# 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

*Research Center Finance & Information Management, University of Augsburg, 86159 Augsburg, Germany*

*E-Mail firstname.lastname@fim-rc.de; corresponding author Henner Gimpel, henner.gimpel@fim-rc.de*

## Abstract

Analytical and empirical research has shown that the aggregation of multiple independent expert judgments significantly improves the quality of forecasts as compared to individual expert forecasts. This "wisdom of crowds" (WOC) has recently sparked substantial research interest. However, previous studies on the strengths and weaknesses of aggregation models have been limited by restricted empirical data and analytical complexity. Based on a comprehensive analysis of the existing knowledge of WOC and different aggregation models, we designed and implemented a simulation model to emulate WOC scenarios with a wide range of parameters. The model has been thoroughly evaluated: We validated its assumptions against propositions derived from relevant literature, built a computational representation, and showed its applicability in a variety of experimental scenarios to validate its behavior. Experimental scenarios include the investigation of aggregation model behavior on a detailed level, the assessment of aggregation model performance, and the exploration of previously undiscovered hypotheses on WOC. The simulation model helps expand the understanding of WOC in fields where previous research was restricted. Additionally, it gives directions for further developing aggregation models, and contributes to a general understanding of the WOC phenomenon.

*Keywords:* simulation, forecasting, expert judgment, expert aggregation, wisdom of crowds

*Highlights:*

- Provides an overview of the literature on the wisdom of crowds and derive propositions.
- Presentation of the first general simulation model for aggregation of expert density judgments.
- Threefold evaluation: investigation of aggregation model, assessment of aggregation model performance, exploration of previously undiscovered hypotheses.
- Highlights limitations of commonly used aggregation models.

*Funding:* This work has in parts been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project 343128888

## 1. Introduction

High quality forecasts are essential for informed decision-making (Sanders, 1997). As such, they play an important role in all kinds of areas such as sales, product and service development, finance, and operations management (Brailsford & Faff, 1996; Dalrymple, 1975; Fildes & Hastings, 1994; Mahajan & Wind, 1988; Scott, 1981; Slack, Chambers, & Johnston, 2007; Urban, Weinberg, & Hauser, 1996). In contrast to the traditional approach of relying on a single forecast, a comprehensive amount of research suggests combining multiple forecasts in order to improve forecast accuracy (Clemen, 1989). Next to the combination of forecasts based on statistical models (Bates & Granger, 1969; Winkler & Makridakis, 1983), this also applies to the

combination of forecasts drawn from human judgment (Ashton & Ashton, 1985; Lawrence, Goodwin, O'Connor, & Oenkal, 2006). Consequently, the aggregation of multiple judgments is an important area in decision analysis research (Hurley & Lior, 2002).

Back in 1907, Galton had already identified the value of aggregating multiple judgments to exploit the individual efforts of a crowd of people (Galton, 1907; Surowiecki, 2005). While individual judgments might be biased (M. Hogarth & Makridakis, 1981; Kahneman & Tversky, 1974) and individuals typically lack the required level of expertise (Van Wesep, 2016), the aggregation of multiple individual judgments can solve those issues. The positive effects of judgment aggregation, with the goal of outperforming an individual's judgment, are commonly referred to as the wisdom of crowds (WOC) phenomenon (Budescu & Chen, 2015; Larrick, Mannes, & Soll, 2011).

The aggregated judgment from a crowd of people can be derived via (consensus-based) group decision processes (e.g. Kittur & Kraut, 2008; Leimeister, 2010; Woolley, Chabris, Pentland, Hashmi, & Malone, 2010) or by aggregating individual judgments (e.g. Ashton & Ashton, 1985; Bates & Granger, 1969; Budescu & Chen, 2015; Clemen & Winkler, 1999; Einhorn, Hogarth, & Klempner, 1977). Looking at the latter, the mathematical way of aggregating multiple judgments (aggregation models) becomes an important driver of the WOC effect. When examining different aggregation models in the context of WOC, data availability plays an important role. Performance-based models (e.g. history-based models, like suggested by Budescu & Chen, 2015) require information (such as previous predictions) to calculate performance measures for the involved experts. Those information sources are so called seed variables (Cooke & Goossens, 2008). Thus, we look at a group of people who individually provide judgments over multiple periods. The individual judgments of the target period are then aggregated into one combined judgment. The data issue becomes particularly problematic when rare circumstances or systematically deviating characteristics lie at the heart of the research question. Consequently, the evaluation of aggregation models places high demands on the corresponding data and only few researchers utilize empirical data (e.g. Budescu & Chen, 2015; Galton, 1907; Herzog & Hertwig, 2011; Wagner & Suh, 2014). To overcome the shortness of adequate empirical data, other researchers use simulation (e.g. Hammitt & Zhang, 2013; Hastie & Kameda, 2005; Keuschnigg & Ganser, 2016; Mannes, Soll, & Larrick, 2014). With the help of simulation, alternating characterizations of the crowd and the relevant environment (referred to as scenarios) can be recreated and the performance of aggregation models can be studied.

