting strategies we considered, the best results occurred with group size in the low double-digits and moderate to strongly biased (i.e., $\beta \geq 10$) weighting strategies.

To better understand why performance appeared to worsen between $k = 25$ and $k = 100$ in the PL condition where $\beta = 100$, we examined the distribution of bootstrapped *MAE* values (Figure 7). As the histogram shows, increasing *k* from 25 to 100 produces a slight increase in cases where *MAE* is between .00 and .05, but a larger increase in the number of cases where *MAE* was between .20 and .25. An inspection of the raw data revealed that four participants across three of the experiments answered .50 for every estimate. This response tendency produced an incidentally coherent judgment with an $MAE \approx .25$. These incidentally coherent cases explain the second peak in the histogram. That is, coherence-weighting strategies assigned a high weight to these incidentally coherent but inaccurate judges. Although rarer than coherent and accurate judges, larger group sizes increased the chance that the algorithm selected at least one of these individuals, reducing the accuracy of the aggregated judgment.

4.2.2. Exogenous ι

Figures 8 and 9 show the results of the coherence-weighted aggregation for *MAE* and *PI*, respectively. As with the general-knowledge domain, the exogenous ι weighting strategy was much less effective than the endogenous 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 followed the broad patterns established with endogenous ι, including the slight worsening of *MAE* for large group sizes at $\beta = 100$. The results show that the best performance occurred with a combination of moderate-to-large group size combined with a moderate to severely biased combination of weighting function and tuning

parameter.

4.3. Discussion

The results demonstrate that aggregators can apply coherence efficaciously to the statistical-evidence domain with nonlinear optimization constraints. The strategy reduced *MAE* by as much as 38.0%; endogenous ι and standardized linear difference function at β = 10 and *k* = 100\ relative to equal-weighted equivalent (Figures 1 & 2). In applying the CAP to a set of probabilities bound by nonlinear constraints (Appendix A & Formula 6), we generated a metric ι that was an effective basis for coherence-weighted aggregation. This shows that practitioners can efficaciously apply the CAP to complex conditional probability estimates with nonlinear constraints.

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 worked best with large group sizes and moderate weighting (β = 10) or moderate group sizes ($10 \lesssim k \lesssim 25$) and severe weighting (β = 100). Incremental changes were minimally positive and sometimes negative for $k > 10$, reflecting strongly diminishing returns (Figures 5-6, 8-9).

The finding that increasing group size past a critical point worsened accuracy (Figures 5-6, 8-9) is particularly interesting. In general, the addition of information should not worsen accuracy. However, coherence weighting restricts the pool of information to a smaller set of coherent individuals, resulting in some information loss. Our findings show this sometimes results in the selection of a coherent but inaccurate assessor (Figure 7). These individuals are conceptually like the *false positives* in the general knowledge domain: they are coherent not because of mathematical rigor or internal consistency (as with *true positives*). Rather, they are coherent due to sub-optimal response biases that incidentally produce coherent responses (Bruine de Bruin et al., 2002; Fischhoff & Bruine de Bruin, 1999). In fact, this is why Karvetski et al. (2013) found that excluding the complementarity constraint from coherence-weighting

schemes improved accuracy; judges who expressed epistemic uncertainty by responding .5 to both $P(A)$ and $P(A^c)$ would otherwise be incidentally coherent. Unfortunately, coherence-weighted aggregation cannot discriminate between false positives and true positives. We return to this issue in the General Discussion.

Regarding our aim of evaluating the usefulness of exogenous coherence weighting, we find that exogenous ι provided an effective basis for coherence-weighted aggregation. Promisingly, weighting by exogenous ι improved *MAE* most of the time in most cases in both the PL and MP conditions (Figure 8-9). 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.

5. General Discussion

The present investigation contributes to our understanding of prescriptive and descriptive theoretical issues concerning coherence-weighted aggregation. First, we provided a proof of concept that coherence-weighting can be effective in two dissimilar task domains where expert judgment is frequently relied on: general knowledge tasks and statistical-evidence tasks. Coherence-weighting reduced *MAE* relative to equal-weighting by as much as 21.4% in the general-knowledge domain (Figure 1), and 38.0% in the statistical-evidence domain (Figure 5). These are large and reliable improvements, comparable to other contemporary performance-weighted aggregation methods such as Cooke's classical method and the contribution weighted method (Budescu & Chen, 2015; Cooke et al., 1988). Second, we showed that the ideal group size and parameterization are task-dependent. Third, we found the exponential function to be the most intuitive and flexible weighting function. Fourth, we show that out-of-sample exogenous ι holds some promise of predicting correspondence on tasks (Figures 3-4, 8-9), particularly in the statistical-evidence domain. Fifth, we show that the CAP (Osherson & Vardi, 2006; Predd et al.,

2008) can be efficaciously applied to sets of probabilities characterized by nonlinear coherence constraints. Importantly, these types of problems are frequently encountered in forecasting where conditional probabilities must be considered (Mandel, 2014). Finally, our results indicate that there are at least two potential reasons that coherence-weighting is effective. For some tasks, coherence and correspondence are scored on similar bases, in which case coherence may be a *causal* determinant of accuracy. For other tasks, aggregators can simply exploit the fact that correct answers are, by definition, coherent, and therefore coherence *diagnoses* potential experts in a crowd. Whether aggregators exploit the *causal* or *diagnostic* relationship has consequences for elicitation and aggregation strategies.

Next, we will discuss these theoretical issues in greater detail before revisiting the issues of ideal implementation with these theoretical developments in mind. We first examine *why* coherence weighting is effective in Section 5.1, then discuss how to best exploit the *wisdom of the coherent* in Section 5.2, then conclude with a discussion of the decision-theoretic framework applying coherence weighting in Section 5.3.

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

We have proposed that there are two relationships between coherence and correspondence that aggregators can capitalize on to exploit the "Wisdom of the Coherent." The first, and most straightforward, is the causal relationship. In these cases, 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 received statistical information (e.g., base rates and diagnostic probabilities) that was sufficient to correctly produce the target judgments using Bayes' theorem. One might say that knowledge of Bayes' theorem [6] suffices to yield both coherent and accurate judgments. However, this obscures the fact that coherence is the basis for scoring accuracy in such cases.

The second relationship we described occurs when coherence *diagnoses* experts in the

opinion pool. Consider the general-knowledge task we examined here (see also Karvetski et al., 2013), which centered on the first man to step foot on the moon. A true expert will be certain that Neil Armstrong was the first person to set foot on the moon, $P(A) = 1$. This expert will be equally certain that Buzz Aldrin was not the first person to set foot on the moon, $P(B) = 0$, and also that one of either Neil Armstrong or Buzz Aldrin was the first man to step on the moon, $P(A \cup B) = 1$. This combination of responses is both correct and coherent, even though the assessor may not even understand the coherence principles applied. Rather, the exploitable relationship here leverages the fact that correct answers are, by definition, coherent.

