Bus Inf Syst Eng 63(4):329–348 (2021)
https://doi.org/10.1007/s12599-020-00664-x

RESEARCH PAPER

A Simulation-Based Approach to Understanding the Wisdom of Crowds Phenomenon in Aggregating Expert Judgment

Patrick Afflerbach · Christopher van Dun · Henner Gimpel · Dominik Parak · Johannes Seyfried

Received: 20 December 2018 / Accepted: 18 May 2020 / Published online: 4 August 2020

Abstract Research has shown that aggregation of independent expert judgments significantly improves the quality of forecasts as compared to individual expert forecasts. This "wisdom of crowds" (WOC) has sparked substantial interest. However, previous studies on strengths and weaknesses of aggregation algorithms have been restricted by limited empirical data and analytical complexity. Based on a comprehensive analysis of existing knowledge on WOC and aggregation algorithms, this paper describes the design and implementation of a static stochastic simulation model to emulate WOC scenarios with a wide range of parameters. The model has been thoroughly evaluated: the assumptions are validated against propositions derived from literature, and the model has a computational representation. The applicability of the model is demonstrated by investigating aggregation algorithm behavior on a detailed level, by assessing aggregation algorithm performance, and by exploring previously undiscovered suppositions on WOC. The simulation model helps expand the understanding of WOC, where previous research was restricted. Additionally, it gives directions for developing aggregation algorithms and contributes to a general understanding of the WOC phenomenon.

Keywords Simulation · Forecasting · Expert judgment · Expert aggregation · Wisdom of crowds

Accepted after two revisions by Natalia Kliewer.

P. Afflerbach · H. Gimpel (✉) · D. Parak · J. Seyfried
Research Center Finance and Information Management, University of Augsburg, 86159 Augsburg, Germany
e-mail: henner.gimpel@fim-rc.de

P. Afflerbach
e-mail: patrick.afflerbach@fim-rc.de

D. Parak
e-mail: dominik.parak@fim-rc.de

J. Seyfried
e-mail: johannes.seyfried@fim-rc.de

C. van Dun
Research Center Finance & Information Management, University of Bayreuth, 95444 Bayreuth, Germany
e-mail: christopher.vandun@fim-rc.de

C. van Dun · H. Gimpel
Project Group Business and Information Systems Engineering, Fraunhofer FIT, 86159 Augsburg, Germany

1 Introduction

High-quality forecasts are essential for informed decision-making (Sanders 1997). As such, they play an important role in areas such as sales, product development, finance, and operations management (Dalrymple 1975; Mahajan and Wind 1988; Fildes and Hastings 1994; Urban et al. 1996; Slack et al. 2007). In contrast to the traditional approach of relying on single forecasts, research suggests combining multiple forecasts to improve accuracy (Clemen 1989). This applies to forecasts based on statistical models (Bates and Granger 1969; Winkler and Makridakis 1983) and forecasts drawn from human judgment (Ashton and Ashton 1985; Lawrence et al. 2006). The aggregation of multiple judgments is an important area in decision analysis research (Hurley and Lior 2002) and strongly impacts IS research (Winter 2009; Bichler et al. 2014).

As early as 1785, the Marquis de Condorcet researched the probability of a group of individuals arriving at a correct judgment and identified competence and diversity of group members as important prerequisites (de Condorcet 1785). In 1907, Galton studied aggregating judgments to

exploit the individual efforts of a crowd of people (Galton 1907; Surowiecki 2005). While individual judgments might be biased (Tversky and Kahneman 1974; Hogarth and Makridakis 1981) and individuals typically lack expertise required for making informed judgments (Van Wesep 2016), the aggregation of multiple judgments can alleviate these issues. The effects of aggregation, with the goal of outperforming individual judgments, are commonly referred to as the wisdom of crowds (WOC; see Appendix for abbreviations) phenomenon (Budescu and Chen 2015; Larrick et al. 2011). In this paper, we follow the definition of Davis-Stober et al. (2014), defining a crowd as wise if a linear combination of individual judgments is on average more accurate than the judgment of a randomly selected individual.

The aggregated judgment from a crowd can be derived via group decision processes (e.g., Kittur and Kraut 2008; Leimeister 2010; Woolley et al. 2010) or by aggregating judgments (e.g., Bates and Granger 1969; Einhorn et al. 1977; Ashton and Ashton 1985; Clemen and Winkler 1999). Looking at the latter, mathematically aggregating judgments (aggregation algorithms; also termed aggregation models) becomes an important driver of the WOC phenomenon. When examining aggregation algorithms in the context of WOC, data availability plays an important role. Performance-based algorithms (e.g., history-based algorithms as suggested by Budescu and Chen 2015) require information (e.g., previous predictions) to calculate performance measures for experts. Those information sources are so-called seed variables (Cooke and Goossens 2008). Thus, we look at a crowd of people who individually provide judgments over multiple periods and the individual judgments of the target period are aggregated into one combined judgment.

Consequently, the evaluation of aggregation algorithms places high demands on corresponding data. To fully understand the mechanics of WOC, data on internal expert characteristics (e.g., expertise, biases) as well as external context factors (e.g., volatility of the forecasted event in general and over time) is needed. These areas are partly or fully unobservable or can only be examined in laboratory settings (e.g., Palley and Soll 2019). Thus, only few researchers use empirical data (e.g., Herzog and Hertwig 2011; Wagner and Suh 2014; Budescu and Chen 2015). They focus on niche domains instead of providing domain-spanning insights due to low generalizability and comparability of results. To overcome this inadequacy of empirical data, researchers use simulation (e.g., Hastie and Kameda 2005; Hammitt and Zhang 2013; Keuschnigg and Ganser 2017). Via simulation, alternating characterizations of the crowd and the environment (i.e., scenarios) are recreated and the performance of aggregation algorithms can be studied. Although simulation-based research has been employed sparsely in the IS discipline, it has recently gained traction (Beese et al. 2019). Simulation models, like all models, are simplifications of reality. They abstract from parts of the context that is present in empirical work. This is both a strength as it enables generalization and a weakness as context is important (Davison and Martinsons 2016; Sarker 2016). Compared to theoretical and empirical analysis, simulation is recognized as a third way of doing science (Harrison et al. 2007). We see these ways as complementary and take the third way to overcome the problem of data availability in empirical investigations.

Discrete density judgments have been addressed in research (Hora et al. 2013; Park and Budescu 2015) and are used by institutions such as the European Central Bank, the Bank of England, and the Federal Reserve Bank of Philadelphia (Tay and Wallis 2000). To our knowledge, only Hammitt and Zhang (2013) have addressed the simulation of discrete density judgments. Nevertheless, the full potential of simulation of discrete density judgments has not been reached yet: There exists no simulation model providing a general framework for modeling experts with all necessary characteristics (e.g., expertise, access to information, biases, uncertainty, …) as well as events with all necessary characteristics (e.g., cues, volatility, observability, …) which can be used to implement and examine existing and new aggregation algorithms.

Guided by WOC theory and simulation-based research, we aim to close this gap and thus promote research on judgment aggregation algorithms. Specifically, based on existing models, we provide a novel model to simulate discrete expert density judgments that (1) is flexible and generalizable, (2) allows for detailed expert and event modeling along the abovementioned characteristics, and (3) can, therefore, be applied independently of domain, context, or used aggregation algorithms. The model can cope with large crowds and provides the flexibility to design versatile scenarios of experts and events. Beyond that, we compile relevant literature on the subject into propositions of the WOC effect and provide an instantiation of our model as an open-source software prototype, which is thoroughly evaluated and can be used for further research. We derive new insights into WOC in the process of evaluating the model.

Section 2 outlines the research method. Section 3 introduces the judgment setting at hand, elaborates on performance measures for evaluating judgments, describes aggregation algorithms, and closes the theoretical background by deriving propositions from WOC literature. Section 4 describes the conceptual simulation model. Sections 5 and 6 follow the evaluation process by Sargent (1987, 2005). First, we compare our conceptual model to the propositions derived from literature. Second, we verify the computerized model. Third, we validate the operational

model by demonstrating that the model leads to new research insights. Finally, Sect. 7 outlines major theoretical and managerial implications as well as limitations.

2 Research Method

Simulation is the concept of using computerized representations of processes, systems, or events to generate insights into their inner workings (Law and Kelton 2007). It has gained support as a means of generating new theory (Davis et al. 2007) and is used in WOC as well as in OR and IS in general (Harling 1958; Petrovic et al. 1998; Law and Kelton 2007; Beese et al. 2019).

The modeling process involves three components: the problem entity, the conceptual model, and the computerized model (Sargent 1987, 2005). As the correctness of the model and its results are of great concern, verification and validation play important roles (Beese et al. 2019). Sargent (1987, 2005) propose three steps: (1) conceptual model validation, (2) computerized model verification, and (3) operational validation.

We conduct the conceptual model validation by extracting relevant theory in the form of propositions on WOC and aggregation algorithms from relevant literature (Sect. 3.4). Propositions represent conceptual truths about the field of study and allow us to assess whether our conceptual model is a reasonable representation of the problem entity (Sargent 2005). Propositions described in this work are not an exhaustive list of WOC phenomena, but rather a set of properties that our model needs to possess.

Subsequently, we derive our model for static stochastic (Monte Carlo) simulation (Banks et al. 2010) in Sect. 4. We finalize the conceptual model validation by evaluating whether our simulation model behaves according to presented propositions (Sect. 5). This includes the validation techniques of *predictive validity*, *event validity*, *extreme condition tests,* and *internal validity* (Beese et al. 2019). We do this based on analytical and logical reasoning and only use simulation when necessary. Consequently, we simultaneously conduct the computerized model verification, which provides strong evidence that the implementation adequately represents the conceptual model. Thus, simulation can be utilized for validation purposes since the technical implementation is adequate. With the goal of creating a correct implementation, we utilize established program design and development approaches (modular programming, object orientation, detailed documentation, etc.) as well as the application of a well-suited programming language (Python; Oliphant, 2007). Additionally, we conduct test simulations and compare them to manually computed results from the model (Kleijnen 1995). The software code is provided open-source to allow for inspection and reuse.

Finally, the operational validation aims to determine whether the model's behavior has the accuracy required for the model's purpose (Sargent 1987, 2005). Most of the elements in the problem entity are non-observable (i.e., empirical data on expert characteristics or rarely occurring circumstances is difficult or impossible to gather). Hence, the comparison to results from empirical data is not feasible in our case. However, the purpose of this model is not to create a detailed replica of the problem entity, but rather an emulation to facilitate data acquisition for scenarios where data is unavailable. We, therefore, assess operational validity by exploring the model behavior in-depth and showing that its results provide new insights into aggregation algorithms and WOC (Sect. 6). This includes the validation techniques of *parameter variability*, *sensitivity analysis,* and *operational graphics* (Beese et al. 2019). Through evidence for applicability and usefulness, operational validity is accepted. In detail, we do this by shedding light on three sets of experimental conditions, namely by (1) exploring how aggregation algorithms weight experts, by (2) exploring the performance of aggregation algorithms under changing conditions and by (3) identifying new suppositions through experimentation.

3 Theoretical Background on the Aggregation of Expert Judgments

There are multiple terms for statements regarding unknown entities, e.g., judgments, predictions, and forecasts. While there are differences (e.g., forecasts are predictions of future entities, judgments are subjective opinions or predictions), we use judgments as the term in our paper since, for the purpose of WOC, the differences are negligible. Expert judgments and their aggregation can be carried out under different circumstances, and the dimensions in which judgment methods can be evaluated are versatile (Carbone and Armstrong 1982). Therefore, a well-defined setting (Sect. 3.1), and an adequate performance measure (Sect. 3.2) must be described. Furthermore, we present an overview of aggregation algorithms (Sect. 3.3). Finally, we derive propositions on WOC and aggregation algorithms from existing literature (Sect. 3.4).