The existing literature that applies simulation models in the context of WOC primarily focuses on point estimates (Hastie & Kameda, 2005; Keuschnigg & Ganser, 2017; Mannes et al., 2014; Wagner & Vinaimont, 2010). Further approaches include the simulation of experts who provide rankings of alternatives (Hurley & Lior, 2002) or probability judgments on binary events (Budescu & Chen, 2015). To our knowledge, only Hammitt and Zhang (2013) have deliberately addressed the simulation of experts providing density judgments. Amongst probability judgments, density judgments offer the most information about the underlying uncertainty. This type of judgment includes probabilities for each potential future value of the variable in question (Tay & Wallis, 2000). Within their simulation model, Hammitt and Zhang (2013) assumed experts to be perfectly calibrated, meaning that their individual error terms are unbiased. This assumption is contrary to established theory, stating that even experts rely on heuristics to provide judgments under uncertainty and thereby are also subject to systematic biases (Kahneman & Tversky, 1974). For example, there is strong evidence for overconfidence in probability judgments, which further interferes with the assumption of perfect calibration (e.g. Brenner, Koehler, Liberman, & Tversky, 1996; McKenzie,

Liersch, & Yaniv, 2008; Teigen & Jørgensen, 2005). In addition, Hammitt and Zhang (2013) only examine aggregation models using two experts, which is very restrictive, as aggregation becomes especially interesting with bigger crowds. Increasing the size of the crowd would imply a rapidly growing complexity. Density judgments are utilized by major institutions for major use cases, for example, by the European Central Bank (ECB), the Bank of England, and the Federal Reserve Bank of Philadelphia (Tay & Wallis, 2000). An example is the European Central Bank's Survey of Professional Forecasters (ECB SPF), regularly querying multiple forecasters to submit density judgments on matters like future inflation, growth of the gross domestic product, and unemployment rates (Bowles et al., 2007). Nevertheless, the simulation of experts providing density judgments to examine WOC is lagging in terms of its potential. Guided by theory on simulation model development and evaluation, we fill this research gap by providing a simulation model to further enhance research on the WOC phenomenon in general, with a focus on judgment aggregation models. As a result, we propose a novel model to simulate expert density judgments, which can cope with large crowds and provides the flexibility to design versatile and detailed, as well as very simplified scenarios of experts, events, and characterizing circumstances. Additionally, we derive new insights into WOC in the process of thoroughly evaluating the model.

Sect. 2 outlines the research method. Sect. 3 introduces the judgment setting at hand, elaborates on performance measures for the evaluation of judgments, describes a broad set of aggregation models, and closes the theoretical background by deriving propositions from existing WOC literature. Sect. 0 describes the conceptual simulation model. Sect. 5, 6 and 7 follow the evaluation process by Sargent (1987, 2005). First, we compare our conceptual model to the propositions derived from WOC 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. 8 outlines the major theoretical and managerial implications as well as the limitations.

## 2. Research Method

Simulation involves creating a stochastic model that corresponds to a real situation (system) and subsequent experimentation with the model (Harling, 1958). The method has gained substantial support as a means of generating new theories (Davis, Eusebgardt, & Binghaman, 2007) and is widely used in operations research (OR; Law & Kelton, 1991). It can be applied in both specific WOC research and traditional OR studies (Harling, 1958; Petrovic, Roy, & Petrovic, 1998).

According to Sargent (1987, 2005), the simulation modelling process involves three major components (Figure 1): the problem entity, the conceptual model, and the computerized model. The problem entity is the system to be modeled (e.g., based on real world observations). The conceptual model is the representation (e.g., mathematical) of the problem entity and is derived in an analysis and modeling phase. The computerized model is the technical implementation of the conceptual model, created in a computer programming and implementation stage. Finally, the computerized model is used to conduct experiments and gain new insights into the problem entity.

As the correctness of the simulation model and its results are of major concern when applying simulation, verification and validation play an important role. Sargent (1987, 2005) proposes three steps: conceptual model validation, computerized model verification, and operational validation. The conceptual model validation consists of two tasks. First, verifying the correctness of the underlying theory and assumptions, and second, ensuring that the model, with all its components, is reasonable for the purpose of the simulation

model (Sargent, 1987, 2005). We conduct this validation by deriving propositions about the behavior of experts and aggregation models, as well as about the WOC phenomenon from the existing literature (Sect. 3.4); then, we evaluate whether our simulation model behaves accordingly (Sect. 5). We do this preferably based on analytical and logical reasoning and only use simulation when necessary.

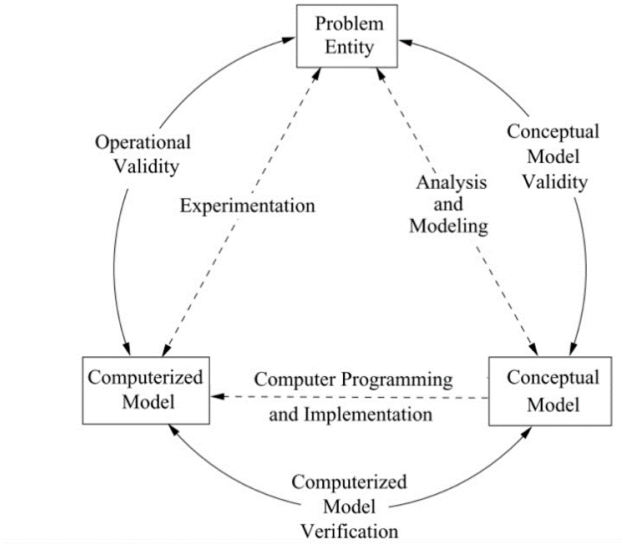


*Figure 1 – Model validation and verification based on Sargent (1987, 2005)*

Consequently, we simultaneously conduct the computerized model verification, which provides strong evidence that the technical implementation adequately represents the conceptual model (Sect. 6). Thus, simulation can also 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 expected results from the conceptual model (Kleijnen, 1995).