From a decision-theoretic standpoint, it is important for aggregators and decision-makers seeking to improve judgment accuracy to consider which relationship they intend to exploit. The two relationships were dissociable in the present research, but they are neither mutually exclusive nor task-specific. For instance, knowledge of Kolmogorov's axioms suffices to yield perfect accuracy in the general-knowledge domain if the participants are certain about at least one query: Neil Armstrong *was* the first man to step on the moon, therefor Buzz Aldrin must not be. Conversely, an expert of a real-life Turtle Island might simply know, as a matter of fact, the different environmental probabilities without necessarily applying Bayes' theorem to arrive at the answer. Thus, a mixture of signals likely exists in all tasks to some degree.

Critically, forecasting is a good example of where both types of bases for exploiting coherence may be present. For instance, accurate forecasting may demand knowledge of factors that are shaping the outcome as well as the statistical knowledge to coherently combine this information. Nevertheless, the nature of the relationship practitioners hope to exploit has implications for how best to extract the wisdom of the coherent among the crowd. This includes the ideal elicitation process before aggregation, the potential utility of exogenous ι, and the optimal aggregator function and group size. Next, we examine these issues in greater detail.

5.2. Exploiting the Wisdom of the Coherent

5.2.1. Elicitation Method

The coherence-correspondence relationship aggregators wish to exploit may affect decisions about the elicitation process. Take, for instance, the decision to structure elicitations in ways that mitigate incoherence or inaccuracy. For example, 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 logical relations salient (Wu et al., 2017); consider-the-opposite strategies such as dialectical bootstrapping (Herzog & Hertwig, 2014); and eliciting confidence intervals to improve best estimates (Hemming et al., 2018; cf. Mandel et al., 2020). If aggregators wish to exploit the causal route, 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 information clarity and achieve low ι. Those who do not understand the relevant probabilistic calculus are unaffected—for better or worse—by the difficulty of the elicitation process.

If aggregators wish to exploit the diagnostic route, however, it may be useful to let individual differences in incoherence "flourish." Aggregators can exploit these differences through coherence-weighted aggregation. In Karvetski et al. (2013) and our re-analyses, maximally spacing logically related judgments made coherence principles, such as additivity, less mentally accessible. This elicitation strategy allows aggregators to harness the wisdom of the coherent more efficiently. The idea of spacing judgments to increase incoherence might seem to be a perverse and counterintuitive strategy. However, when considered alongside recalibration and aggregation methods, its advantages become clear. The reasoning is that the relationship between coherence and correspondence could be incidental or diagnostic. For grouped judgments, it is more likely that a judge 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. For spaced judgments, it is less likely that judges will be aware of the logical constraints on the set of

probabilities. In this case, coherent assessors are more likely to be coherent simply because they know the correct answer. Accordingly, weighting by coherence will separate the wheat from the chaff.

Interestingly, several methods hold promise for capitalizing on both *causal* and *diagnostic* relationships. For instance, one study showed that allowing judges to opt in to (or out of) judgments improved aggregated accuracy (Bennet et al., 2018). The researchers suggested that metacognitive assessments of one's expertise can predict accuracy. For the diagnostic signal, much of the benefit came from coherent and certain individuals (e.g., probabilities of exactly 0 and 1), exploiting a similar principle. For the causal route, we might expect judges with no knowledge of Bayes' 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 incidental coherence that arrives incidentally via response biases such as straight-lining or responding .5 to convey one's utter epistemic uncertainty. Aggregators must also consider the prevalence of *false* experts within the opinion pool who may be bullish and overconfident about incorrect responses, also achieving incidental coherence (Tetlock, 2005).

Because a function cannot distinguish between these false experts and true experts, and because apparently lazy mid-lined responses might, in fact, be epistemically justified, future research on coherence-weighting could profitably focus on elicitation procedures that reduce confounds of the coherence signals. As we have shown, failure to identify the cases and the conditions where incidentally coherent responses are common could impose costs to accuracy. For example, information about response consistency (reliability) can improve accuracy through repeated elicitation (Miller & Steyvers, 2017). For the diagnostic route, reliability is a proxy for certainty and expertise. For the causal route, reliability is a proxy for the correct application of relevant probabilistic axioms and theories. In both cases, false positives are likely to be minimized. Regardless of the signal practitioners wish to exploit, they should take care not to

structure their elicitation processes such that incidental coherence is trivial to achieve, such as with the mid-lined responses in Wu et al. (2017) or the complementary constraint in Karvetski et al. (2013).

Although we have emphasized the utility of coherence-weighting as a tool that can be used even in data-poor environments, nothing prevents decision-makers from combining it with other techniques to further improve the accuracy of aggregated estimates such as other performance-weighted methods (Collins et al., in press; Bolger & Rowe, 2015; Budescu & Chen, 2015; Himmelstein et al., 2021). Furthermore, practitioners can combine ensembles of methods such as competitive (Lichtendahl et al., 2013) or structured (Fraser et al., in press) elicitation methods; enhancing the salience of private versus public information (Larrick et al., 2012), choosing smaller, wiser crowds (Soll et al., 2010); trimming opinion pools to account for under- and over-confidence (Yaniv, 1997; Jose et al., 2014); up-weighting assessors who update estimates frequently in small increments (Atanasov et al., 2020); or extremizing judgments (Baron et al., 2014; Hanea et al., 2021; Satopää & Ungar, 2015) to further improve accuracy. The latter is particularly appealing in general-knowledge tasks where the "outcomes" are, by definition, extreme. Unlike history- and disposition-based methods, these methods do not require a more data-rich environment than ones where coherence methods may be applied.

5.2.2. The Utility of Exogenous ι

The distinction between ι that diagnoses accuracy versus ι as a causal determinant of accuracy has the benefit of explaining the exogenous ι results intuitively. To the extent that coherence is a causal determinant of accuracy on a task, it is not surprising that ι on that task would be a useful indicator of correspondence on a closely related task. For instance, the ability to derive coherent and accurate Bayesian probabilities on one Bayesian task (e.g., one involving the estimation of likelihoods) is a useful indicator of the 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 ι is to the task at hand. This implies that a topic-specific conceptualization of exogenous ι could be particularly effective, which is an issue that future research could investigate.

Finally, although our present aggregation models cannot account for its effect, the variance of exogenous ι may have the potential for identifying experts. Homogeneity in coherence (or incoherence) may indicate someone who does (or does not) understand the relevant probabilistic axioms. By contrast, heterogeneity—a judge who is sometimes coherent and sometimes not—may represent an individual who is sometimes certain and coherent and, in other cases, strategically incoherent; e.g. mid-lining responses to express uncertainty or changing one's mind partway through elicitation. This aspect of exogenous ι deserves further investigation.

5.2.3. Choosing the 'Best' Function and Group Size

The present research compared the efficacy of several functions used in previous studies (Fan et al., 2019; Karvetski et al., 2013; Wang et al., 2011; see Table 1). The present findings clarify, 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* in both task domains. Nevertheless, we draw the following lessons from these findings. First, both the weighting function and β work in tandem to determine how quickly the aggregation weight approaches zero as ι increases. In much the same way that β = 100 is more biased than β = 10, the rank function is more biased than the standardized linear difference function. As we have seen in our re-analysis of Wu et al. (2017), the most biased weighting strategies are not always optimal. In fact, across both re-analyses we conducted, more biased strategies often resulted in lower *PI*. In this way, the exponential function is more flexible than either the rank or standardized linear difference function given that its responsiveness to changes in β allows 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 standardized linear difference function in both studies. The rank function was occasionally too biased by default, whereas the standardized linear difference function exhibited non-monotonic improvements.