3.1 Judgment and Aggregation Setting

The judgment task, as described in the introduction, involves a crowd of experts who individually provide judgment on a particular event. Following the origin of the WOC phenomenon (Galton 1907), we consider individual judgments (such as in Davis-Stober et al. 2014; Lee et al. 2011; Mannes et al. 2014) and do not account for group dynamics. Experts form their judgment about the event

based on their observation of cues. Experts have access to different cues of potentially different quality (based on Brunswik's lens model as in Karelaia and Hogarth 2008).

Existing literature that applies simulation in the context of WOC primarily focuses on point estimates (Hastie and Kameda 2005; Wagner and Vinaimont 2010; Mannes et al. 2014; Keuschnigg and Ganser 2017). Further approaches include rankings of alternatives (Hurley and Lior 2002) or probability judgments on binary events (Budescu and Chen 2015). Experts may also choose to provide information on the certainty of their judgment. Generally, the incorporation of this probability component is favorable in uncertain environments (Fischer 1981) and has gained popularity (Bröcker and Smith 2007). Density judgments bear the most such information as they include probabilities for all potential values of the variable in question and are, therefore, one of the most general forms of judgment (Tay and Wallis 2000). Therefore, we focus our work on discrete density judgments. An illustrative example is predicting inflation rates, e.g., next year's inflation rate for the Eurozone: A possible well-ordered and ordinal set of future values is $\{(-\infty, -0.03], (-0.03, 0], (0, 0.03], (0.03, \infty)\}$. Experts provide their judgment by assigning a probability to each interval.

To our knowledge, only Hammitt and Zhang (2013) have addressed discrete density judgments. Within their simulation model, Hammitt and Zhang (2013) assume experts to be perfectly calibrated, meaning that their individual error terms are unbiased. This assumption is contrary to established theory, stating that even experts are biased and rely on heuristics to provide judgments under uncertainty (Tversky and Kahneman 1974). For example, there is strong evidence for overconfidence in probability judgments, which interferes with the assumption of perfect calibration (e.g., Brenner et al. 1996; McKenzie et al. 2008). In addition, Hammitt and Zhang (2013) only simulate two experts, which is restrictive, as aggregation becomes especially interesting with bigger crowds.

At the point of judgment, the realization of the event cannot be witnessed. After some time, the realization becomes observable and can be compared to expert judgments for ex-post performance measurement (Hammitt and Zhang 2013). Via performance measures, quality differences between experts can be derived. If an expert has already provided previous judgments, the performance of these judgments can be considered when aggregating new judgments (e.g., as in Budescu and Chen 2015).

3.2 Performance Measurement

Judgments can be evaluated via criteria such as accuracy, ease of interpretation, cost, time, and robustness. As accuracy is the most important (Carbone and Armstrong 1982), we take a look at performance in terms of judgment accuracy. The accuracy of a judgment defines how close its estimate lies to the realized value. It can only be assessed ex-post. In most situations, decision makers may not only be interested in mean accuracy, but also in the corresponding variance. Thus, besides mean accuracy, variance is a secondary performance criterion. For density judgments, accurate judgment centers much of the probability on the realization and shows low dispersion. In decision theory, a scoring rule measures the accuracy of such probabilistic judgments (Gneiting and Raftery 2007). In general, scoring rules penalize deviations from the true set of probabilities (Bickel 2007) and can thus be used as performance measures for judgments. In this context, a proper scoring rule assigns the best score to the true probability distribution (Murphy 1970).

The Ranked Probability Score (RPS; Epstein 1969) is a commonly used proper scoring rule for measuring similarity of discrete probability distributions. To assess judgment accuracy, the RPS measures the mean squared difference between the cumulative distribution functions of the judgment and realization. Therefore, the better the prediction's calibration, the lower the RPS. Formally, it is defined as:

$$RPS = a - b \cdot \sum_{i \in I} (F_i - O_i)^2 \tag{1}$$

where I defines the ordered set of possible outcomes of the event, F_i represents the value of the cumulative distribution function of the prediction for outcome i, and O_i indicates the corresponding cumulative distribution function of the true observation (step function with 0 for values less than the realization and 1 for values equal to or greater than the realization). Without transformation, the RPS assigns zero to the best prediction (cumulative function equals step function), and $|I| - 1$ to the worst one. Via a and b, the score can be linearly transformed to a defined value range. This paper uses the RPS on a scale of 0 to 100.

3.3 Aggregation Algorithms

Galton (1907) used the median judgment to aggregate opinions of the crowd. This approach is often seen as the origin of aggregation algorithms (also known as aggregation models or aggregation rules). We differentiate approaches by three basic characteristics. First, does the algorithm rely on past predictions from each expert or other external information (history-based) or can it be used ad-hoc? Second, does the algorithm include all members of the crowd in the weighting and aggregation, or does it select a sub-set of experts from the crowd? Third, does the weighting of the selected crowd deviate from an equal weighting? While the characteristics touch upon different

aspects of aggregation algorithms, they are not independent. Weighting will typically require historical information to determine weights. Likewise, requiring historical information but not using it for selection or weighting is not sensible. Further, selection can be seen as assigning weights of zero. Despite these interdependencies rendering some combinations (namely Yes–Yes–Yes, Yes–No–Yes, No–No–No) irrelevant, we believe that these perspectives help characterize aggregation algorithms.

With the goal of selecting algorithms with differing characteristics, we selected six aggregation algorithms from literature (Table 1). Additional algorithms such as Copula algorithms (Jouini and Clemen 1996), Cooke's model (Cooke and Goossens 2008; Colson and Cooke 2017), or the newly developed pivoting (Palley and Soll 2019) could be included in the future.

Following Budescu and Chen (2015), the simple average of all expert judgments is termed Unweighted Model (UWM). In literature, names like "equally weighted mean" or "simple average" are used. For every possible outcome of an event, the UWM computes the mean of probabilities p_i assigned to it by the experts:

$$UWM: p_i = \frac{1}{|N|}\sum_{n\in N} p_{i,n}, \quad \forall i \in I \tag{2}$$

where $p_{i,n}$ is the probability assigned by expert n to the event outcome i, and N is the set of all experts. Based on the UWM, other algorithms have been developed that weight experts by including a measure of the experts' past performance. The Brier Weighted Model (Budescu and Chen 2015; Chen et al. 2016), e.g., computes the Brier Score (Brier 1950) for every expert, averaging it over all historical events. Based on this, weights (w_n) are allocated to the experts. The sum of all weights is equal to 1. The better an expert's average BS, the higher his proportionate weight. We consider the Performance Weighted Model (PWM) a generalization of the Brier Weighted Model. To use it with ordinal data, the algorithm is based on the RPS as a scoring rule for past performance.

$$PWM: p_i = \sum_{n\in N} w_n \cdot p_{i,n}, \forall i \in I, \quad \text{with}$$
$$w_n = \begin{cases} \dfrac{RPS_n}{\sum_{m\in N} RPS_m} & \text{if } \sum\limits_{m\in N} RPS_m \neq 0 \\ \dfrac{1}{|N|} & \text{otherwise} \end{cases} \tag{3}$$

The PWM only considers the absolute historical performance of expert n, described by RPS_n. As an enhancement to this, Budescu and Chen (2015) developed the Contribution Weighted Model (CWM), also known as attractivity-based weighting in philosophy (Schurz 2008). In the CWM definition by Budescu and Chen (2015), an expert's contribution is defined as the difference in aggregated performance with and without said expert. Here, the performance measure is the BS of the simple average of the crowd. The change in the crowd's performance is the difference in the BS of the crowd with and without the target expert. This difference is averaged over all historical events for each expert n, resulting in CON_n. A positive value means a positive contribution of the expert and therefore induces a positive weight, whereas a negative contribution induces a weight of 0, because the expert in question is expected to impair the judgment. We describe this with the use of the characteristic function $1_{[]}$ which is set to 1 if the condition in the subscript is satisfied, or to 0 if otherwise. Budescu and Chen (2015) also argue that apart from the BS, other scoring schemes can also be applied, which enables the application of the RPS in our paper.

Table 1 Overview of aggregation algorithms

Model name	Ad-Hoc	Selection	Weighting	Description	Source
Unweighted Model (UWM)	Yes	No	No	Simple average of all judgments	Clemen and Winkler (1986), Budescu and Chen (2015)
Random Expert Model (REM)	Yes	Yes	No	Random selection of one expert via prob. distribution	Davis-Stober et al. (2014)
Performance Weighted Model (PWM)	No	No	Yes	RPS-based weighting	Budescu (2006)
Best Expert Model (BEM)	No	Yes	No	Selection of the best (in terms of RPS) expert(s)	Hammitt and Zhang (2013)
Contribution Model (CM)	No	Yes	No	Contribution-based selection	Budescu and Chen (2015)
Contribution Weighted Model (CWM)	No	Yes	Yes	Contribution-based weighting and selection	Budescu and Chen (2015)

$$CWM: p_i = \sum_{n \in N} w_n \cdot p_{i,n}, \forall i \in I, \quad \text{with}$$
$$w_n = \begin{cases} \frac{CON_n \cdot 1_{[CON_n > 0]}}{\sum_{m \in N} CON_m \cdot 1_{[CON_m > 0]}} & \text{if } \exists m \in N : CON_m > 0 \\ \frac{1}{|N|} & \text{otherwise} \end{cases} \tag{4}$$

The CWM will ensure that experts who judged well on past events – while the rest of the crowd judged poorly – will receive high weights. The weighting in the CWM can thus be described as a measure of the relative performance of an expert in a crowd.

The so-called Contribution Model (CM) is also based on the principle of contribution as a relative performance measure. It weights all experts with a positive contribution score equally (Budescu and Chen 2015). Consequently, it produces less extreme weights compared to the CWM and does not depend as strongly on individual experts.

$$CM: p_i = \sum_{n \in N} w_n \cdot p_{i,n}, \forall i \in I,$$
$$\text{with} w_n = \begin{cases} \frac{1_{[CON_n > 0]}}{\sum_{m \in N} 1_{[CON_m > 0]}} & \text{if } \exists m \in N : CON_m > 0 \\ \frac{1}{|N|} & \text{otherwise} \end{cases} \tag{5}$$

An algorithm using more extreme weights is the Best Expert Model (BEM), also described as "imitate the best" (Schurz 2008). It only selects the expert(s) with the highest historical performance. Oftentimes this will be a single best expert obtaining weight 1 (Hammitt and Zhang 2013).

$$BEM: p_i = \sum_{n \in N} w_n \cdot p_{i,n},$$
$$\forall i \in I, \text{ with } w_n = \frac{1}{|\arg\max_{m \in N}(RPS_m)|} \tag{6}$$

For evaluation purposes, we also include the Random Expert Model (REM; Davis-Stober et al. 2014) as a benchmark. The algorithm randomly selects one expert n from the crowd via a probability distribution and weights them with 1. While Davis-Stober et al. (2014) allow for multiple distributions to be used, we use a uniform distribution to reduce complexity.

$$REM: p_i = p_{i,n}, \quad \forall i \in I \tag{7}$$

Literature has not settled on one superior aggregation algorithm and instead promotes the application of multiple algorithms (Hammitt and Zhang 2013). Opinions about the degree to which a specific weighting outperforms the unweighted mean vary. On the one hand, there is evidence that the performance of the UWM is often relatively close to that of a comparable benchmark using a non-equal weighting (e.g., Clemen and Winkler 1986; Einhorn et al. 1977; Flandoli et al. 2011). On the other hand, studies also support the superior performance of weighting-based algorithms (Cooke and Goossens 2008; Hammitt and Zhang 2013; Budescu and Chen 2015; Chen et al. 2016). Consequently, aggregation algorithms leave room for exploration.