Finally, the operational validation aims to determine whether the model's output behavior has the accuracy required for the model's intended purpose (Sargent, 1987, 2005). Most of the elements in the problem entity are practically non-observable (i.e., a collection of empirical data on elements like expert characteristics or very 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 was previously unavailable. We therefore assess operational validity by exploring the model behavior in-depth and showing that it creates results that provide new insights into aggregation models and the WOC phenomenon (Sect. 7). Through extensive evidence for applicability and usefulness, operational validity is accepted. We do this by shedding light on three sets of experimental conditions, namely how aggregation models weight experts by exploring the performance of aggregation models under systematically changing conditions, and identifying new hypotheses by conducting several experiments.

## 3. Theoretical Background on the Aggregation of Expert Judgments

There are multiple terms for expressions about an entity that is at least partially unknown such as judgments, forecasts, opinions, and predictions. Within our paper, we use judgments as the general term comprising those different notations. Expert judgments and their aggregation can be carried out under very different circumstances, and the dimensions in which judgment methods can be evaluated are versatile (Carbone & Armstrong, 1982). Therefore, a well-defined judgment setting, an adequate performance measure, and corresponding judgment aggregation models must be described. Finally, the behavior of the underlying simulation model must be in line with existing results on WOC and aggregation models to examine conceptual model validity. Thus, we derive propositions on WOC and aggregation models from existing literature.

### *3.1. Judgment and Aggregation Setting*

The expert judgment task at hand, as briefly described in the introduction, involves a crowd of multiple experts who individually provide their judgment on a particular event. Following the origin of the WOC phenomenon (Galton, 1907), we consider individual judgments of experts (such as in Davis-Stober, Broomell, Budescu, & Dana, 2014; Lee, Zhang, & Shi, 2011; Mannes et al., 2014) and thus do not account for group dynamics. Each expert forms his judgment about the future state of the event based on his observation of relevant cues. Experts have access to different cues of potentially different quality (based on Brunswik´s lens model as in Karelaia & Hogarth, 2008). Depending on the clarity of the observed cues and their individual uncertainty, experts may choose to provide information on the certainty of their judgment. Generally, the incorporation of the probability component is favorable in uncertain environments (Fischer, 1981) and has also gained popularity (Bröcker & Smith, 2007).

Looking at probability judgments, density judgments bear the most information as they include probabilities for each potential future value of the variable in question (Tay & Wallis, 2000). Consequently, the expert assigns probabilities to all predefined possible future states of the event. In practice, similar procedures are carried out by the ECB (European Central Bank, 2017), the Bank of England, and the Federal Reserve Bank of Philadelphia (Tay & Wallis, 2000). At the point of judgment, the future state of the event cannot be witnessed. After some time, the realization of the event becomes observable and the expert judgments can then be compared to the realized state of the event for ex-post performance measurement (Hammitt & 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 judgments (e.g., as in Budescu & Chen, 2015). Thus, a decision maker who is provided with judgments from different experts can leverage their previous performance to aggregate the judgments. The aggregation of judgments is conducted via aggregation models that either rely on past performance to create some sort of expert weighting (history-based aggregation models) or that do not consider past performance and can thus be used ad-hoc.

A large-scale application of a similar procedure can be observed in the European Central Bank Survey of Professional Forecasters (ECB SPF). Since 1999, the ECB has conducted a quarterly survey on future macroeconomic development within the euro area. They ask a panel of roughly 75 experts from academia, major banks, and other institutions about their judgment on variables such as inflation, gross domestic product growth, and unemployment. Parts of those judgments are provided as density judgments (Bowles et al., 2007). In the first quarter of 2017, for example, 57 forecasters responded. Looking at the question

concerning the inflation rate for the year 2018, the future states of the inflation ranged from below -0.5 percent to above 3.5 percent. The overall range was cut in eight equally sized buckets of 0.4 percent each, and two buckets representing values smaller than -0.5 percent and higher than 3.5 percent. For every bucket, the forecasters assigned probabilities, adding up to 100 percent over the 10 buckets. Those probabilities were then averaged for each bucket to arrive at an aggregated probability distribution (European Central Bank, 2017). This example shows the application of the described expert aggregation setting in practice. The question at hand is whether a simple average is best or whether other aggregation models might be better, depending on the exact scenario, event, experts, etc.

### *3.2. Performance Measurement*

Judgments can be evaluated via a versatile range of criteria such as accuracy, ease of interpretation, cost, time, and robustness. As accuracy is the most important (Carbone & Armstrong, 1982), we take a closer 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 some situations, decision makers may not only be interested in the expected mean accuracy, but also in the corresponding variance of the accuracy. Thus, besides the mean accuracy, the variance of the accuracy is a secondary criterion characterizing judgment performance. When looking at density judgments, an accurate judgment assigns a bulk of the probability to values close to the realization, and ideally centers on the realization with a low dispersion. In decision theory, a scoring rule measures the accuracy of such probabilistic judgments. Scoring rules apply to tasks in which individuals assign probabilities to a set of mutually exclusive discrete outcomes. A score can be thought of as a measure that calibrates a probabilistic judgment (Gneiting & 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 is a function that assigns the highest score to the true probability distribution. The use of a proper scoring rule encourages the expert to maximize the expected reward by accurately providing their best judgment. Also, a scoring rule is strictly proper if it is uniquely optimized by the true probabilities (Murphy, 1970).