Curiously, regarding the standardized 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 ω = 0 to the most incoherent judge, reducing the effective size of the opinion pool by 1. This has a proportionally large biasing effect for small group sizes, equivalent to a step function at $k = 2$ (i.e., choose the most coherent responder). That much of the benefits of coherence-weighting are the result of excluding incoherent judges in the opinion pool suggests that aggregators might efficaciously combine small (or select) crowd aggregation strategies (Mannes et al., 2014) with coherence-weighting. Aggregators could employ a select crowd with a step function that assigns a weight of 1 to judges that suffices a coherence criterion (e.g., ι = 0) and a weight of 0 to those that do not. In other words, aggregating only those judges that were coherent or coherent. Given its bias, the performance of such a model is similar to our ranked function where β = 100 Future research should investigate the efficacy of this 'chasing the *coherent*' strategy and the potential accuracy tradeoffs of permitting small coherence violations (i.e., critical values where ι > 0).

Regarding group size, we find that a larger crowd is generally better. However, gains in accuracy were associated with diminishing returns, similar to other studies of aggregation (Han & Budescu, 2019; Mandel et al., 2018). Beyond a small-to-moderate group size of about ten judges, the addition of a single extra participant often produced minimal benefits. This is in line with research showing that small-to-medium-sized groups are ideal during aggregation (Han & Budescu, 2019; Mannes et al., 2014; Navajas et al., 2018). The results of Turtle Island also show that large groups increased the risk of selecting at least one "false positive"—an individual who was incidentally coherent. Indeed, incidental coherence also diluted the effectiveness of

coherence-weighted aggregation in the general knowledge domain (Karvetski et al., 2013). 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. Simple procedures such as scanning for straight-liners and omitting their judgments could mitigate the problem, though this is likely to be insufficient. Rather, we stress that practitioners who wish to employ coherence-weighting elicitation adopt a priori steps to reduce incidental coherence at the elicitation stage.

5.3. Toward a Decision-Theoretic Framework for Optimizing Probability Judgment

The foregoing discussion provides a proof of concept for the efficacy of coherence-weighted aggregation in two distinct domains where expert judgment is frequently called upon. It also provides the basis for sketching a decision-theoretic framework for optimizing probability judgments. Central to this framework is the view that decisions about judgment optimization strategies must focus broadly on *ensembles* of relevant factors, which include, but 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. Researchers must better understand how the pillars of optimization—information representation, judgment elicitation, and post-judgment recalibration and aggregation—interact amongst themselves and task characteristics (Zellner et al., 2021). Currently, few studies take this multi-interventionist approach, and we encourage further research along these lines.

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.

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., Mandel, D. R., & Budescu, D .V. (in press). Performance-weighted aggregation: Ferreting out the wisdom within the crowd. In M. Seiffert (Eds.), *Judgment in Predictive Analytics*. Springer, NY.

Collins, R. N., Mandel, D. R., Karvetski, C. W., Nelson, J. D., & Wu, C. M. (2022, Dec 12). *The wisdom of the coherent: Improving correspondence with coherence-weighted aggregation*. Retrieved From https://osf.io/46fdm/.

Cooke, R., Mendel, M., & Thijs, W. (1988). Calibration and information in expert resolution; a classical approach. *Automatica*, *24*(1), 87–93. https://doi.org/10.1016/0005-1098(88)90011-8

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.

Fraser, H., Bush, M., Wintle, B. C., Mody, F., Smith, E., Hanea, A., Gould, E., Hemming, V., Hamilton, D. G., Rumpff, L., Wilkinson, D. P., Pearson, R., Thorn, F. S., Ashton, R., Willcox,

A., Gray, C. T., Head, A., Ross, M., Groenewegen, R., Marcoci, A., … & Fidler, F. (in press). Predicting reliability through structured expert elicitation with the repliCATS (Collaborative Assessments for Trustworthy Science) process. *PLOS ONE*.

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. PLoS ONE, 16(9): e0256919. https://doi.org/10.1371/journal.pone.025691

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.

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton: Princeton University Press. https://doi.org/10.1515/9781400830312

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 (MAE) 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 optimization 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 Processes, 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