3.4 WOC in Expert Aggregation

To build and evaluate the conceptual model, the problem entity must be understood. For that purpose, we derive propositions from literature on the behavior of WOC and corresponding aggregation algorithms. Propositions represent conceptual truths about the field of study and allow us to assess whether the conceptual model is a reasonable representation of the problem entity (Sargent 2005). In order to build a general simulation model for WOC, it is important to recreate a general understanding of the phenomenon via propositions that represent the common knowledge on WOC. Proposition 1 defines the characteristics and quality of a single expert, the basic element of WOC. Proposition 2 postulates the existence of WOC, while Propositions 3 to 5 examine the WOC effect in more detail. Finally, Propositions 6 and 7 focus on aggregation algorithms.

Proposition 1 *The optimal expert possesses all information, no bias, and no individual uncertainty.*

Experts can differ in the amount of relevant information they possess and in their ability to infer useful judgments from information. Hammitt and Zhang (2013) define expert quality with two key figures: informativeness and calibration. Experts with high informativeness form judgments with a comparatively low variance around a mean value. Calibration describes the extent to which realizations from an expert's probability distribution occur with the implied frequency (i.e., the extent to which p % of realizations actually fall within the p-percentile). A bias is a systematic displacement of the mean value and can, e.g., express extreme optimism or pessimism. Experts with a high bias are poorly calibrated. Thus, an expert's performance depends on the amount and quality of information, as well as a potential bias and variance.

Proposition 2 *The wisdom of crowds exists and is robust to the application in different scenarios and aggregation algorithms.*

Abstracting from single experts, the essential characteristic of WOC lies in improving overall judgment performance by aggregating multiple judgments and thus reducing the influence of incomplete information and biases (Surowiecki 2005). Even when members of a crowd are

biased, the aggregation of multiple judgments can make the crowd wise. Based on Davis-Stober et al. (2014) we define a crowd as wise if a linear combination of individual judgments is on average more accurate than a randomly selected individual. This holds true even for the UWM, and under unfavorable conditions, such as correlated judgments or highly and unidirectionally-biased crowds. Apart from its robustness to various scenarios (e.g., highly biased crowds), the existence of WOC is robust to different kinds of judgment aggregation approaches (Davis-Stober et al. 2014). Consequently, an aggregated judgment should, on average, be superior to a random one.

Proposition 3 *There is a linear combination of expert judgments that, on average, performs at least as good as the deterministic best expert.*

Even under extreme circumstances, it is nearly always favorable to rely on the weighted crowd or selected subcrowd than the single best individual. Davis-Stober et al. (2014) show that a linear combination of judgments is, on average, at least as good as the selection of one expert, even if this is the best expert. One explanation can be found in the bias/variance trade-off. By averaging multiple judgments, the variance of the predictions is reduced to a level that compensates for the potentially induced bias. Another reason for this is the probability of including more expertise in the judgment by aggregating multiple opinions.

Proposition 4 *On average, the performance of aggregation algorithms increases with crowd size.*

Crowd size influences the performance of aggregation algorithms. Thinking of an unbiased expert judgment as the true value plus a random error (as done by Hammitt and Zhang, 2013), according to the law of large numbers, an increasing number of experts will stabilize the aggregated judgment around the true value and decrease variance (Einhorn et al. 1977). The effect of increasing aggregation performance with increasing crowd size has been shown analytically (Hogarth, 1978), empirically (Chen et al. 2016), and via simulation (Wagner and Vinaimont 2010). However, it is important to assume that the increase in crowd size originates from randomly selected experts and not from specifically characterized experts (e.g., unqualified ones). This means that their errors are randomly distributed, i.e., there is no systematic bias in the expert population.

Proposition 5 *The more similar experts are, the harder it is to create a wise crowd.*

Apart from crowd size, other characteristics also play substantial roles. The best performance of WOC can be achieved when judgments systematically differ as much as possible (Davis-Stober et al. 2014) because this maximizes available information (Budescu 2006). Systematically differing judgments are a result of experts having access to different information sources or having different characteristics such as biases. Even if each expert holds only little information, the overall crowd might have access to all sources (Herzog and Hertwig 2011). This characteristic is called information diversity and partly explains the WOC effect. As a result, when adding a new expert to a crowd, it is on average best to choose the maximally different one from the existing crowd. This implies that experts' judgments should be collected independently (i.e., without communication between experts). Budescu and Chen (2015) have shown that the higher the similarity between experts in a crowd, the more experts are necessary to achieve the same level of judgment accuracy. Taking this to the extreme, the wisest crowd contains negatively correlated experts (Davis-Stober et al. 2014).

Proposition 6 *Much ($> 50\%$) of the advantage of weighting algorithms can be attributed to the initial selection of experts and only subordinately to subsequent weighting.*

Many aggregation algorithms (e.g., PWM, CM, CWM) use external information to impose a selection or weighting of experts. Their advantage lies in their ability to identify knowledgeable experts (Budescu and Chen 2015). A larger unweighted crowd, including good and bad experts, might be outperformed by the selection and weighting of good ones (Einhorn et al. 1977; Dana et al. 2015). That being said, Budescu and Chen (2015) remark that the quality of the CWM's performance is predicated on its efficiency in selecting the important experts in a crowd (removing all other experts from the crowd). The subsequent weighting of remaining experts only accounts for about 40% of the advantage over other algorithms.

Proposition 7 *History-based weighting profits from a large amount of seed events. The performance converges asymptotically.*

Budescu and Chen (2015) state that the CWM performs better, the more historical events are available to evaluate historical performance. This assumption also applies to other history-based algorithms (e.g., PWM, BEM). Adding past events decreases the error rate of algorithms trying to identify historical expert performance. However, performance of history-based algorithms will not increase significantly when provided with more than $\sim$ 25 historical events (Budescu, Chen, Lakshmikanth, Mellers, & Tetlock, 2016).

4 Simulation Model

Data availability plays an important role in WOC research. This makes applying simulation models particularly interesting. When examining WOC, data on the judgment of experts and the corresponding realization of the judged event are required. Consequently, our simulation model is static and stochastic (Banks et al. 2010) – also known as Monte Carlo simulation – and contains two key elements: stochastic events (which are to be judged) and experts (who provide those judgments). To simulate circumstances (called scenarios), stochastic events as well as experts are characterized by adjustable parameters, which influence the quality of the judgment and the volatility of the events. The simulation of events and expert judgments is conducted for multiple points in time to acquire the necessary history of predictions and realizations for history-based algorithms. A representation of the simulation model is depicted in Fig. 1.

An event is described as a probabilistic future incident or state. Typical examples are future stock prices, future sales of a new product, or next year's inflation rate. All these are not directly observable ex ante, but their future value is influenced by a multitude of factors. Following Hastie and Kameda (2005), we call factors that hint at the future event value cues. Examples of cues impacting the above events could be a firm's historical profits and business plan, results from a market survey, or a recent decision in monetary policy. Thus, we model an event X at time t as the weighted average of a set J of different cues $C_{t,j}$ with corresponding weights $v_{t,j} > 0$:

$$X_t = \frac{1}{\sum_{j\in J} v_{t,j}} * \sum_{j\in J} v_{t,j} * C_{t,j}, \quad C_{t,j} \sim N\left(\mu_{t,j}, \sigma^2_{t,j}\right) \tag{8}$$

Following Hastie and Kameda (2005) and Keuschnigg and Ganser (2017), we model the relation between cues and event X as a weighted average. Cues are random variables which, in our instance of the model, are normally distributed. By differing in their $\mu_{t,j}$, cues can be more or less representative of the underlying event (i.e., more or less close to the expected value of the event) and hence differ in quality. While cues and experts in reality are – more often than not – somewhat correlated (Broomell and Budescu 2009), we model cues as independent variables since inter-expert correlation can also be achieved via access to the same cues (Morris 1986).

We assume that there is one event per time step. The set of all events is thus defined as $X = \{X_{-T}, X_{-T+1}, \ldots, X_0\}$. The events are not correlated. The events X_{-T} to X_{-1} are called seed events and represent events that have already occurred in the past. Their realizations and judgment data are already fully available. Target event X_0 is to be judged.

In general, there are three possible ways of how experts make judgments. Experts can provide point estimates (de Menezes et al. 2000), interval estimates (e.g., confidence intervals; McKenzie et al. 2008), or a discrete probability distribution (Yates et al. 1991). We apply the latter, which is extensively addressed in research (e.g., Clemen and Winkler 1999; Genest and Zidek 1986; Hammitt and Zhang 2013), or practical forecasting applications (e.g., European Central Bank 2017). Consequently, we select the RPS as an adequate scoring rule for measuring judgment performance.

The second key model component are experts. Let N denote the set of experts. Our simulated experts can differ in three aspects: whether or not they have access to some or all cues related to an event, their individual uncertainty, determining the width of the individual probability distribution, and a bias, which affects the mean of the distribution. Access to cues means that experts know about the realized value of cue $C_{t,j}$. Hence, experts form judgments by calculating the weighted average of all realized cues

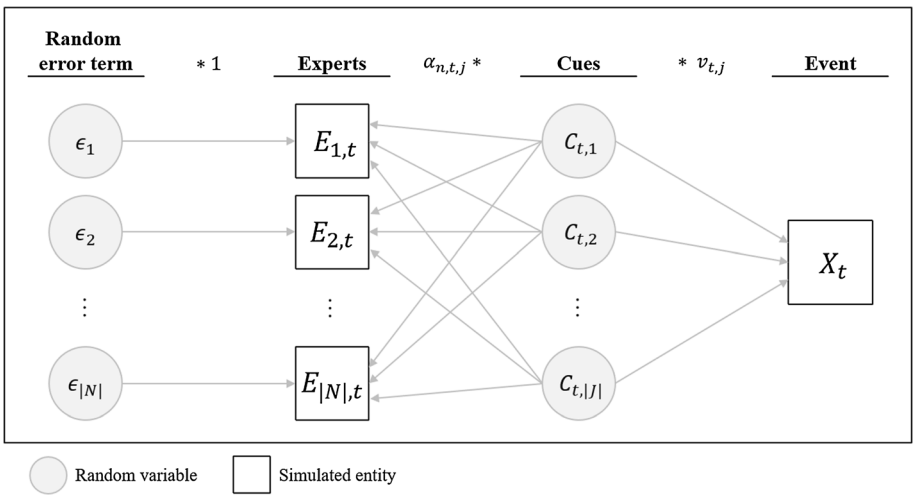

Fig. 1 Overview of event and expert simulation model for time t

known to them, while ignoring cues they do not know about. Experts might not accurately perceive or process the informational cues, thereby adding a random error parameterized by bias (mean $\neq 0$) and uncertainty (variance > 0). Access to a cue is described by $\alpha_{n,t,j}$. If expert n observes cue $C_{t,j}$, $\alpha_{n,t,j}$ defines the weight the expert allocates to the cue. Otherwise, it is 0. The random error can be modeled with a probability distribution. Following Hammitt and Zhang (2013), we use normal distributions as an example: The uncertainty is described by variance σ_n^2 of the distribution, while the bias is the offset μ_n. Adding up these requirements to a stochastic formula, the judgment $E_{n,t}$ of expert n for the event at time t is modeled as follows:

$$E_{n,t} = \left(\frac{1}{\sum_{j \in J} \alpha_{n,t,j}} * \sum_{j \in J} \alpha_{n,t,j} * C_{t,j} \right) + \varepsilon_n, \quad \text{with} \varepsilon_n \sim N\left(\mu_n, \sigma_n^2\right) \tag{9}$$

The set I of numerical intervals is defined freely within the range of possible outcomes. To simulate a discrete probability distribution (probabilities for a well-ordered set of intervals), we draw multiple times upon the expert's probabilistic judgment $E_{n,t}$ and calculate the relative frequency of a hit in an interval.