The Ranked Probability Score (RPS; Epstein, 1969) is a commonly used scoring rule for ordinal and cardinal data. The RPS measures the mean squared difference between the cumulative distribution functions of the judgment and the actual outcome. Therefore, the lower the RPS is for a prediction, the better the prediction is calibrated. 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 underlying event, $F_i$ represents the value of the cumulative distribution function of the prediction for outcome $i$, and $O_i$ indicates the corresponding distribution function of the true observation. Without transformation, the RPS assigns a score of zero to the best prediction, and a score of $|I| - 1$ to the worst one. Via the use of $a$ and $b$, a linear transformation can also be used to normalize the score.

Another widely used scoring rule is a specialization of the RPS called the Brier Score (Brier, 1950). The Brier Score (BS) measures the mean squared difference between the predicted probability assigned to the possible outcomes and the actual result. The BS is only appropriate for binary or nominal outcomes that cannot be put in a specific order, but is inappropriate for ordinal or cardinal variables, which take on three or

more values. This is because the BS assumes that all possible outcomes are equivalently distant from one another. Thus, we use the RPS.

### *3.3. Aggregation Models*

In what is believed to be the first publication on the WOC phenomenon, Galton (1907) used the median judgment to aggregate the opinions of the crowd to one judgment. This approach can be seen as the origin of aggregation models (also known as aggregation algorithms or aggregation rules). Over the years and with rising interest in the WOC phenomenon, a variety of aggregation models were developed. We differentiate those approaches by three basic characteristics. First, does the model rely on past predictions from each expert (history-based) or can it be used ad-hoc? Second, does the model include all members of the crowd in the aggregation, or does it select a sub-set of experts from the crowd? Third, does the weighting of the model for the selected crowd deviate from an equal weighting? Based on those three characteristics, we identify six aggregation models from literature to focus on particularly relevant models and cover a broad range of possible characteristics (Table 1). Additional models that could be included involve Copula models (Jouini & Clemen, 1996) and Cooke's model (Cooke & Goossens, 2008).

| **Model name** | **Ad-Hoc** | **Selec-tion** | **Weight-ing** | **Source** |
|---|---|---|---|---|
| Unweighted Model (UWM) | Yes | No | No | Clemen & Winkler (1986), Budescu & Chen (2015) |
| Random Expert Model (REM) | Yes | Yes | No | Davis-Stober, Broomell, Budescu, & Dana (2014) |
| Performance Weighted Model (PWM) | No | No | Yes | Budescu, Chen, Lakshmikanth, Mellers, & Tetlock (2016) |
| Best Expert Model (BEM) | No | Yes | No | Hammitt & Zhang (2013) |
| Contribution Model (CM) | No | Yes | No | Budescu & Chen (2015) |
| Contribution Weighted Model (CWM) | No | Yes | Yes | Budescu & Chen (2015) |

*Table 1 - Overview of aggregation models*

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

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

where $p_{i,n}$ is the probability assigned by expert $n$ to the event outcome belonging to interval $i$, and $N$ is the set of all experts. Based on the UWM, several other models have been developed that weight experts by including a measure of the experts' past performance. The past performance can, for example, be measured with scoring rules. The Brier Weighted Model (Budescu & Chen, 2015; Budescu et al., 2016) is an example

of such an aggregation model; it computes the Brier Score for every expert, averaging it over all historical events. Based on this score, weights ($w_n$) are allocated to the experts. The sum of all weights is always equal to 1. The better the average BS of an expert, the higher his proportionate weight. We consider the Performance Weighted Model (PWM) as a generalization of the Brier Weighted Model. To use it with ordinal data, our model is based on the RPS as a scoring rule for past performance.

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

The PWM only considers the average absolute historical performance of an expert $n$, described by $RPS_n$, for all seed events for that expert $n$ submitted a judgment. As an enhancement to this model, Budescu and Chen (2015) developed the Contribution Weighted Model (CWM). Here, the term contribution is defined as the difference in aggregated performance with and without the target expert. This captures the effect of inclusion or exclusion of each person in the crowd. In the CWM definition by Budescu and Chen (2015), 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 $\mathbb{1}_{[\,]}$ which is set to 1 if the condition in the subscript brackets 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. While we do not use it here, it is possible to set a threshold higher than zero to eliminate experts with a low positive contribution (Chen, Budescu, Lakshmikanth, Mellers, & Tetlock, 2016). The CWM is described as:

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

The CWM will ensure that an expert who judged well on past events – while the rest of the crowd judged poorly – will receive a high weight. 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 equally weights all experts with a positive contribution score (Budescu & Chen, 2015). Consequently, it produces less extreme weights compared to the CWM and does not depend as strongly on individual experts.

$$CM\colon p_i = \sum_{n \in N} \mathrm{w_n} \cdot p_{i,n},\ \forall \mathrm{i} \in \mathrm{I},$$

$$\text{with } \mathrm{w_n} = \begin{cases} \frac{\mathbb{1}_{[CON_n > 0]}}{\sum_{m \in N} \mathbb{1}_{[CON_m > 0]}} \ if\ \exists \mathrm{m} \in \mathrm{N}\colon \mathrm{CON_m} > 0 \\ \frac{1}{|N|}\ otherwise \end{cases} \tag{5}$$

A model using more extreme weights is the Best Expert Model (BEM). It employs the RPS as a performance measure and then only selects the expert(s) with the highest performance, oftentimes this will be a single best expert obtaining weight 1 (Hammitt & Zhang, 2013).