Mean Absolute Error in the General-knowledge Domain Using the Endogenous Metric

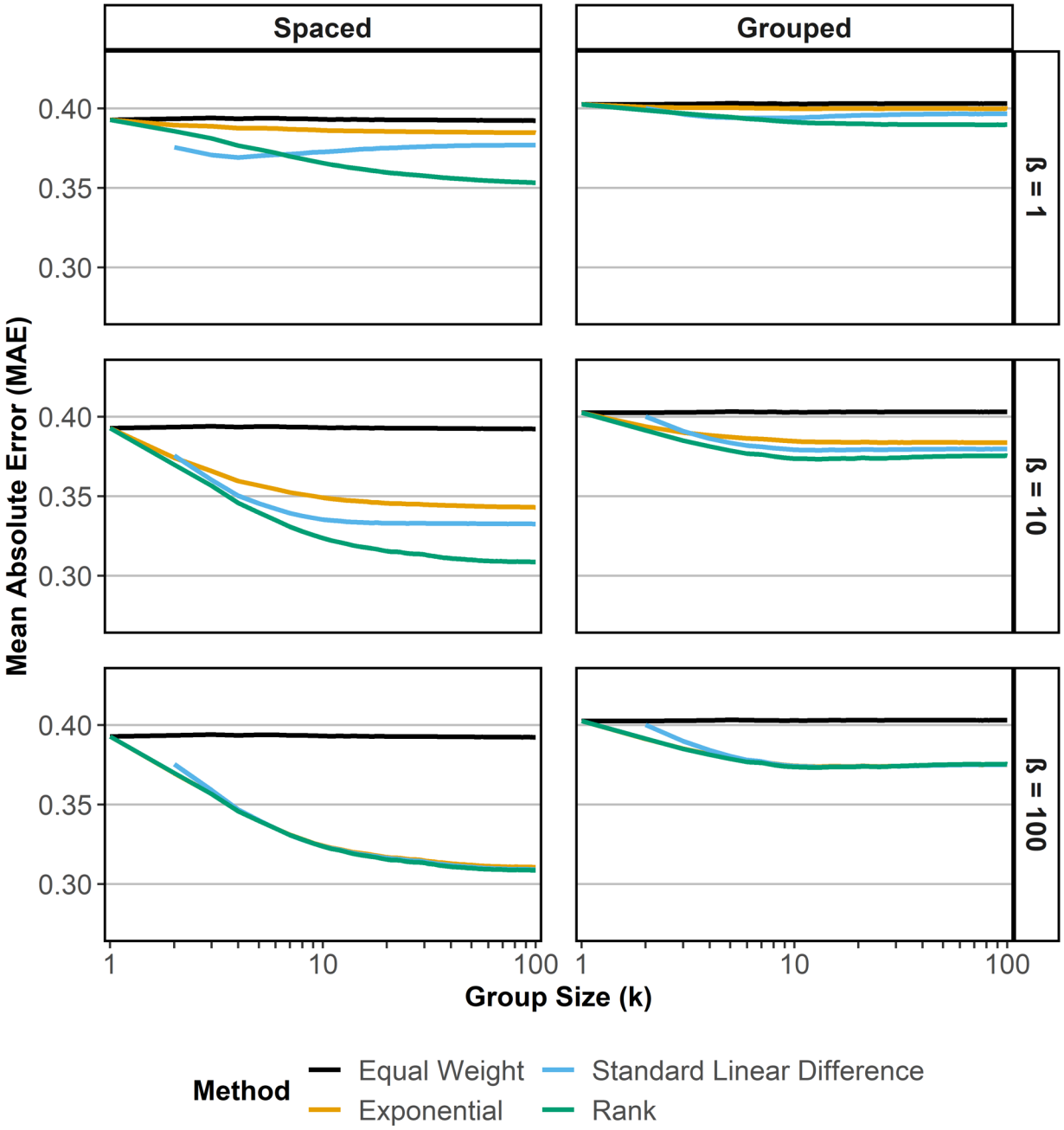
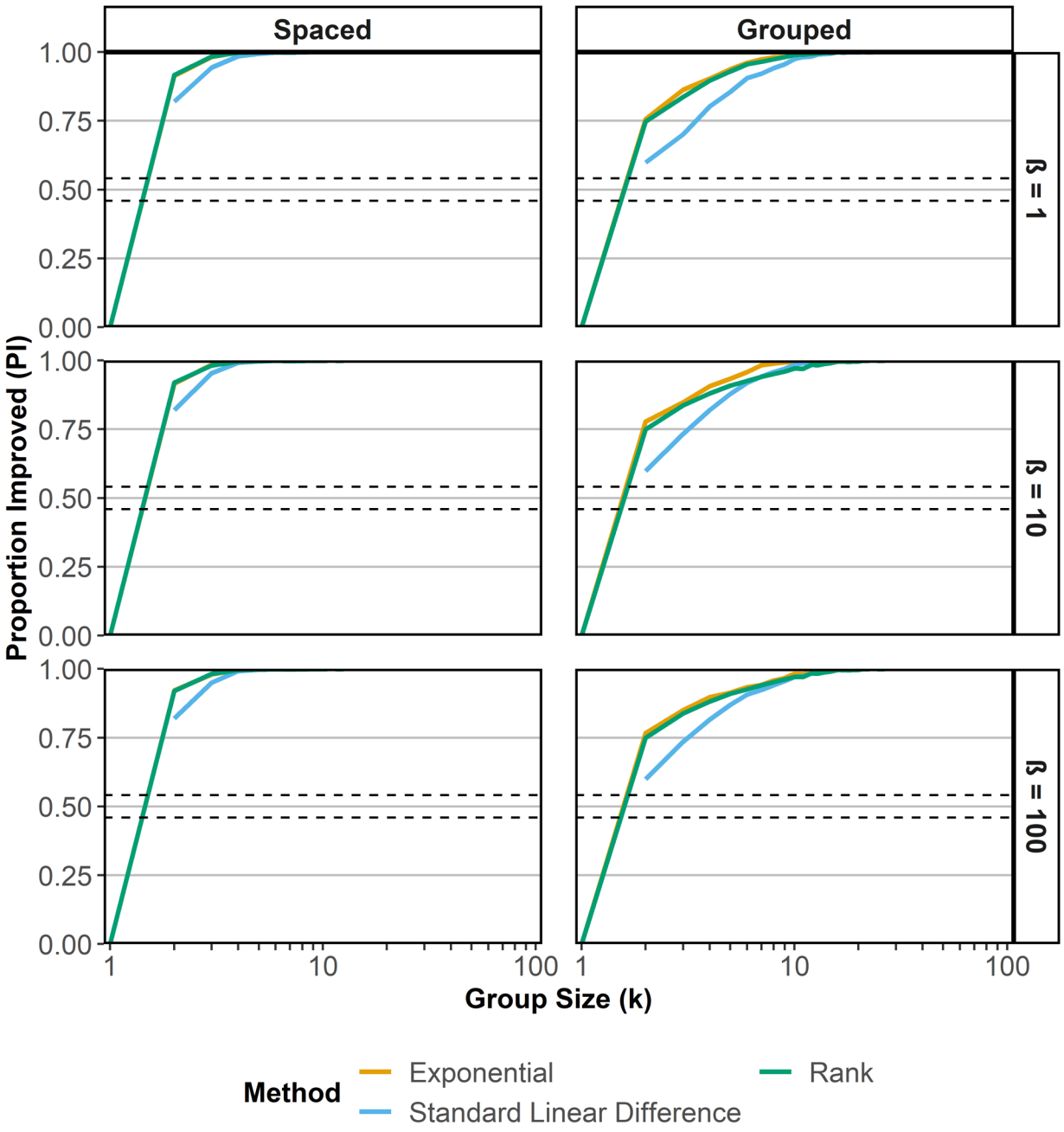


Figure 2

Proportion Improved in the General-knowledge Domain Using the Endogenous Metric



Note. The dashed lines are the *lower* (y = .459) and *upper bound* (y = .541) of the 99% CI of a random event using a binomial distribution with 1,000 events.