As in all Monte Carlo simulations, the procedure of deriving judgments must be repeated multiple times per scenario in order to create a sufficiently stable probability distribution that can be used to assess the outcome.

A common empirical analysis would consider effect size and statistical significance to appraise statistical relationships. While effect size is important in our model, statistical significance is less so. The methods of frequentist statistical hypothesis testing were designed for low-replication empirical data – they are inappropriate when comparing outputs of simulation models (White et al. 2014). One reason is that, for a given effect size, p-values depend on the number of replications under analysis, which can be arbitrarily high in simulation. This can produce minuscule p-values regardless of the effect size (White et al. 2014). Excessive sample size increases "the sensitivity of statistical tests possibly to the point of absurdity" and surfaces statistically significant results on contextually inconsequential differences (Lee et al. 2015, paragraph 2.2). Thus, we suggest setting the number of simulation runs per scenario sufficiently large for obtaining meaningful estimates of outcome distributions and effect sizes but to refrain from significance testing (Lee et al. 2015). Different metrics for variance stability may be employed for determining minimum sample sizes, that is, the number of required simulation runs – e.g., confidence interval bound variance (Law and Kelton 2007), coefficient of variation (Lorscheid et al. 2012), or windowed variance (Lee et al. 2015). These procedures may be applied to all scenarios under consideration and the required number of simulation runs is the maximum deriving from any of these scenarios. Once this minimum number of simulation runs is determined, one should not test for significance of results but merely interpret the contextual significance of differences.

In the following, if not specifically stated otherwise, we reduce the degrees of freedom in our model to limit the complexity of the simulation: (1) we do not change events or the experts' access to information over time, meaning that the $\alpha_{n,t,j}$ and $v_{t,j}$ stay constant with changing t, (2) we use unweighted averages of cues both for events and experts, meaning that all $v_{t,j}$ are 1 and all $\alpha_{n,t,j}$ are either 0 or 1. In other words, experts do not learn over time and do not weight cues.

For evaluation, we focus on three types of scenarios that represent different stylized configurations of crowds. For illustration, we use the notation of a $|N| \times |J|$ matrix A_t containing the $\alpha_{n,t,j}$:

$$\mathrm{A}_t = \begin{pmatrix} \alpha_{1,t,1} & \cdots & \alpha_{1,t,j} \\ \vdots & \ddots & \vdots \\ \alpha_{n,t,1} & \cdots & \alpha_{n,t,j} \end{pmatrix} \forall t \tag{10}$$

The first symbolic scenario contains experts that all have access to different cues. They do not share access to cues. Instead, each expert owns a different piece of information. In matrix notation, this generates a diagonal matrix ($|N|$ experts, $|J| = |N|$ cues). Our second scenario represents experts with varying levels of expertise. The best-informed expert has access to all cues, while the worst-informed expert has no cues. This manifests in a triangular matrix with an additional row of zeros for the uninformed expert ($|N| = |J| + 1$). Finally, we consider so-called information clusters: We assume that groups of several experts share the same cues and therefore form clusters of similar knowledge. A matrix representation of this case would contain several well-defined areas of ones and zeros and will henceforth be called cluster matrix.

5 Conceptual and Computerized Model Validation

Validating the conceptual model involves comparing it to the corresponding problem entity in order to determine whether the model adequately represents commonly accepted characteristics. To answer this, we show that the propositions derived from literature (Sect. 3.4) hold within our model. We partially validate the conceptual model via analytical reasoning, and indirectly via simulation. This will show that the simulation model is valid as a representation of the problem entity and as a means of understanding the characteristics of aggregation algorithms and WOC in general. In the following, we assume for ease and

brevity that events are unweighted averages of cues ($v_{t,j} = v_{t,k} \forall j,k \in J$) and that experts are aware of this (i.e., they only estimate unweighted averages). This reduces the number of scenarios and hence limits computational complexity while allowing us to vary the expert's information level via access to the cues.

Proposition 1 states that the optimal expert possesses all information, no bias, and no uncertainty. An expert is considered optimal if he always allocates a 100% probability to the interval containing the future realization of the event. Consider two experts: A and B, who have complete information ($\alpha_{n,t,j} = 1$) and no bias ($\mu_A = 0; \mu_B = 0$). The uncertainty of A is lower than that of B ($\sigma_A^2 < \sigma_B^2$). From lower uncertainty, it follows that on average, A's allocated probabilities will scatter less, and A will assign more probability to intervals close to the mean (i.e., the realization of the event). Consequently, A is better than B. Now consider new characteristics for A and B: Both have no bias and uncertainty, but A has access to more cues than B. Since the realization of the event is the average of all cues, access to more cues increases the probability of being close to the realization. Therefore, A is better than B. Finally, consider A and B as experts with all information and no uncertainty. When A is less biased than B, A will be closer to the realization. Again, A is better than B. We can conclude that an expert with less uncertainty, less bias, and access to more information is generally better. Proposition 1 holds for our model since the optimal expert must possess all available cues ($\alpha_{n,t,j} = 1$), no bias ($\mu = 0$), and no individual uncertainty ($\sigma = 0$).

Proposition 2 does not focus on individual expertise, but rather on the existence of crowd wisdom. Based on Davis-Stober et al. (2014) we define a crowd as wise if a linear combination of individual judgments is, on average, more accurate than randomly selected individuals. We want to show that aggregation algorithms (like UWM, PWM, CWM, and CM) are, on average, more accurate than randomly drawn experts from the crowd (REM). Looking at a crowd of N experts, it is fair to assume that they infer their judgment based on at least partly different cues ($\exists \alpha_{n,t,j} = 0$). Thus, by aggregating judgments of multiple experts, more cues are considered than for a single randomly selected expert, and the overall judgment becomes more informed. Similar effects are caused by the expert-specific error term. By aggregating judgments of multiple experts and thus aggregating their error terms, the overall deviation from the value implied by the cues is reduced because the standard deviation of an average of independently distributed random variables is smaller than that of a single random variable. As a result, in our model, the aggregation of multiple experts improves judgment accuracy and leads to wise crowd-based judgments. We can demonstrate the validity and robustness of this property by simulating several different scenarios, measuring the average RPS performance of aggregation algorithms. We use scenarios where we vary expertise, uncertainty or bias. Aggregation algorithm performances in exemplary scenarios are depicted in Fig. 2. The RPS scores for all algorithms that aggregate multiple experts (UWM, PWM, CWM, and CM) are, on average, higher than that of a random expert (REM). Thus, we can show that the wisdom of crowds exists in our model and is robust in a wide range of scenarios and aggregation algorithms. However, extreme scenarios do exist where the REM outperforms other aggregation algorithms.

A stronger assumption is formulated in Proposition 3, which suggests that for every scenario, there is some linear aggregation of judgments that, on average, performs at least as good as not only a random expert, but as the deterministic best expert. Via Jensen's inequality, Davis-Stober et al. (2014) have proven that this proposition holds in theory. To test it for our simulation, we specify $w = (w_1, w_2, \ldots, w_n)$ as the vector of weights assigned to experts N while linearly aggregating their judgments. Without loss of generality, we assume the deterministic best expert to be weighted with w_1. Then, the selection of the best expert results in $w_{BEM} = (1, 0, \ldots, 0)$. Consider an extreme scenario where one expert holds all information while all other experts are badly calibrated and uninformed. Here, using w_{BEM} as aggregation will, on average, outperform all other aggregation algorithms. However, this is an artificial scenario. It is reasonable to assume that a best expert in a realistic scenario is not holding all information and is therefore not judging perfectly, as this is highly unlikely in the real world. Such scenarios contain no perfect experts and more than one expert holds relevant information. Optimal weights will deviate from w_{BEM} towards a more equal weighting, thus outperforming the deterministic best expert.

Proposition 4 assumes that increasing crowd size impacts the performance of WOC positively. Imagine a scenario with $|N|$ randomly characterized experts and $|J|$ cues. Independently of all other $|J| - 1$ cues, the probability p_j of having access to one particular cue j is the same for every expert and greater than zero. Therefore, with probability $(1 - p_j)^{|N|}$, the overall crowd does not have access to the cue. If we now add another randomly characterized expert, the probability of adding that particular cue to the pool of available information for the first time is $(1 - p_j)^{|N|} \cdot p_j > 0$. This implies that with positive probability a new expert is valuable to the crowd since he might add new cues to the crowd's knowledge base. If not, he is not useful as a carrier of new information but can still reduce overall variance of the aggregated judgment since

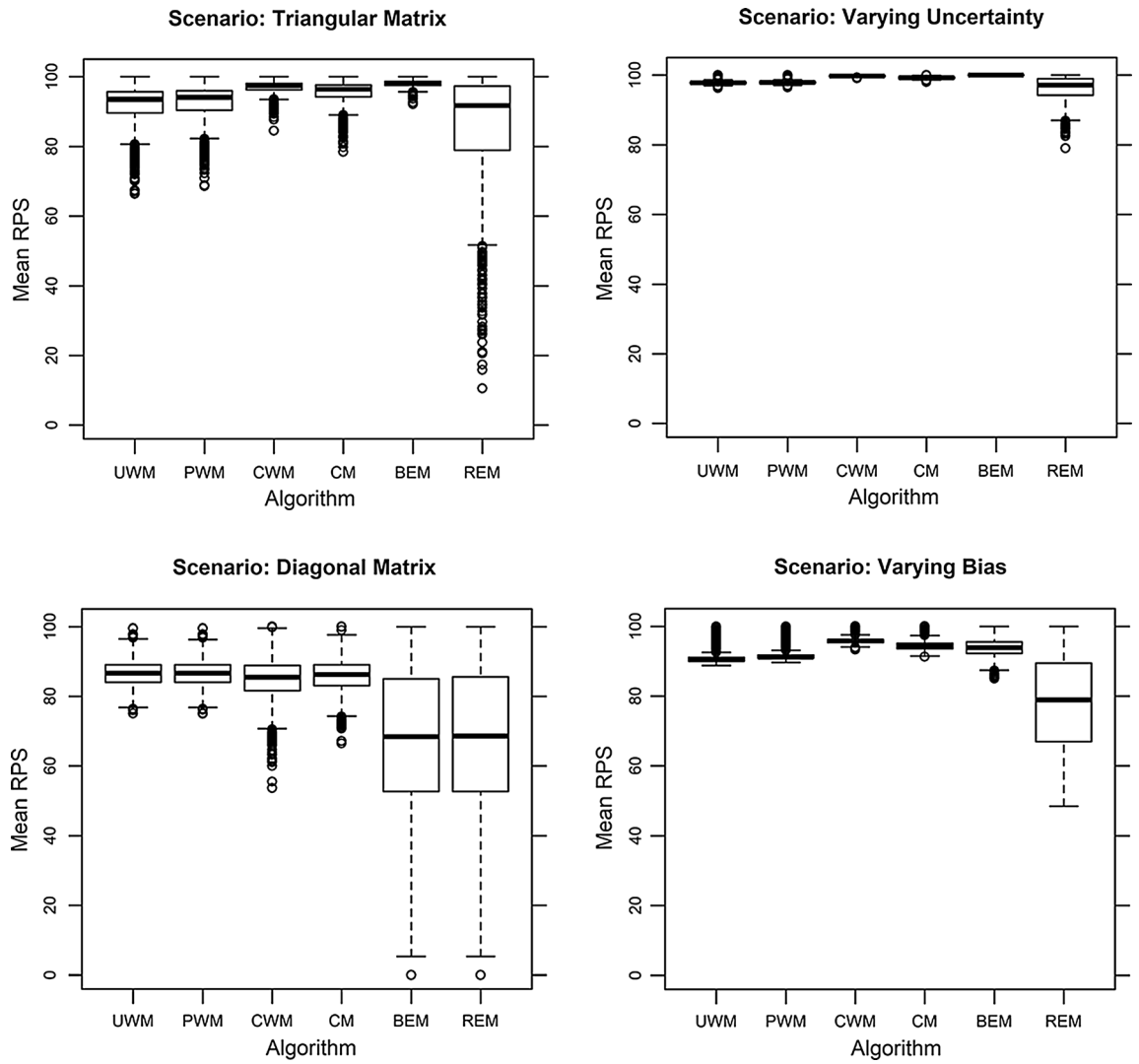

Fig. 2 Aggregation algorithm performance in different scenarios

we assume no systematic bias in the expert population. In extreme cases only, experts can decrease the crowd's performance (e.g., by being heavily biased). Altogether, a new expert generally increases crowd performance by adding new information or reducing judgment variance. The effect diminishes with increasing crowd size.