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

As a result, the BEM solely relies on the judgment of the identified "best expert" or multiple jointly best experts and thereby only takes advantage of the crowd by pinpointing the supposedly best performing expert(s) from the crowd. For evaluation purposes, we also include the Random Expert Model (REM; Davis-Stober et al., 2014) as a benchmark. The model randomly selects one expert $n$ from the crowd via a uniform distribution and weights them with 1.

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

Looking at those and other aggregation models, thus far, the literature has not settled on one superior aggregation model and promotes the application of multiple models (Hammitt & Zhang, 2013). Opinions about the degree to which a specific weighting induced by an aggregation model 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 & Winkler, 1986; Einhorn et al., 1977; Flandoli, Giorgi, Aspinall, & Neri, 2011). On the other hand, studies also support the superior performance of weighting-based models (Budescu & Chen, 2015; Budescu et al., 2016; Cooke & Goossens, 2008; Hammitt & Zhang, 2013). Consequently, the evaluation of aggregation models leaves room for further exploration.

### *3.4. WOC in Expert Aggregation*

To build and evaluate the simulation model, a conceptual model must be designed based on the problem entity. For that purpose, we derive propositions from existing literature on the general behavior of WOC and the corresponding aggregation models. This allows us to assess whether the conceptual representation of the problem entity is reasonable for the intended purpose and thus validate the conceptual model (Sargent, 2005). Proposition 1 defines the characteristics and quality of a single expert as the basic element of a WOC application. Proposition 2 postulates the general existence of the WOC effect, while Propositions 3 to 5 examine the WOC effect in more detail. Finally, Propositions 6 and 7 focus on aggregation models as another central element of crowd wisdom.

**Proposition 1:** The optimal expert possesses all the 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 this information. Hammitt and Zhang (2013) define this quality measure as a mix of two key figures: informativeness and calibration. Experts with high informativeness form judgments with a comparatively low variance around a mean value. Calibration defines an expert's biasedness, where bias is a systematic displacement of the mean value and can, for example, express extreme optimism or pessimism. Experts with a high bias are poorly calibrated. Thus, an expert's individual performance depends on the amount and quality of observed information, as well as a potential bias which influences variance.

**Proposition 2:** The wisdom of crowds exists and is robust to the application in different scenarios (e.g., a unidirectionally-biased crowd) and aggregation models.

Abstracting from single experts, the essential characteristic of WOC lies in its power to improve overall judgment performance by aggregating multiple, individual judgments. The power of aggregation to boost judgment performance originates in reducing the influence of incomplete information and biases, which impair individual judgments (Surowiecki, 2005). Thus, even when individual members of a crowd are riddled with biases, the aggregation of multiple individual judgments can make the crowd wise. Based on Davis-Stober et al. (2014) we define a crowd as wise if a linear combination of the individual judgments is on average more accurate than a randomly selected individual. This holds true even for the simplest model, the UWM, and under unfavorable conditions, such as correlated judgments or highly and unidirectionally-biased crowds. Apart from its robustness to characteristics of possible scenarios (e.g., a highly biased crowd), the existence of the WOC phenomenon 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 sub-crowd 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 only one expert, even if this is the best expert. One explanation can be found in the bias/variance trade-off. By utilizing the average of 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 expert opinions.

**Proposition 4:** The performance of aggregation models increases with crowd size.

Another factor that influences the performance of aggregation models is crowd size. Thinking of an unbiased expert judgment as the true value plus a random error (e.g., 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 the underlying variance (Einhorn et al., 1977). The basic effect that the performance of aggregation models increases with the size of the crowd has been shown analytically (Hogarth, 1978), empirically (Budescu et al., 2016), and via simulation (Wagner & Vinaimont, 2010).

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

Apart from the size of the crowd, its characteristics also play a substantial role. The best performance of WOC can be achieved when experts' judgments systematically differ as much as possible (Davis-Stober et al., 2014), because this maximizes the available information (Budescu, 2006). Even if each expert alone holds a small amount of the information, the total crowd might have access to all sources (Herzog & Hertwig, 2011). This crowd characteristic is called information diversity and partly explains the success of WOC applications. As a result, when adding a new expert to a crowd, it is best to choose the maximally different one from the existing crowd. This implies that experts with different backgrounds should be selected, and that their judgments should be collected independently (i.e., without communication between the 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:** The advantage of weighting models can be largely attributed to the selection of experts and only subordinately to subsequent weighting.

Many aggregation models (e.g., PWM, CM, and CWM) are based on the idea of utilizing external information to impose a selection or weighting of experts. Their advantage lies in their ability to identify experts with a high amount of knowledge within a crowd (Budescu & Chen, 2015). As such, their performance is not only affected by the crowd's size, but also by its characteristics, which enable the selection and weighting of experts with expertise. A larger unweighted crowd, including good and bad experts, might be outperformed by the selection and weighting of only good ones (Dana, Broomell, Budescu, & Davis-Stober, 2015; Einhorn et al., 1977). That being said, Budescu and Chen (2015) remark that the quality of the CWM's performance is mostly predicated on its efficiency in selecting the important experts in a crowd. The subsequent weighting procedure only accounts for about 40% of the advantage over other models.

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

Budescu and Chen (2015) state that the CWM performs better, the more historic events are available to evaluate the experts' historic performance. This assumption also applies to other history-based models (e.g., PWM, BEM). Adding past events reduces the volatility of an expert's average historic performance and thereby decreases the error rate of models trying to identify his true performance. One can detect that the performance of history-based models will not increase significantly when provided with more than a certain number of historic events (about 25; Budescu, Chen, Lakshmikanth, Mellers, & Tetlock, 2016). This is because the model's ability to identify expertise is already quite reliable for only those 25 events.