Figure 3

Mean Absolute Error in the General-knowledge Domain Using the Exogenous Metric

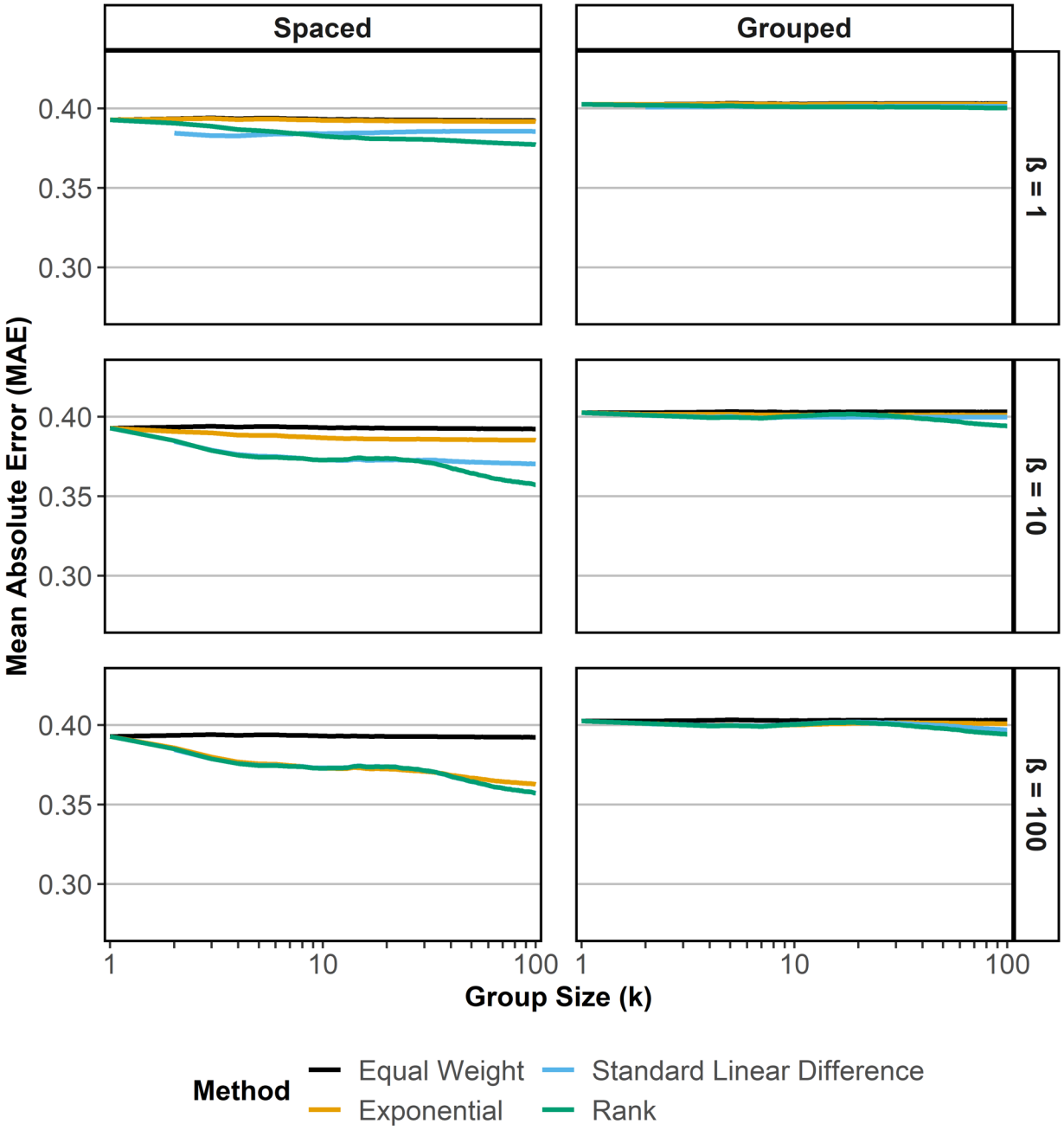
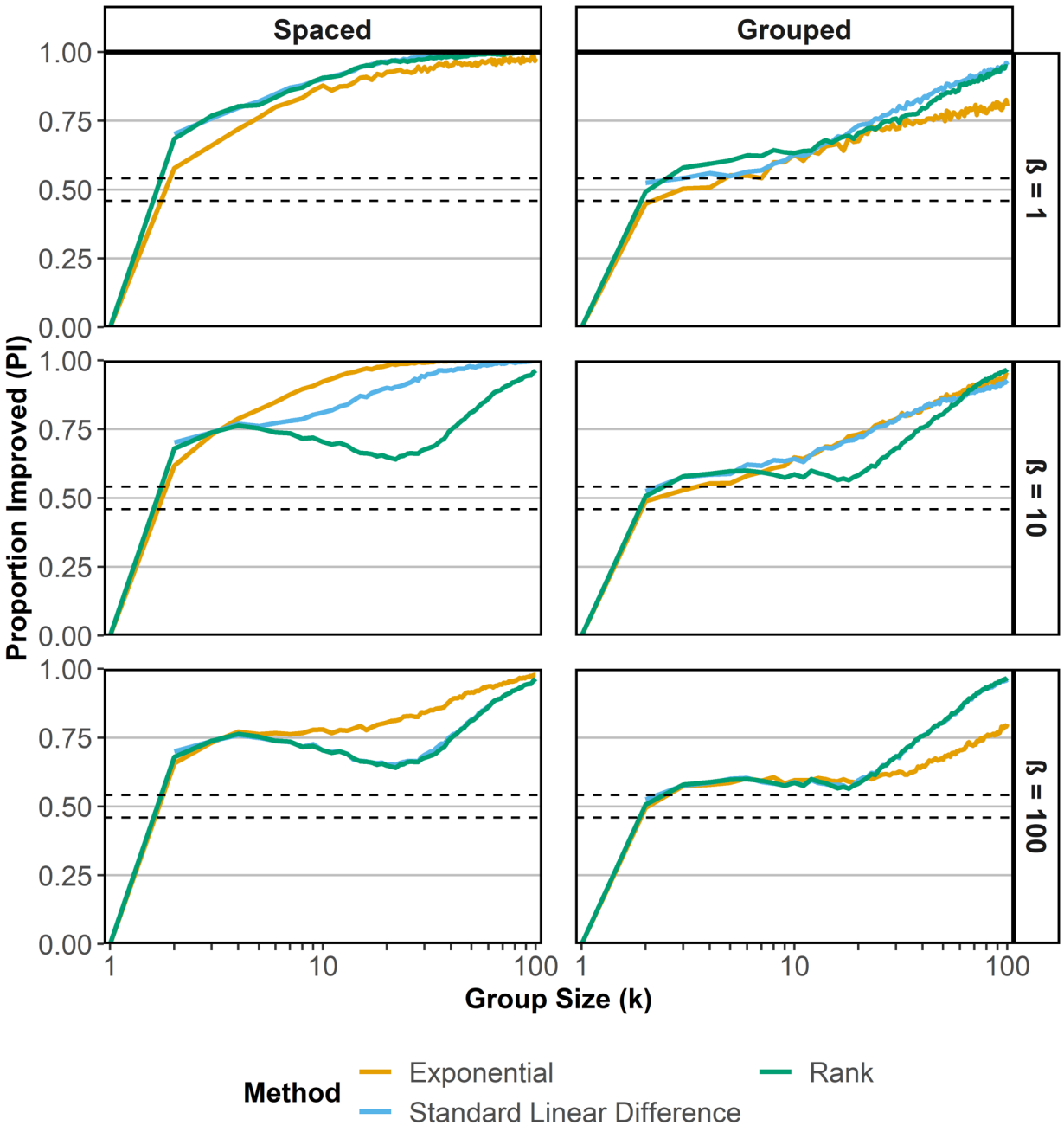


Figure 4

Proportion Improved in the General-knowledge Domain Using the Exogenous Metric



Note. The dashed lines are the *lower* (y = .459) and *upper bound* (y = .541) of the 99% CI of a random event using a binomial distribution with 1,000 events.