Proposition 5 describes the assumption that WOC is based on maximizing available information; it suggests that aggregation algorithm performance is better when acting on heterogeneous crowds. A heterogeneous crowd contains differently characterized experts (i.e., experts that have access to different cues). This means that the crowd has access to more cues overall, while a homogeneous crowd only has access to a limited information pool. As in Proposition 4, we reason that every expert added to the crowd has a positive probability of adding new cues to the crowd and thus increasing performance if not all cues are already available in the crowd. If all cues are available, new experts might still diversify the crowd's error.

Proposition 6 suggests that weighting algorithms benefit primarily from selecting knowledgeable experts and only subordinately from subsequent weighting. As such, selecting experts is generally more important than trying to additionally weight them according to their level of expertise. Via simulation, we can confirm that the performance advantage of weighting algorithms can largely be attributed to the selection of experts. We look at many different scenarios where there is heterogeneity of expertise in the crowd and use the CM as a modification of the CWM with equal weights for the selected experts. The CM's performance is, on average, closer to the CWM's performance than to the UWM's (Fig. 2). From this, we conclude that selection is more important than the actual weighting of remaining experts. Therefore, the proposition holds.

Table 2 Mean and variance of RPS scores for the CWM in scenarios with different seed amounts

Number of seeds	5	15	25	50
Mean RPS	96.459	96.830	96.875	96.904
Variance RPS	4.957	3.323	2.964	2.871

Finally, Proposition 7 focuses on the algorithms' ability to extract information on expert performance from historical events. We assume that the performance of history-based algorithms generally increases with the amount of available seed events (i.e., the mean RPS will increase, or the variance will decrease). Additionally, we expect the performance to converge asymptotically with increasing seed events because the measurement of an expert's historical performance will stabilize. We test scenarios with 5, 15, 25, and 50 seed events and compare the CWM's resulting performance measurements (Table 2). In particular, the decreasing and simultaneously converging variance of the performance supports our assumption.

In sum, all WOC propositions hold true within our model. Therefore, it is reasonable to assume that the conceptual model is valid.

Validating the correctness of the computerized model requires assurance that the programming and implementation of the conceptual model are correct (Sargent 1987, 2005; Kleijnen 1995). The computerized implementation of the model has been designed and implemented in a top-down approach using standard software design and development procedures. It is implemented in the general-purpose programming language Python, which is often used in statistics and simulation (Oliphant 2007). Every component and function of the conceptual model has been mapped to separate modules in the computerized model, thereby ensuring program modularity. Every module has been tested: First, all necessary simulation functions have been executed with dummy scenarios. Afterward, individual modules and the whole model have been tested using static as well as dynamic testing approaches. Their output was compared to manually computed results of the conceptual model. We can conclude from positive results that the computerized model is representing the conceptual model correctly. All information has been documented to assert future expandability. The use of the computerized model for evaluating the conceptual model with respect to selected propositions further supports the validity of the computerized model. The software is provided as open-source code to allow for further validation and reuse of our computerized model[1].

[1] https://github.com/chaOtis/simulating-woc/

6 Operational Model Validation

Operational validity is concerned with examining the model's applicability by ensuring that it creates accurate results that are useful for the intended purpose. This section provides evidence for applicability and usefulness. We show that the simulation model can be used to investigate the behavior of aggregation algorithms (Sect. 6.1), to assess the performance of aggregation algorithms under circumstances that are hard to investigate empirically (Sect. 6.2), and to explore suppositions for further research (Sect. 6.3).

We use the simulation model to conduct experiments by constructing scenarios and measuring the behavior of aggregation algorithms. For this purpose, we define a reference setting for experimental scenarios consisting of seed variables, outcome intervals I, and the number of simulation runs; it defines the basic setting that we use for all experiments (unless specified otherwise), which ensures comparability of different scenarios. The algorithms can rely on 25 seed variables for each expert. We derive selectable intervals from the range of the event distribution $X_t \sim N(\mu_t, \sigma_t^2)$. Of all possible intervals, $|I| - 2$ intervals are equidistantly distributed within $[\mu_t - 2\sigma_t, \mu_t + 2\sigma_t]$, and two remaining intervals are open intervals towards $-\infty$ and $+\infty$, respectively. Again, we assume events to be unweighted averages of cues. We determine the minimum number of necessary simulation runs to obtain sufficiently stable distributions with the windowed variance method (Lee et al. 2015). If results are compared between different scenarios, 10,000 cycles are sufficient; if not, 3000 cycles are sufficient. The event specifications, the size of the crowd, and individual characteristics are defined per scenario, as they fundamentally define experiments. This allows us to create a range of scenarios to examine WOC and, with it, the simulation model's ability to derive new research findings.

6.1 Understanding Aggregation Algorithms in Depth – Expertise Diversity and Seed Events

One application of simulation is to further understand the particulars of aggregation algorithms. As the inner workings of aggregation algorithms are difficult to understand from the outside, a deeper analysis is required (Clemen and Winkler 1986). When creating decision models based on aggregation algorithms and scenarios, it is crucial to understand which algorithm will work best (e.g., Hammitt and Zhang 2013).

We create two scenarios, fashioned to illustrate discrepancies in the aggregation algorithms' weighting. Each scenario consists of ten experts who only differ in their

access to cues. The information matrix of one scenario is triangular (i.e., expert $n \in \{1, \ldots, 10\}$ has access to exactly n out of ten cues) while the matrix of the second scenario is diagonal (i.e., each expert has access to a different one of the cues). The triangular matrix portrays a heterogeneous distribution of expertise in the crowd, while the diagonal matrix depicts similar experts in terms of expertise. Figure 3 shows the cumulative average weights for both scenarios as a function of the share of experts. For example, 20% of experts in the triangular scenario possess $\sim$ 70% of weights when using the CWM. A steep incline in the curve signals the allocation of substantial weight to few experts.

Since the UWM distributes equal weights to experts independent of scenario characteristics, the cumulative weights always proceed linearly. The PWM assesses historical performance based on the RPS and assigns weights accordingly. Under heterogeneous expertise, this leads to a slightly unequal weighting. The CWM and CM both select experts. Thus, the full weight is allocated to a subset of experts. This effect is stronger for the CWM than for the CM since it also weights the selected experts. The BEM is not displayed here as it only selects one expert in every application. Figure 4 shows the corresponding performance scores. As high performance and low variance are desirable, it suggests a clear ranking of aggregation algorithms for the triangular matrix scenario with BEM being best, followed by CWM, CM, PWM, and UWM as worst. In other words: The more unequal the weighting, the better the performance in this scenario.

Expertise and experts' weights are clearly heterogeneous in the triangular scenario. However, in the diagonal scenario, on average, all experts show equal performance, and no superior one can be found. Consequently, weighting algorithms (PWM, CWM, CM) compute an equal weighting for all experts (Fig. 3) and achieve mean performance scores similar to the UWM (Fig. 4). However, while performance is comparable, the variance of weighting algorithms is higher in scenarios with little differentiation in the expertise.

We conclude that the performance of weighting algorithms depends on a certain variation in a crowd's expertise. For similar experts, aggregation algorithms using performance measures can even perform worse than equally weighting algorithms since they falsely introduce weighting, despite no expert having superior expertise. Furthermore, the performance of the CM is strongly related to that of the CWM. Both algorithms benefit from diverse crowds and will generally outperform the UWM if there is special expertise within the crowd. The BEM performs strongly for crowds including well-informed experts and poorly otherwise.

In a second step, we inspect aggregation algorithms' dependence on seed events. Apart from the U21WM, all selected algorithms depend on identifying good experts based on historical performance. Consequently, the number of observable seed events influences algorithm performance (Cooke and Goossens 2008; Eggstaff et al. 2014, Budescu 2006). However, a deeper understanding of the coherences, especially concerning the CWM, still needs to be obtained.

We investigate the necessary number of seed variables by analyzing the standard deviation of the CWM's weighting in scenarios where different amounts of seed events are available. We define the standard deviation of the weighting as the standard deviation of an expert's weight across simulation runs, averaged across experts. Low standard deviation signals that aggregation algorithms reliably calculate almost the same weights in every run. This indicates that enough seed events are present for algorithms to form a stable weighting. We focus on two scenarios, each containing five experts: a diagonal matrix and a matrix with two fully informed and three uninformed experts (cluster matrix). To limit complexity, we use powers of 2 as seed amounts: 4, 8, 16, 32, 64, 128, and 256. Higher numbers will seldomly occur in a real-world context.

Table 3 shows the standard deviation of the CWM's weighting for each scenario and seed amount. The standard deviations are generally lower for the cluster matrix because the uninformed experts are mostly deselected by the CWM, and their weights seldomly deviate from zero. The data shows that quadrupling the number of seed variables decreases the standard deviations by at least 30%, on average even by 45%. Not surprisingly, more seed events are better for the CWM.

Even for 256 seed events, the weights calculated by the CWM in the diagonal matrix scenario vary substantially more than for only 4 seed events in the block matrix scenario. Thus, when judging the reliability of weights, the diversity of expertise appears more important than the number of seed events.

The results indicate that diversity of expertise is essential for weighting aggregation algorithms to work, that more seed events are better, but that diversity trumps number of seed events.