## 4. Simulation Model

Data availability plays an important role in WOC research and even more so if one wishes to examine the impact of rarely occurring circumstances, or measure the sensitivity of aggregation model performance in relation to specific characteristics. This makes applying simulation models particularly interesting for WOC research. When examining WOC phenomena, data on the judgment of experts on a certain event and the corresponding realization of the event are required. Consequently, our simulation model contains two key elements: stochastic events (which are to be judged) and experts (who provide those judgments). To simulate situations under different circumstances (called scenarios), the stochastic events as well as the experts are characterized by several adjustable parameters, which influence the quality of the judgment and the volatility of the events. As the history-based models require data on past realizations, the simulation of events and expert judgments is conducted for multiple points in time to acquire the necessary history of predictions and realizations.

An event is described as a probabilistic incident or state in the future. Examples for such events can be the future price of a stock, the future sales of a new product, or next year's inflation rate. All these examples 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 value of the event cues. Examples of cues impacting the above events could be a firm's historic profits and forward-looking business plan, results from a market survey, or a recent decision in monetary policy. While we theoretically allow all cues to be observable, non-observable cues can be modeled via experts' access to those cues, as described later. 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 $v_{t,j} > 0$:

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

Modeling the relation between the cues and the event $X$ as a weighted average, we follow Hastie and Kameda (2005) and Keuschnigg and Ganser (2016). Cues are random variables with a certain probability distribution. In our instance of the model, normal distributions are used exemplarily. When cues differ in their $\mu_{t,j}$, they can be more or less representative of the underlying event and hence differ in quality. Without loss of generality, we assume that there is one event per time step. The set of all events is thus defined as $X = \{X_{-T}, X_{-T+1}, \dots, X_0\}$. 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. They can, for example, be used to evaluate past expert performance. $X_0$ is to be estimated as a target event. Time indices $t, t+1$ etc. are ordered but not necessarily equally spaced. For instance, for annual events, the target period 0 might be next year, but could also be in two or three years from now.

In general, there are three possible ways of how experts make judgments. Experts can provide a point estimate (de Menezes, W. Bunn, & Taylor, 2000), an interval estimate (e.g. confidence intervals; McKenzie et al., 2008), or assign probabilities to a range of intervals, thus creating a discrete probability distribution (Yates, McDaniel, & Brown, 1991). As for our modeling, we apply the latter, which is extensively addressed in research (e.g. Clemen & Winkler, 1999; Genest & Zidek, 1986; Hammitt & Zhang, 2013), or practical forecasting applications (e.g., European Central Bank, 2017). Consequently, we select the RPS as an adequate scoring rule for measuring the performance of judgments; unlike other scoring rules, it is designed to work

with cardinal data and therefore applies in situations where the calibration of discrete probability distributions is being scored.

The second key component are the experts. The set of experts is denoted by $N$. Our simulated experts can differ in three key 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 an individual bias, which affects the mean of the distribution. Access to cues means that an expert knows about the realized value of a cue $C_{t,j}$. Hence, experts form judgments by calculating the weighted average of all realized cues known to them, while ignoring any cues they do not know about. Experts might have unjustified preconceptions or not accurately perceive and process the informational cues, thereby adding a random error parameterized by bias (mean ≠ 0) and uncertainty (variance > 0). Mathematically, the access to a cue is described by the variable $\alpha_{n,t,j}$. If expert $n$ observes the cue $C_{t,j}$, $\alpha_{n,t,j}$ defines the weight the expert allocates to the cue. Otherwise, it is 0. The random error term can be modeled with a probability distribution. Following Hammitt and Zhang (2013), we use a normal distribution as example: The expert-specific uncertainty is described by the variance $\sigma_n^2$ of the distribution, while the bias is the offset $\mu_n$. Adding up these requirements to a stochastic formula, the judgment $E_{n,t}$ of an expert $n$ for the event at time $t$ is modeled as follows:

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

The set $I$ of intervals can be defined freely within the range of possible outcomes. To derive a discrete probability distribution (probabilities for a 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. The procedure of deriving judgments must be repeated multiple times per scenario in order to create a large enough sample to statistically analyze the results. We therefore set the number of simulation runs per scenario individually.

To limit the complexity, we do not change the experts' access to information over time, meaning that the $\alpha_{n,t,j}$ stay constant with changing $t$, although in principle this could be relaxed. A real-world interpretation of this can be a series of similar events that must be estimated every year (e.g., inflation or GDP growth) and experts who have access to certain types of information that remains available to them over time. For evaluation, we focus on three basic types of scenarios that represent different stylized configurations of crowds. For illustration, we use the notation of a $|N| \times |J|$ matrix $\mathrm{A}_t$ containing the $\alpha_{n,t,j}$:

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

The first symbolic scenario contains several experts that all have access to different cues. They do not share access to any of the cues. Instead, each expert owns a different piece of information. In matrix notation, this generates a diagonal matrix. Our second scenario represents a hierarchical case containing experts with varying levels of expertise. The best-informed expert has access to all cues, while the worst-informed expert has no cues available. This manifests in a triangular matrix with an additional row of zeros for the fully uninformed expert. 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 a cluster matrix.

## 5. Conceptual Model Validation

Validating the conceptual model involves comparing the conceptual model to the corresponding problem entity, and addresses the question of whether the model adequately represents commonly accepted characteristics. To answer this question, 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 models and the WOC phenomenon in general. In the following sections, we assume for ease and brevity of the discussion 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 possible scenarios and hence limits computational complexity, while still allowing us to vary the expert's level of information via their access to the cues.