Figure 5

Mean Absolute Error in the Statistical-evidence Domain Using the Endogenous Metric

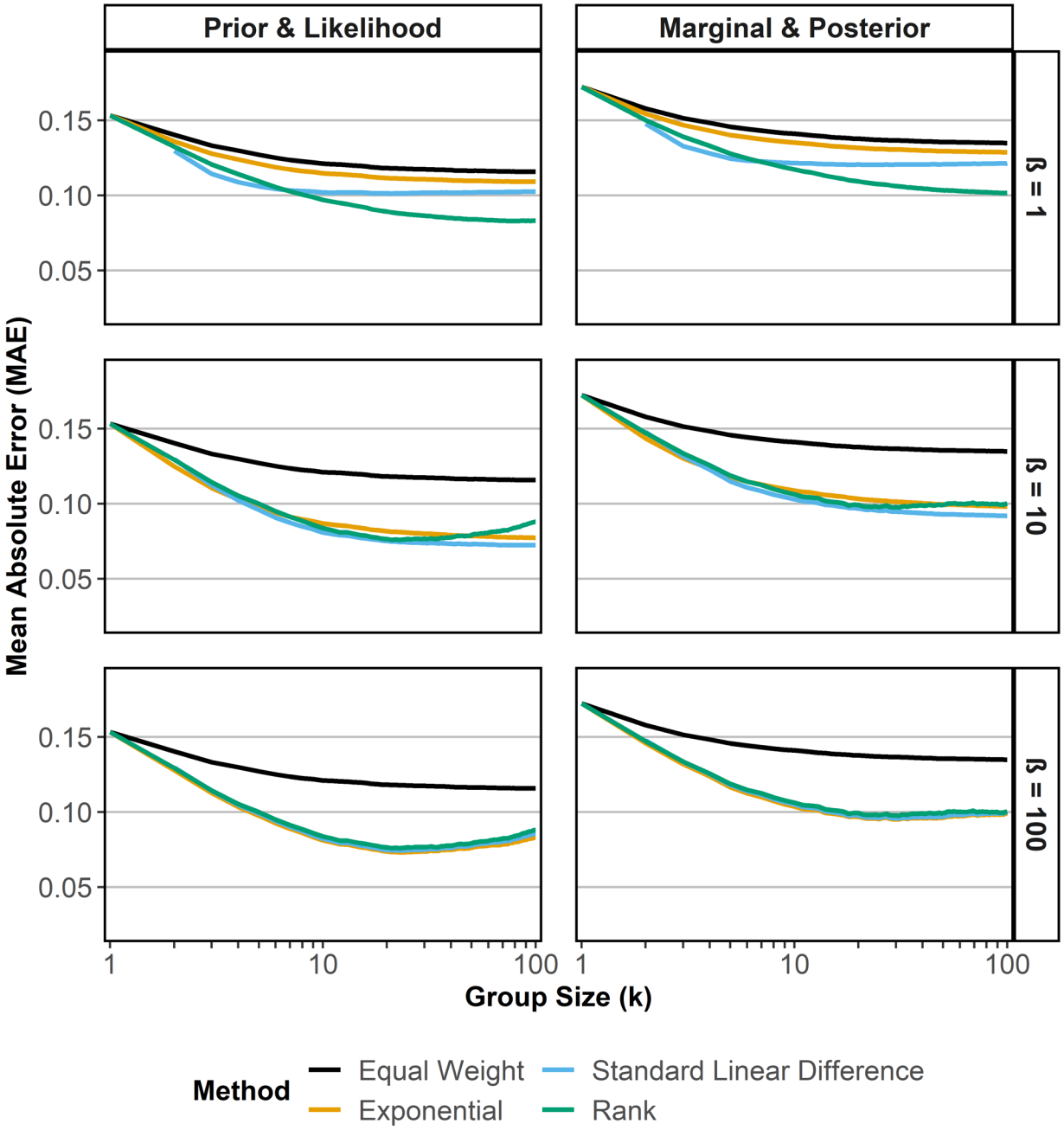
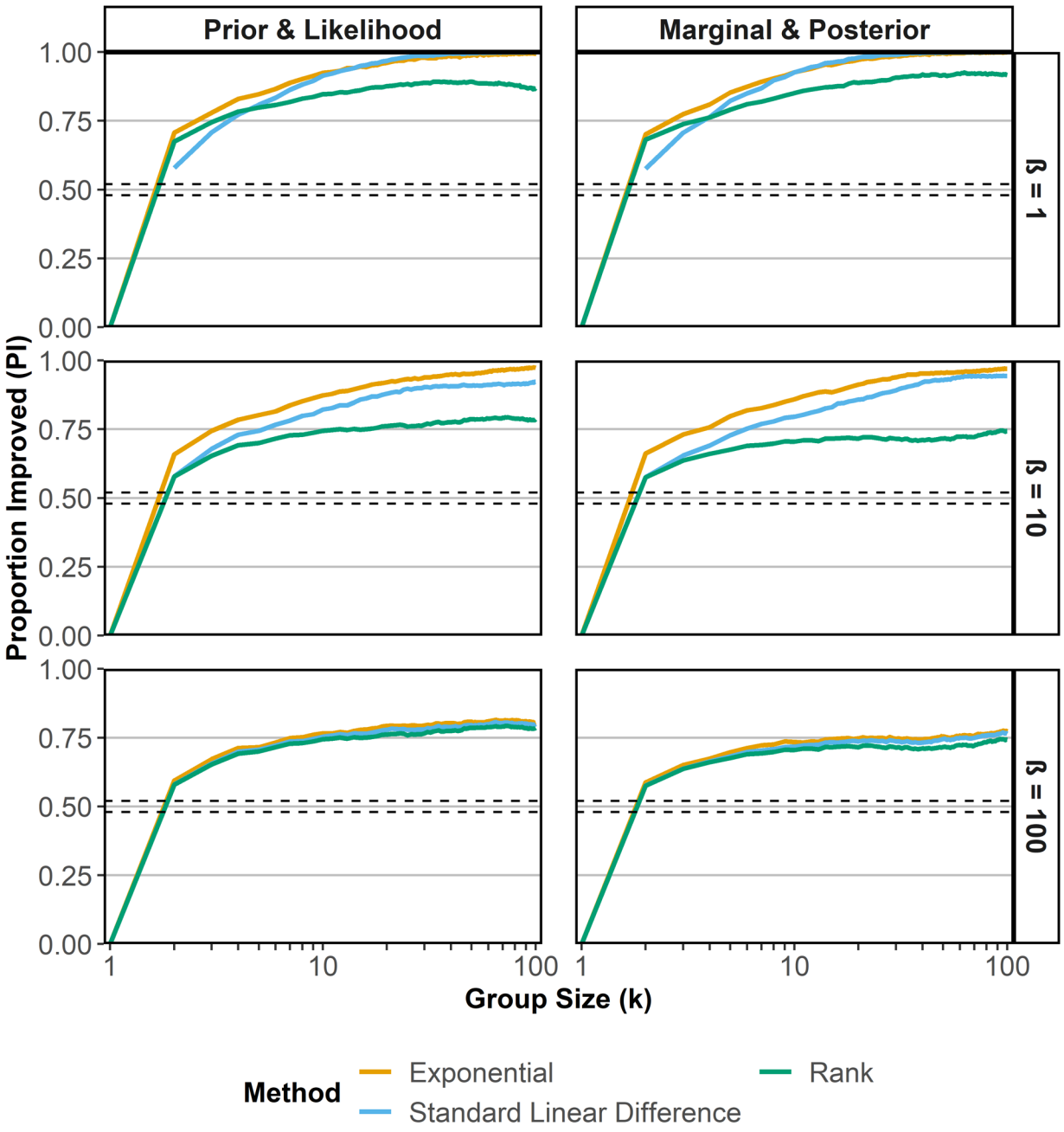