6.2 Evaluating Algorithm Performance – Structural Breaks

The aggregation of judgments is concerned with improving judgment accuracy. In general, literature on WOC and aggregating judgments is mostly concerned with quantitatively assessing the aggregation algorithms' performance (Clemen 1989). This assessment is foremost conducted

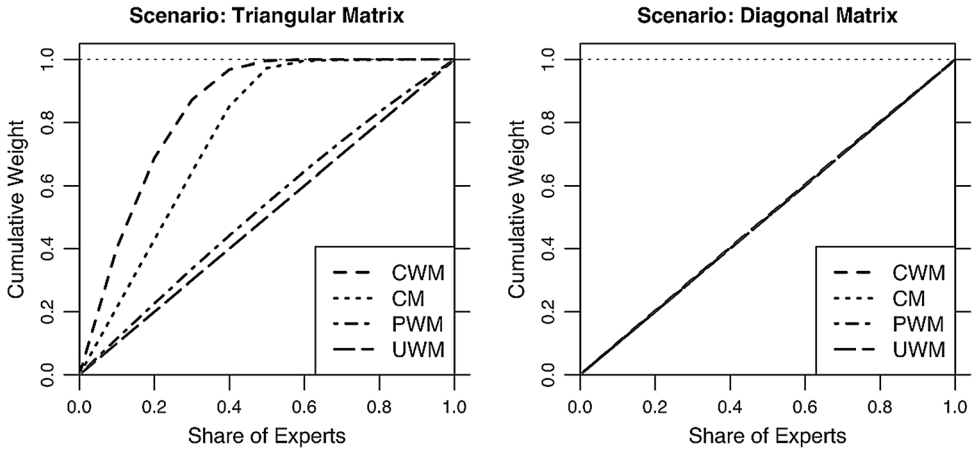

Fig. 3 Cumulative weighting in two different scenarios

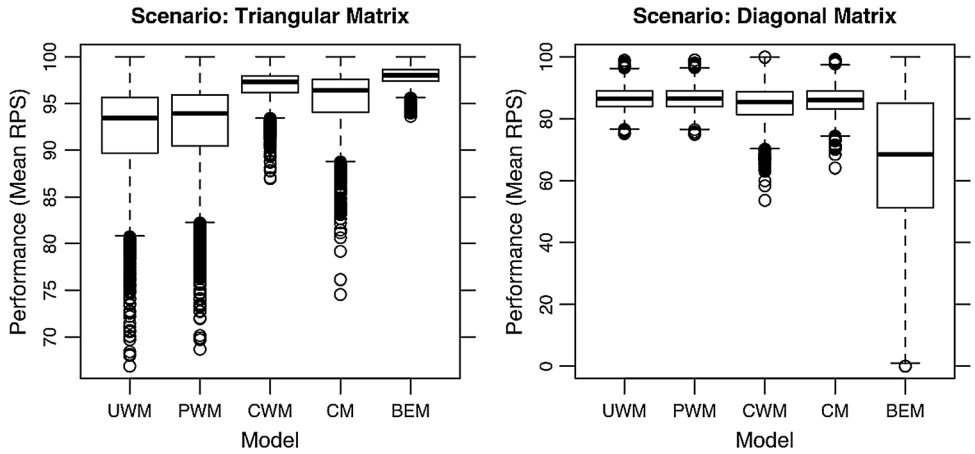

Fig. 4 Boxplot of RPS values for two different scenarios

Table 3 Standard deviation of the CWM's weighting, depending on the number of seeds available

Number of seeds	4	8	16	32	64	128	256
Diagonal Matrix	0.168	0.147	0.112	0.081	0.060	0.050	0.042
Cluster Matrix	0.025	0.020	0.006	0.004	0.003	0.002	0.002

empirically (e.g., Budescu 2006; Clemen and Winkler 1999; Cooke and Goossens 2008). This is reasonable for evaluating performance under externally given circumstances. However, when measuring performance of algorithms for specific circumstances, this approach reaches its limits. A general example for such circumstances are structural breaks in time series, i.e., situations where underlying characteristics (e.g., that describe an industry) change fundamentally and remain in this new state. With quickly changing technological landscapes, fast-moving industries and volatile global financial markets, structural breaks are especially relevant in practice.

The underlying assumption of history-based aggregation algorithms is that experts who performed well in the past are likely to perform well in the future. With structural breaks, this hypothesis might not be true. Consider experts providing judgments on which technology will emerge as new market leader. Experts in this market might be clustered in groups: Experts in group 1 bet on the success of the incumbent technology, while experts of group 2 bet on the emerging technology's success. While the emerging technology is still a niche product, experts favoring the incumbent will deliver accurate judgments. Yet as soon as the emerging technology has its breakthrough, it rapidly gains market share and eventually replaces the incumbent. This break can be simulated by changing cue properties. Since this usually happens quickly and experts tend to stick to their judgments, group 1 will now deliver inaccurate

judgments, while group 2 delivers accurate ones. A famous example for such a structural break was the rise of digital photography (Lucas and Goh 2009).

Since history-based aggregation algorithms use existing seeds to weight experts, the point in time of the structural break impacts aggregated judgments. We evaluate algorithm performance depending on the time of the structural break. Figure 5 shows the mean RPS of the algorithms as a function of the break's time when simulating a scenario such as the one described above.

The first structural break (at $t = -25$) is equivalent to the information switch before the first period of the simulation model's history; thus, algorithms only observe experts providing their post-structural-break estimates. Thus, the algorithms behave as if there were no structural break. On the other side of the spectrum (break at $t = -1$), the algorithms only observe one historical period that occurs after the structural break.

Figure. 5 clearly shows that the performance of the history-independent UWM remains constant over time. As expected, history-dependent algorithms show a decrease in performance the later the structural break takes place. Among these algorithms, the point in time and the extent of the impact on the performance substantially differ. The BEM is the first algorithm to show a substantial drop in performance, followed by CWM and CM. The PWM performs similar to the UWM and is characterized by a constant decrease. Comparing the algorithms to the UWM as benchmark, the performance increase of history-based algorithms in case of an early structural break is far lower than the performance decrease in case of a late structural break. Furthermore, for early structural breaks, the history-based algorithms seem to be very close to each other, and the CWM can hold its performance advantage against the UWM longest.

The performance decrease of history-based algorithms is tied to their weighting. The later the structural break takes place, the more pre-structural break information is included in the weights. Since the experts' performance switches, the included information is flawed, and the algorithms allocate above-average weights to experts who perform worse. This leads to a decrease in performance. The extent to which algorithms react to structural breaks thus depends on the strength of their weighting. Depending on the algorithms' specific weighting logic, the intensity of the weighting differs substantially.

Moreover, the simulation brought unexpected behavior of the CWM to light. Looking at periods t = − 25 to t = − 19, the RPS increases. This behavior is unexpected since the algorithm can access the most representative data on the experts if the structural break takes place at $t = -25$; thus, one would expect performance to be highest there. The later the structural break takes place, the more flawed information is incorporated into the calculations of the weights. Taking a closer look at the weights supports the assumption that flawed information leads to less extreme weighting which results in more moderate judgments and increases the average performance. Thus, it appears that in periods $t = -25$ to -19, the CWM suffers from overfitting and benefits from a slight reversal.

Generally, history-based weighting allows for a high possibility of long-term success, coupled with risk of short-term errors in volatile scenarios. These findings have not yet been established empirically, presumably as structural breaks only occur seldomly, especially in combination with available judgement data. Future research should address the underlying question of how to balance short-term versus long-term success.

6.3 Exploring New Suppositions

Experimentation is a core application of simulation models. Effective experimentation supports discovering new theory (Davis et al. 2007). Simulation methods enable

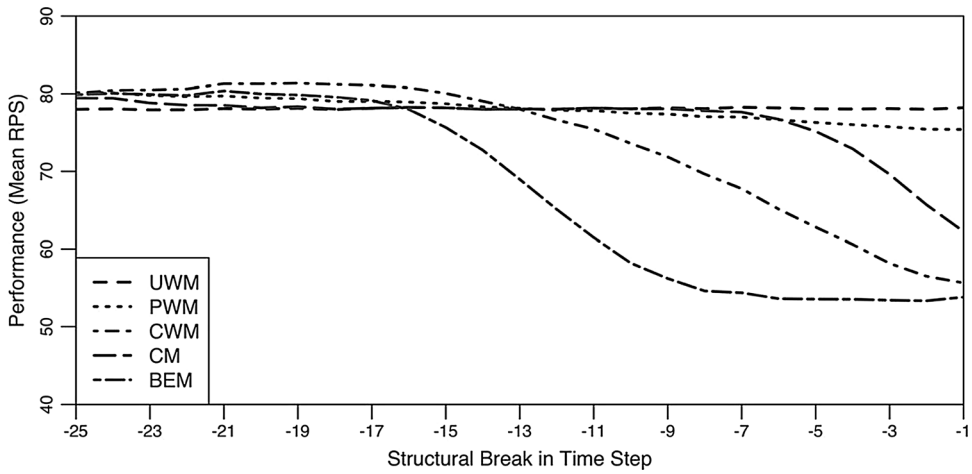

Fig. 5 Impact of structural breaks on the performance of aggregation algorithms

experimentation across a wide range of conditions. By varying assumptions and values in our model, we identified two new suppositions that demand further exploration. These suppositions are a first step in establishing new theory in the WOC field and focus on the optimal composition and characterization of expert crowds. We purposefully call them suppositions to set them apart from aforementioned propositions and from hypotheses as these are typically used in empirical analyses. First, we address a specific issue of the CWM, which can lead to flawed assessments of expert performance and impair the CWM's judgment performance. Second, we examine the expert-specific as well as the crowd's overall uncertainty and try to identify conditions for optimality.

The CWM measures expert performance relative to the crowd's performance. Therefore, even reasonably good experts can be deselected. Also, a reasonably uninformed expert can increase the crowd's performance by balancing a bias held by the majority of experts (i.e., by bracketing the true value; Larrick and Soll 2006; Herzog and Hertwig 2009) and might, therefore, be selected. Imagine a scenario with five experts and three cues $\{c_{t,1}, c_{t,2}, c_{t,3}\}$. Let one expert have access to $c_{t,1}$. The other four experts are reasonably well-informed, but similar to one another (access to $c_{t,2}$ and $c_{t,3}$ each). Thus, the first expert has exclusive access to $c_{t,1}$. In such a scenario, we expect the CWM to distribute much of the weight among the four well-informed experts, but still select the first expert because of his access to a cue that is rarely observed. However, simulation results in CWM weights and performance scores as depicted in Fig. 6.

In 84% of runs, the CWM allocates a weight of 1 to expert 1, while well-informed experts (experts 2-5) are weighted with 0. Consequently, we see that the performance of the CWM and CM is considerably lower than that of other algorithms, as it mostly only considers the uninformed expert's judgment (Fig. 6).

This might happen for two reasons. First and foremost, the contribution score of expert 1 to the crowd is extremely positive as he adds information about a rare cue. This makes selecting him and allocating a relatively high weight reasonable. Secondly, the contributions of experts 2–5, as calculated by the CWM formulation of Budescu and Chen (2015), are mostly negative. Since they are similar to each other, excluding one of them from the crowd will considerably increase crowd performance because it will lower the excess weight of said experts in an unweighted mean. A negative contribution leads to the deselection of the respective expert. Thus, the effect appears due to the isolated perspective of the CWM on a single expert's performance relative to the crowd. It becomes stronger the more similarly experts are characterized and the stronger groups of similar experts are in a crowd. Simultaneously, when experts are characterized diversely, the effect will disappear. Of course, we show an artificial scenario with only five experts. However, the effect holds true to a somewhat lesser extent in scenarios with additional experts. We therefore state our first supposition as follows:

Supposition 1: High similarity of experts in a crowd can lead to the CWM deselecting said experts which in turn leads to an unfavorable forecasting performance.