**Proposition 1** states that the optimal expert possesses all the information, no bias, and no individual uncertainty. An expert is considered optimal if he always allocates a probability of 100% to the interval, including the future realization of the event. Consider two experts: A and B, who have complete information (all $\alpha_{n,t,j} = 1$) 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 the lower uncertainty, it follows that on average, A's allocated probabilities will scatter less, and A will assign more probability to intervals close to his mean (i.e., the realization of the event). Consequently, A is a better expert 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 its cues, access to more cues increases the probability of being close to the realization. Therefore, expert A is better than B. Finally, consider A and B as experts with all information and no uncertainty. When expert A is less biased than expert B, he will be closer to the realization and will allocate never less and sometimes more probability to the interval that contains it. Again, A is better than B. We can conclude that an expert with less uncertainty, a lower bias and access to more information is generally better. Therefore, Proposition 1 holds for our model since the optimal expert must possess all available cues (all $\alpha_{n,t,j} = 1$), no bias ($\mu = 0$), and no individual uncertainty ($\sigma = 0$).

**Proposition 2** does not focus on the individual expertise of crowd members, but rather on the existence of crowd wisdom. Based on Davis-Stober, Broomell, Budescu, and Dana (2014) we define a crowd as wise if a linear combination of the individual judgments is on average more accurate than a randomly selected individual. Therefore, we want to show that aggregation models (like UWM, PWM, CWM and CM) are on average more accurate than a randomly drawn expert 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 the judgments of multiple experts, more cues are considered than for a single randomly selected expert, and the overall judgment becomes more informed. Since the RPS scoring rule is distance-sensitive and punishes positive as well as negative deviations, the stabilization of the judgment improves the judgment accuracy due to the integration of a broader range of cues. Similar effects are caused by the expert-specific error term. By aggregating the judgments of multiple experts and thus also aggregating their specific error term, the overall deviation from the value implied by the cues is reduced. This happens because the standard deviation of an average of independently distributed random variables is smaller than the standard deviation of a single error term. As a result, in our model, the aggregation of multiple experts improves judgment accuracy and leads to a wise crowd-based judgment. We can further demonstrate the validity and

robustness of this property by simulating several different scenarios, measuring the average RPS performance of the aggregation models. We use scenarios where we vary expertise, uncertainty or bias. Performance of the aggregation models in some exemplary scenarios are depicted in Figure 2. The RPS scores for all models that truly aggregate multiple experts (UWM, PWM, CWM, and CM) are on average higher than the individual RPS of a random expert (REM). Thus, we can show that the wisdom of crowds exists in our conceptual model and is robust to the application in a wide range of scenarios and aggregation models. However, extreme scenarios do exist where the REM outperforms other aggregation models.

A somewhat stronger assumption is formulated in **Proposition 3**, which suggests that for every scenario of expert crowds and events, 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 $w = (w_1, w_2, \dots, w_n)$ as the vector of weights assigned to the set $N$ of experts 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 weights $w_{BEM} = (1, 0, \dots, 0)$. Consider an extreme scenario where one expert holds all the information while all other experts are badly calibrated and uninformed. In this scenario, using $w_{BEM}$ as aggregation will on average outperform all other aggregation models. However, this is an artificial scenario. More realistic scenarios contain no optimal expert and more than one expert holds relevant information. Therefore, optimal weights will tend to deviate from $w_{BEM}$ in the direction of a more equal weighting, and there is a linear combination of expert judgments that on average outperforms the deterministic best expert.

**Proposition 4** assumes that an increasing crowd size impacts the performance of the WOC effect positively. Imagine a scenario with $|N|$ randomly characterized experts (i.e., experts with access to a random number of cues between zero and $J$ each). The probability $p_j$ of having access to a particular cue $j$ is the same for every expert and greater than zero. Therefore, with a probability of $(1 - p_j)^N$, the overall crowd does not have access to the cue at all. 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 be able to add new cues to the crowd's knowledge base. In the complementary event, he is not useful as a carrier of new information but can still reduce overall variance of the aggregated judgment. In extreme cases only, experts can decrease the crowd's performance (e.g., by being heavily biased). Altogether, a new expert generally increases the crowd's performance by adding new information or reducing judgment variance. The effect diminishes as the size of the crowd increases.

**Proposition 5** is a formulation of the assumption that WOC is based on maximizing the available information; it suggests that the performance of aggregation models is better when acting on a heterogeneous crowd. A heterogeneous crowd contains experts that are characterized differently (i.e., 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 very limited information pool. Like Proposition 4, we can reason that every expert added to the crowd has a positive probability of adding new cues to the crowd and thus increasing its performance if not all cues are already available in the crowd. If the crowd's information pool is already complete, a new expert might at least diversify the crowd's error.