Figure 6

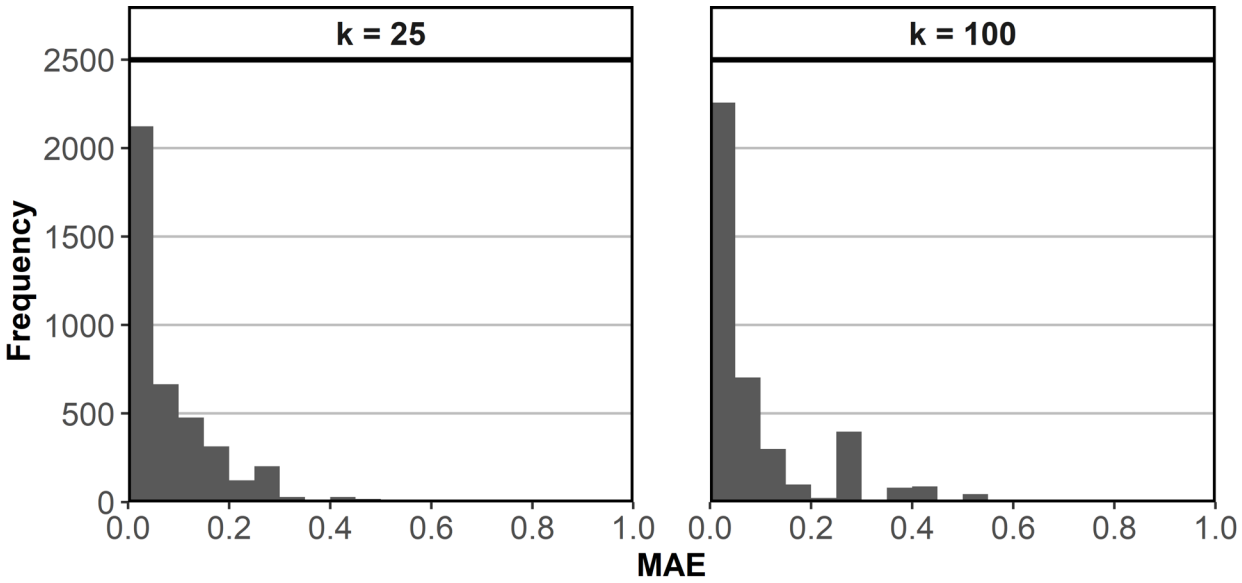
Proportion Improved in the Statistical-evidence Domain Using the Endogenous Metric



Note. The dashed lines are the *lower* (y = .480) and *upper bound* (y = .520) of the 99% CI of a random event using a binomial distribution with 4000 events.

Figure 7

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



Note: 4000 bootstraps total

Figure 8

Mean Absolute Error in the Statistical-evidence Domain Using the Exogenous Metric

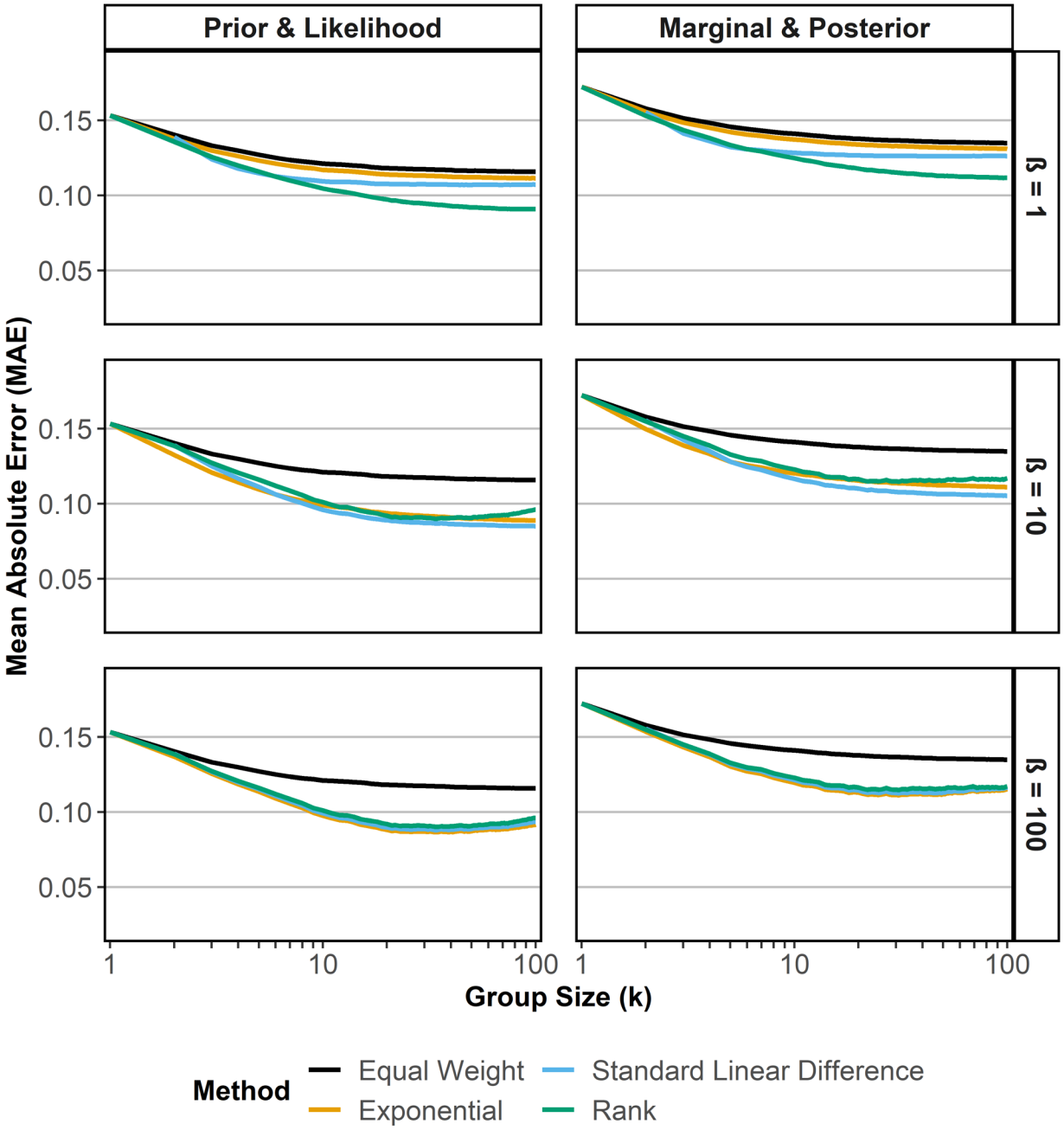
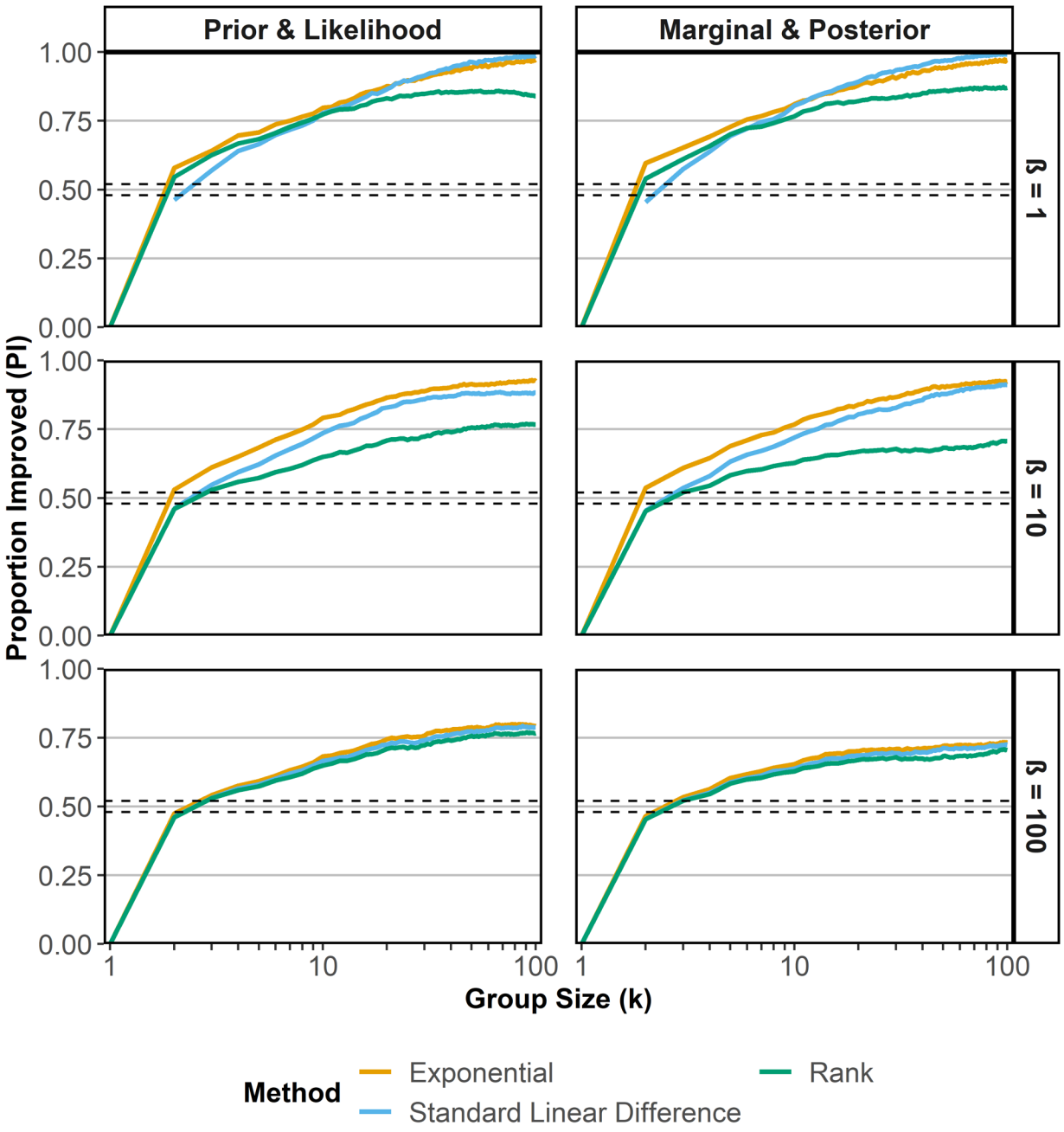