In a second experiment, we inspect the experts' individual uncertainty σ_n and how it affects judgment performance. We focus on the optimal individual uncertainty (i.e., the uncertainty value that maximizes an expert's individual judgment performance). First, which factors influence the value of the optimal uncertainty, and how strong is the impact of deviating from the optimal value on expert performance? We build a scenario with three cues ($c_{t,j} \sim N(0,1) \forall j \in J = \{1,2,3\}$). We use a brute-force approach to compute the optimal individual uncertainty for an expert while varying the expert-specific bias μ_n and the number of available cues as described by $\sum_{j=1}^{|J|} \alpha_{n,t,j}$. Subsequently, we let the expert deviate from this optimal value to see how strong the impact of uncertainty is on expert performance. The optimal uncertainty values for each set of parameters as well as the performance scores are shown in Fig. 7. μ_n is defined within reasonable borders.

Optimal uncertainty become lower the better the expert is calibrated. An expert with access to all cues and no bias will perform best if his uncertainty is 0, as this will nullify his error term (see Proposition 1). When deviating in both parameter dimensions (bias, cue access), the optimal uncertainty values become higher. In scenarios where an expert's stochastic judgment is far from the event realization through bias or missing cues, the expert benefits from variance in the error term. This is explained by the nature of the error term: It scatters in both directions. Thus, it might cancel out the deviation from the real value. With the complementary probability, it will increase the deviation. However, the impact of this complementary event is limited as the last interval in each direction is open towards infinity, and thus does not penalize extreme deviations. The negative impact of deviation from the optimal uncertainty on performance becomes stronger, the better an expert is otherwise calibrated.

Supposition 2: Well-calibrated experts perform best if their individual uncertainty is low, while badly calibrated experts can profit from a higher individual uncertainty that can cancel out their bias.

As before, these effects have not yet been described in empirical data, presumably because it is difficult or

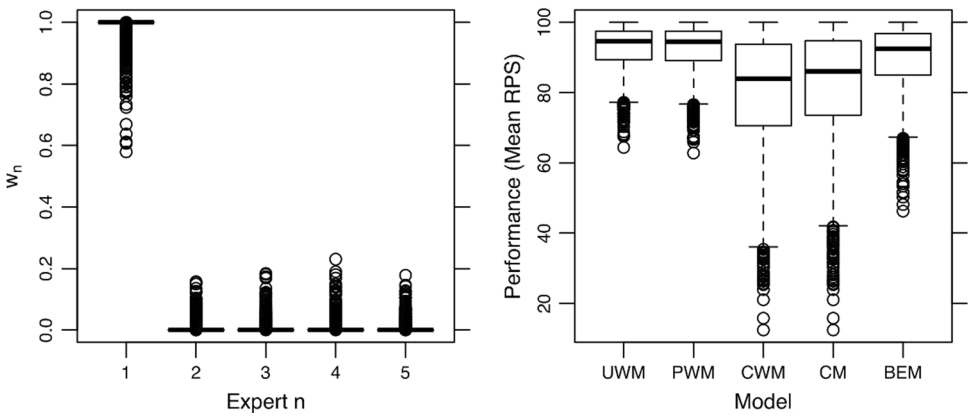

Fig. 6 Weighting phenomenon of the CWM

impossible to disentangle justified judgment form idiosyncratic error in real-life settings.

7 Discussion and Conclusion

Simulation is an important toolkit in WOC research as data availability is a limiting factor. We propose a novel model to simulate expert density judgments, with the aim of shedding light on expert judgment, aggregation algorithms and WOC in general. To do so, we first deduct propositions on WOC from literature and design a model to simulate WOC scenarios. After completing all verification steps, we conclude that the model and its implementation are valid representations of the real-world problem entity. With its help, we gain exemplary new insights into WOC.

This paper contributes to WOC research with four major aspects. First and foremost, the conceptual simulation model is a novel representation of experts providing discrete density judgments. While institutions like the European Central Bank are using density judgments as forecasting input, there is currently no simulation model available for this form of judgment. Our model sets itself apart from judgment models such as Hammitt and Zhang (2013) who have only incorporated two experts with distinct characteristics.

Second, we compile relevant literature into propositions on WOC. With their help, it is possible to reach a deeper understanding of WOC and its characteristics. The propositions are designed to act as a foundation for further research and can be utilized as verification criteria for models with similar backgrounds.

We show that the model is applicable and valid by creating a computerized implementation and conducting validation and verification steps based on an established framework. Researchers can employ the instantiation to produce findings in the field of WOC. For example, the model supports iteratively specifying and testing new aggregation algorithms under a variety of potential circumstances. It enables researchers who want to understand and compare existing algorithms as it breaks boundaries imposed by empirical data.

Lastly, we conduct experiments to assess the operational validity of the model by deriving new insights. The findings from these experiments build a deeper understanding of the judgment and aggregation process. We list aggregation algorithms, both established (e.g., UWM, PWM) and relatively newly designed (e.g., CWM, CM), and identify their drivers for weighting and performance, such as diversity of expertise or availability of seed events. When comparing performances in a range of scenarios, strengths and weaknesses in special situations (e.g., structural breaks) become noticeable. The degree to which aggregation algorithms are influenced by structural breaks varies substantially. The more extreme aggregation algorithms weight crowd members, the higher the performance decrease in structural breaks. Additionally, the observed scenario implies greater damage for weighting-based algorithms in case of recent structural breaks in comparison to benefits in case of a very distant structural break. This analysis demonstrates that simulation of scenarios and algorithms can trigger unexpected findings (e.g., potential overfitting by the CWM) and suggest routes for improvements. We also conduct experiments to create suppositions on select WOC elements. For example, we demonstrate the CWM's difficulties in scenarios with similar experts. Under certain conditions, knowledgeable experts are excluded from the crowd, while most of the weight is assigned to an unknowledgeable expert. In addition, we elaborate on the concept of individual uncertainty and measure its impact on performance. Depending on expert characteristics, the optimal individual uncertainty differs. The less informed an expert is, the higher the ideal individual uncertainty.

We show that the choice of aggregation algorithm depends highly on the underlying scenario. Factors include expert characteristics (such as individual uncertainty), crowd characteristics (such as diversity of expertise), and

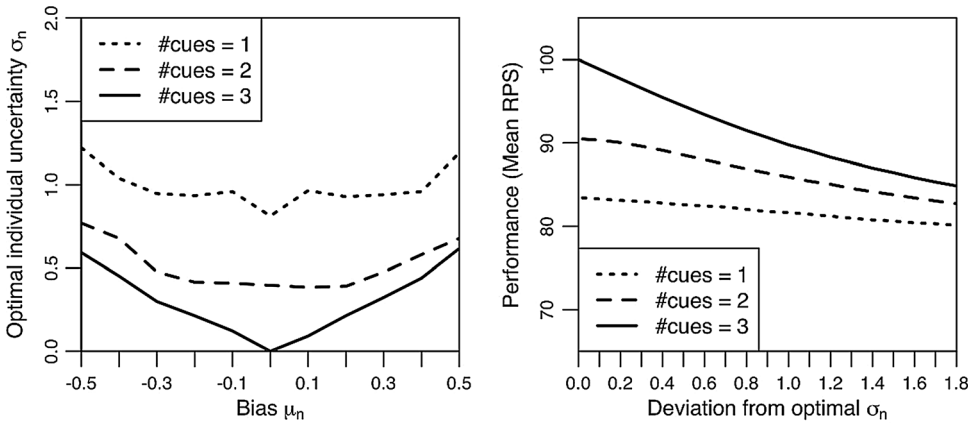

Fig. 7 Optimal individual uncertainty and its implications

event characteristics (such as availability of seed events or probability of structural breaks). These considerations, in combination with our simulation model, can help practitioners choose the right algorithm for a specific scenario. Furthermore, we have highlighted practical risks of weighting algorithms. While weighting can improve performance, events such as structural breaks may have radical consequences.

The results in this paper are beset by limitations. As with all simulations, our model is a less complex representation of the real world and therefore simplifies certain aspects of it. For experimentation, we chose normal distributions for individual uncertainty and cues. Furthermore, our model assumes all cues to be equally important and that an expert either has full or no access to a cue. Experts do not learn from their mistakes and do not switch the cues their judgment is based on. Both the probability distribution of the expert judgments and events are symmetrical normal distributions. The simulation of density judgments via multiple drawings from normal distributions implies high computational complexity, which also limits the intricacy of the model. In some experiments, we use extreme scenarios that are unlikely to occur in reality and might limit the explanatory power of the results. In summary, our simulation model is part of a distinct third way complementing theoretical analysis and empirical analysis. We believe that the complementary strength of these three approaches can jointly contribute to understanding WOC and aggregation algorithms.

Future work should, therefore, focus on comparing empirical and simulated data and strive towards further theorizing. Subsequently, researchers can use the simulation model for experimentation to broaden our knowledge base of WOC and aggregation algorithms. In addition, the simulation model can be enhanced and expanded to achieve a more sophisticated view of the real world. To enhance the model's practical applicability, it may be parameterized with common expert and crowd characteristics. This includes learning experts that adjust their behavior over time based on their historical performance, which could substantially change the performance of all aggregation algorithms.

Acknowledgment Open Access funding provided by Projekt DEAL. This work has in parts been funded by Deutsche Forschungsgemeinschaft (DFG).

Appendix

See Table 4.

Table 4 List of abbreviations (sorted by occurrence)

Abbreviation	Description
WOC	Wisdom of Crowds
RPS	Ranked Probability Score
BS	Brier Score
UWM	Unweighted Model
PWM	Performance Weighted Model
CWM	Contribution Weighted Model
CM	Contribution Model
BEM	Best Expert Model
REM	Random Expert Model

References

Ashton AH, Ashton RH (1985) Aggregating subjective forecasts: some empirical results. Manag Sci 31:1499–1508

Banks J, Carson II, Nelson BL, Nicol DM (2010) Discrete-event system simulation, 5th edn. Prentice Hall, Upper Saddle River

Bates JM, Granger CWJ (1969) The combination of forecasts. J Oper Res Soc 20:451–468

Beese J, Haki MK, Aier S, Winter R (2019) Simulation-based research in information systems. Bus Inf Syst Eng 61:503–521. https://doi.org/10.1007/s12599-018-0529-1

Bichler M, Hess T, Krishnan R, Loos P (2014) Emerging research areas in business and information systems engineering. Bus Inf Syst Eng 6:1–2. https://doi.org/10.1007/s12599-013-0309-x

Bickel JE (2007) Some comparisons among quadratic, spherical, and logarithmic scoring rules. Decis Anal 4:49–65

Brenner LA, Koehler DJ, Liberman V, Tversky A (1996) Overconfidence in probability and frequency judgement: a critical examination. Organ Behav Hum Decis Process 65:212–219

Brier GW (1950) Verification of forecasts expersses in terms of probaility. Mon Weather Rev 78:1–3. https://doi.org/10.1126/science.27.693.594

Bröcker J, Smith LA (2007) Scoring probabilistic forecasts: the importance of being proper. Weather Forecast 22:382–388. https://doi.org/10.1175/waf966.1

Broomell SB, Budescu DV (2009) Why are experts correlated? Decomposing correlations between judges. Psychometrika. https://doi.org/10.1007/s11336-009-9118-z

Budescu DV (2006) Confidence in aggregation of opinions from multiple sources. In: Fiedler K, Juslin P (eds) Information sampling and adaptive cognition. Cambridge University Press, Cambridge, pp 327–352

Budescu DV, Chen E (2015) Identifying expertise to extract the wisdom of crowds. Manag Sci 61(2):267–280

Carbone R, Armstrong JS (1982) Evaluation of extrapolative forecasting methods: results of a survey of academicians and practitioners. J Forecast 1:215–217. https://doi.org/10.1002/for.3980010207

Chen E, Budescu DV, Lakshmikanth SK et al (2016) Validating the contribution-weighted model: robustness and cost-benefit analyses. Decis Anal 13(2):128–152