Figure 9

Proportion Improved in the Statistical-evidence Domain Using the Exogenous Metric



Note. The dashed lines are the *lower* (y = .480) and *upper bound* (y = .520) of the 99% CI of a random event using a binomial distribution with 4000 events.

Table 1

List of Functions for Translating the Incoherence Metric ι 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*

Condition	Given	Derive	Constraints
Prior & Likelihood	*Prior* $P(B)$ *Likelihood* $P(E \mid B)$ $P(E \mid F)$ $P(L \mid B)$ $P(L \mid F)$	*Marginal* $P(E)$ $P(L)$	<u>Non-Negativity/Bounds</u> (*Inequality*) $0 \leq \{P(B), P(E), P(L), P(E \mid B), P(E \mid F), P(L \mid B), P(L \mid F)\} \leq 1$ <u>Non-Linear/Bayesian</u> (*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 $P(B \mid E)$ $P(B \mid D)$ $P(B \mid L)$ $P(B \mid M)$	<u>Non-Negativity/Bounds</u> (*Inequality*) $0 \leq \{P(B), P(E \mid B), P(E \mid F), P(L \mid B), P(L \mid F), P(B \mid E), P(B \mid D), P(B \mid L), P(B \mid M)\} \leq 1$ <u>Non-Linear/Bayesian</u> (*Equality*) $0 = P(B \mid E) - \{P(B) \times P(E \mid B)\} / \{[P(B) \times P(E \mid B) + [1 - P(B)] \times P(E \mid F)\}$ $0 = P(B \mid D) - \{P(B) \times [1 - P(E \mid B)]\} / \{[P(B) \times [1 - P(E \mid B)] + [1 - P(B)] \times [1 - P(E \mid F)]\}$ $0 = P(B \mid L) - \{P(B) \times P(L \mid B)\} / \{[P(B) \times P(L \mid B) + [1 - P(B)] \times P(L \mid F)\}$ $0 = P(B \mid M) - \{P(B) \times [1 - P(L \mid B)]\} / \{[P(B) \times [1 - P(L \mid B)] + [1 - P(B)] \times [1 - P(L \mid F)]\}$
Marginal & Posterior	*Marginal* $P(E)$ $P(L)$ *Posterior* $P(B \mid E)$ $P(B \mid D)$ $P(B \mid L)$ $P(B \mid M)$	*Prior* $P(B)$	<u>Non-Negativity/Bounds</u> (*Inequality*) $0 \leq \{P(B), P(E), P(L), P(B \mid E), P(B \mid D), P(B \mid L), P(B \mid M)\} \leq 1$ <u>Non-Linear/Bayesian</u> (*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 $P(E \mid B)$ $P(E \mid F)$ $P(L \mid B)$ $P(L \mid F)$	<u>Non-Negativity/Bounds</u> (*Inequality*) $0 \leq \{P(E), P(L), P(E \mid B), P(E \mid F), P(L \mid B), P(L \mid F), P(B \mid E), P(B \mid D), P(B \mid L), P(B \mid M)\} \leq 1$ <u>Non-Linear/Bayesian</u> (*Equality*) $0 = P(E \mid B) \times \{P(E) \times P(B \mid E) + [1 - P(E)] \times P(B \mid D)\} - P(E) \times P(B \mid E)$ $0 = P(E \mid F) \times \{P(E) \times [1 - P(B \mid E)] + [1 - P(E)] \times [1 - P(B \mid D)]\} - P(E) \times [1 - P(B \mid E)]$ $0 = P(L \mid B) \times \{P(L) \times P(B \mid L) + [1 - P(L)] \times P(B \mid M)\} - P(L) \times P(B \mid L)$ $0 = P(L \mid F) \times \{P(L) \times [1 - P(B \mid L)] + [1 - P(L)] \times [1 - P(B \mid L)]\} - P(L) \times [1 - P(B \mid L)]$