Clemen RT (1989) Combining forecast: a review and annotated bibliography. Int J Forecast 5:559–583

Clemen RT, Winkler RL (1986) Combining Economic Forecasts. J Bus Econ Stat 4:39–46. https://doi.org/10.2307/1391385

Clemen RT, Winkler RL (1999) Combining probability distributiond from experts in risk analysis. Risk Anal 19:155–156

Colson AR, Cooke RM (2017) Cross validation for the classical model of structured expert judgment. Reliab Eng Syst Saf 163:109–120. https://doi.org/10.1016/j.ress.2017.02.003

Cooke RM, Goossens LLHJ (2008) TU Delft expert judgment data base. Reliab Eng Syst Saf 93:657–674. https://doi.org/10.1016/j.ress.2007.03.005

Dalrymple DJ (1975) Sales forecasting methods and accuracy. Bus Horiz 18:69–73

Dana J, Broomell SB, Budescu DV, Davis-Stober CP (2015) The composition of optimally wise crowds. Decis Anal 12:130–143. https://doi.org/10.1287/deca.2015.0315

Davis JP, Eusebgardt KM, Binghaman CB (2007) Developing theory through simulation methods. Acad Manag Rev 32:480–499. https://doi.org/10.5465/amr.2007.24351453

Davison RM, Martinsons MG (2016) Context is king! Considering particularism in research design and reporting. J Inf Technol 31:241–249. https://doi.org/10.1057/jit.2015.19

Davis-Stober CP, Budescu DV, Dana J, Broomell SB (2014) When is a crowd wise? Decision 1:79–101

de Condorcet N (1785) Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Reprint by Cambridge University Press, Cambridge

de Menezes LM, Bunn WD, Taylor JW (2000) Review of guidelines for the use of combined forecasts. Eur J Oper Res 120:190–204. https://doi.org/10.1016/s0377-2217(98)00380-4

Eggstaff JW, Mazzuchi TA, Sarkani S (2014) The effect of the number of seed variables on the performance of Cooke's classical model. Reliab Eng Syst Saf 121:72–82. https://doi.org/10.1016/j.ress.2013.07.015

Einhorn HJ, Hogarth RM, Klempner E (1977) Quality of group judgment. Psychol Bull 84:158–172

Epstein ES (1969) A scoring system for probability forecasts of ranked categories. J Appl Meteorol 8:985–987

European Central Bank (2017) ECB survey of professional forecasters. https://www.ecb.europa.eu/stats/ecb_surveys/survey_of_professional_forecasters/html/index.en.html

Fildes R, Hastings R (1994) The organization and improvement of market forecasting. J Oper Res Soc 1–16

Fischer GW (1981) When oracles fail—a comparison of four procedures for aggregating subjective probability forecasts. Organ Behav Hum Perform 110:96–110

Flandoli F, Giorgi E, Aspinall WP, Neri A (2011) Comparison of a new expert elicitation model with the classical model, equal weights and single experts, using a cross-validation technique. Reliab Eng Syst Saf 96:1292–1310. https://doi.org/10.1016/j.ress.2011.05.012

Galton F (1907) Vox populi—the wisdom of crowds. Nature 75:450–451. https://doi.org/10.1038/075450a0

Genest C, Zidek JV (1986) Combining probability distributions: a critique and an annotated bibliography. Stat Sci 1:147–148. https://doi.org/10.1214/ss/1177013831

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102:359–378. https://doi.org/10.1198/016214506000001437

Hammitt JK, Zhang Y (2013) Combining experts' judgments: comparison of algorithmic methods using synthetic data. Risk Anal 33:109–120

Harling J (1958) Simulation techniques in operations research—a review. Oper Res 6:307–319. https://doi.org/10.1126/science.183.4130.1141-a

Harrison JR, Lin Z, Carroll GR, Carley KM (2007) Simulation modeling in organizational and management research. Acad Manag Rev 32:1229–1245. https://doi.org/10.5465/amr.2007.26586485

Hastie R, Kameda T (2005) The robust beauty of majority rules in group decisions. Psychol Rev 112:494–508. https://doi.org/10.1037/0033-295x.112.2.494

Herzog SM, Hertwig R (2009) The wisdom of many in one mind: improving individual judgments with dialectical bootstrapping. Psychol Sci. https://doi.org/10.1111/j.1467-9280.2009.02271.x

Herzog SM, Hertwig R (2011) The wisdom of ignorant crowds: predicting sport outcomes by mere recognition. Judgm Decis Mak 6:58–72

Hogarth RM (1978) A note on aggregating opinions. Organ Behav Hum Perform 21:40–46

Hogarth M, Makridakis S (1981) Forecasting and planning: an evaluation. Manag Sci 27:115–138

Hora SC, Fransen BR, Hawkins N, Susel I (2013) Median aggregation of distribution functions. Decis Anal. https://doi.org/10.1287/deca.2013.0282

Hurley WJ, Lior DU (2002) Combining expert judgment: on the performance of trimmed mean vote aggregation procedures in the presence of strategic voting. Eur J Oper Res 140:142–147. https://doi.org/10.1016/s0377-2217(01)00226-0

Jouini MN, Clemen RT (1996) Copula models for aggregating expert opinions. Oper Res 44:444–457

Karelaia N, Hogarth RM (2008) Determinants of linear judgment: a meta-analysis of lens model studies. Psychol Bull 134:404–426. https://doi.org/10.1037/a0013550

Keuschnigg M, Ganser C (2017) Crowd wisdom relies on agents' ability in small groups with a voting aggregation rule. Manag Sci 63:mnsc.2015.2364. https://doi.org/10.1287/mnsc.2015.2364

Kittur A, Kraut RE (2008) Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In: Proceedings of the 2008 ACM conference on Computer Supported Cooperative Work, ACM, pp 37–46

Kleijnen JPC (1995) Verification and validation of simulation models. Eur J Oper Res 82:145–162. https://doi.org/10.1109/wsc.2000.899697

Larrick RP, Soll JB (2006) Intuitions about combining opinions: misappreciation of the averaging principle. Manag Sci 52:111–127. https://doi.org/10.1287/mnsc.1060.0518

Larrick RP, Mannes AE, Soll JB, Krueger JI (2011) The social psychology of the wisdom of crowds. In: Krueger JI (ed) Frontiers of social psychology. Social judgment and decision making. Psychology Press, Hove, pp 227–242

Law AM, Kelton DW (2007) Simulation modeling & analysis. McGraw Hill, Boston

Lawrence M, Goodwin P, O'Connor M, Oenkal D (2006) Judgmental forecasting: a review of progress over the last 25 years. Int J Forecast 22:493–518. https://doi.org/10.1016/j.ijforecast.2006.03.007

Lee MD, Zhang S, Shi J (2011) The wisdom of the crowd playing the price is right. Mem Cognit 39:914–923

Lee JS, Filatova T, Ligmann-Zielinska A, et al (2015) The complexities of agent-based modeling output analysis. JASSS. https://doi.org/10.18564/jasss.2897

Leimeister JM (2010) Collective intelligence. Bus Inf. Syst Eng 2:245–248

Lorscheid I, Heine BO, Meyer M (2012) Opening the "Black Box" of simulations: increased transparency and effective communication through the systematic design of experiments. Comput Math Organ Theory. https://doi.org/10.1007/s10588-011-9097-3

Lucas HC, Goh JM (2009) Disruptive technology: how Kodak missed the digital photography revolution. J Strateg Inf Syst 18:46–55. https://doi.org/10.1016/j.jsis.2009.01.002

Mahajan V, Wind Y (1988) New product forecasting models. Int J Forecast 4:341–358. https://doi.org/10.1016/0169-2070(88)90102-1

Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. J Pers Soc Psychol 107:276–299. https://doi.org/10.1037/a0036677

McKenzie CRM, Liersch MJ, Yaniv I (2008) Overconfidence in interval estimates: what does expertise buy you? Organ Behav Hum Decis Process 107:179–191. https://doi.org/10.1016/j.obhdp.2008.02.007

Morris PA (1986) Combining probability distributions: a critique and an annotated bibliography: comment. Stat Sci 1:141–144. https://doi.org/10.1108/eb038541

Murphy AH (1970) The ranked probability score and the probability score: a comparison. Mon Weather Rev 98:917–924

Oliphant TE (2007) Python for scientific computing. Comput Sci Eng 9:10–20. https://doi.org/10.1109/mcse.2007.58

Palley AB, Soll JB (2019) Extracting the wisdom of crowds when information is shared. Manag Sci 65(5):1949–2443. https://doi.org/10.1287/mnsc.2018.3047

Park S, Budescu DV (2015) Aggregating multiple probability intervals to improve calibration. Judgm Decis Mak 10(2):130–143

Petrovic D, Roy R, Petrovic R (1998) Modelling and simulation of a supply chain in an uncertain environment. Eur J Oper Res 109:299–309. https://doi.org/10.1016/s0377-2217(98)00058-7

Sanders NR (1997) The status of forecasting in manufacturing firms. Prod Invent Manag J 38:32–35

Sargent RG (1987) An overview of verification and validation of simulation models. In: Proc 19th Conf Winter Simulation Conference, pp 33–39. https://doi.org/10.1145/318371.318379

Sargent RG (2005) Verification and validation of simulation models. In: Proc 37th Winter Simulation Conference, pp 130–143. https://doi.org/10.1109/wsc.2000.899697

Sarker S (2016) Building on Davison and Martinsons' concerns: a call for balance between contextual specificity and generality in IS research. J Inf Technol 31:250–253. https://doi.org/10.1057/s41265-016-0003-9

Schurz G (2008) The meta-inductivist's winning strategy in the prediction game: a new approach to Hume's problem. Philos Sci 75:278–305. https://doi.org/10.1086/592550

Slack N, Chambers S, Johnston R (2007) Operations management, 5th edn. Pearson Education, Essex

Surowiecki J (2005) The wisdom of crowds. Anchor

Tay AS, Wallis KF (2000) Density forecasting: a survey. J Forecast 19:235–254

Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185(4157):1124–1131

Urban GL, Weinberg BD, Hauser JR (1996) Premarket forecasting of really-new products. J Market 60(1):47–60

Van Wesep ED (2016) The quality of expertise. Manag Sci 62:2937–2951. https://doi.org/10.2139/ssrn.2257995

Wagner C, Suh A (2014) The wisdom of crowds: impact of collective size and expertise transfer on collective performance. In: Proceedings Annual Hawaii International Conference on System Sciences, pp 594–603. https://doi.org/10.1109/hicss.2014.80

Wagner C, Vinaimont T (2010) Evaluating the wisdom of crowds. Proc Issues Inf Syst XI:724–732

White JW, Rassweiler A, Samhouri JF et al (2014) Ecologists should not use statistical significance tests to interpret simulation model results. Oikos. https://doi.org/10.1111/j.1600-0706.2013.01073.x

Winkler RL, Makridakis S (1983) The combination of forecasts. J Royal Stat Soc 146:150–157

Winter R (2009) What in fact is fundamental research in business and information systems engineering? Bus Inf Syst Eng 1:192–199. https://doi.org/10.1007/s12599-008-0024-1

Woolley AW, Chabris CF, Pentland A et al (2010) Evidence for a collective intelligence factor in the performance of human groups. Sci 330(6004):686–688

Yates JF, McDaniel LS, Brown ES (1991) Probabilistic forecasts of stock prices and earnings: the hazards of nascent expertise. Organ Behav Hum Decis Process 49:60–79. https://doi.org/10.1016/0749-5978(91)90042-r