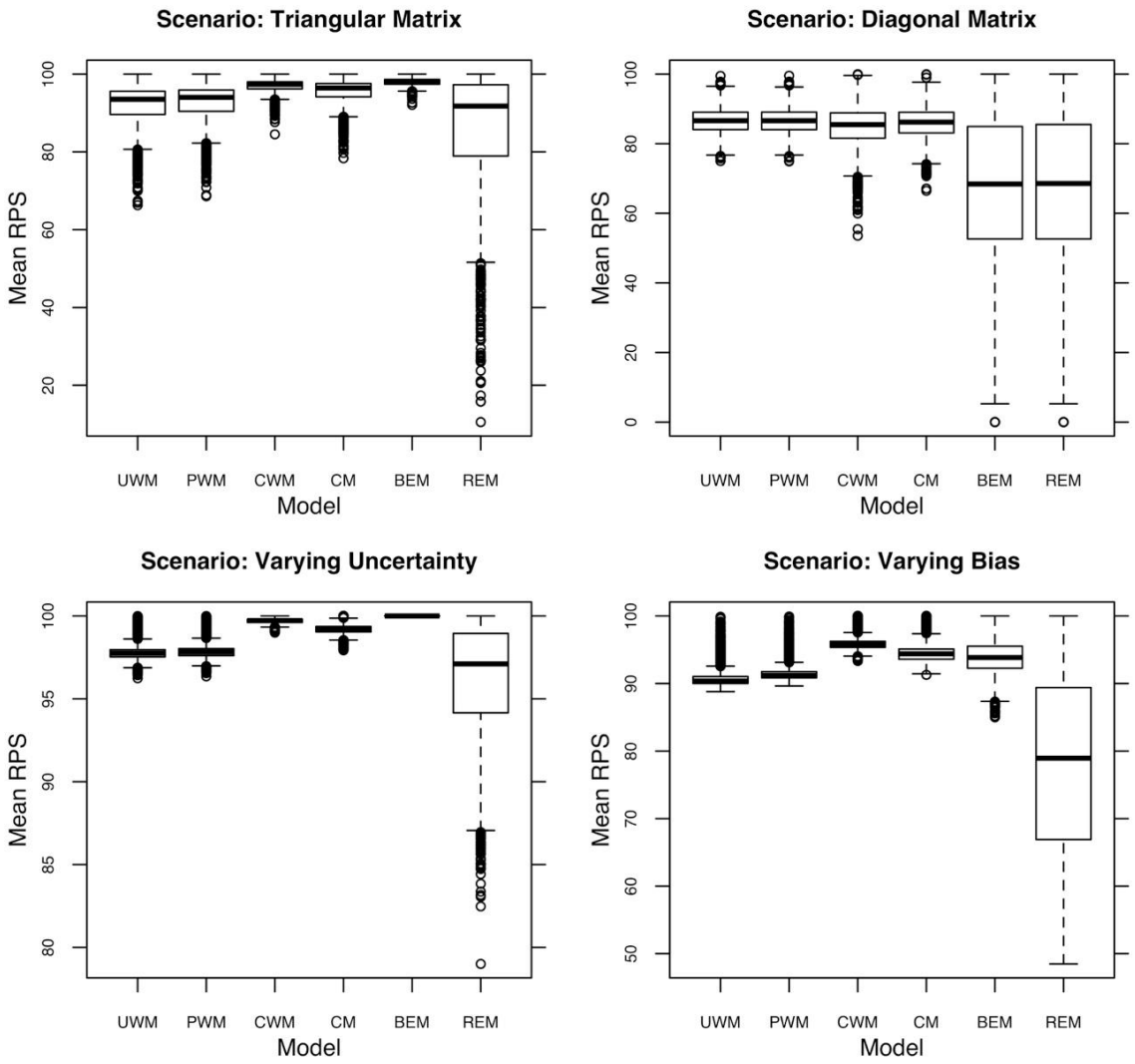


*Figure 2 - Aggregation model performance in different scenarios*

**Proposition 6** suggests that weighting models benefit primarily from selecting knowledgeable experts and only subordinately from the subsequent weighting. As such, selecting the right experts is generally more important than subsequently trying to additionally weight them according to their level of expertise. By means of simulation, we can confirm that the performance advantage of weighting models can largely be attributed to the selection of the experts. Therefore, we look at many different scenarios where there is heterogeneity of expertise in the crowds, 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 that of the UWM (Figure 2). We follow that the selection process is more important than the actual weighting of the remaining experts. Therefore, the proposition holds.

Finally, **Proposition 7** focuses on the model's ability to extract information on expert performance from historic events. We assume that the performance of history-based weighting models generally increases with the amount of available seed events (i.e., the RPS will increase, or the variance of the model's performance will decrease). Additionally, we expect the performance to converge asymptotically with increasing seed events because the measurement of an expert's historic 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.

| **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 |

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

In sum, all propositions for the WOC phenomenon hold true within our model in general. Therefore, it is reasonable to conclude that the conceptual model is valid.

## 6. Computerized Model Verification

Validating the correctness of the computerized model requires assurance that the computer programming and implementation of the conceptual model are correct (Kleijnen, 1995; Sargent, 1987, 2005). The computerized implementation of the conceptual 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 higher-order programming language Python, which is often used in statistics and simulation (Oliphant, 2007). Every component of the conceptual model as well as simulation functions have been mapped to separate modules in the computerized model, thereby ensuring program modularity. Every module has been tested thoroughly: First, all necessary simulation functions have been executed with dummy scenarios. Afterwards, all 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 simulation model. We can follow from the positive results that the computerized model is representing the conceptual model correctly. All information about module and method functionality 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 code will be provided as open source software upon publication of this paper.

## 7. Operational Model Validation

Operational validity is concerned with examining the model's behavior and applicability by ensuring that it creates accurate results that are helpful for the intended purpose. We follow the operational validation process as outlined in Sect. 2 by deriving extensive evidence for applicability and usefulness.

When applying judgment aggregation models, we are concerned with how the models function and how they perform. Additionally, the enhancement of existing and the creation of new models requires the identification and examination of hypotheses as cornerstones for new developments. Therefore, we define three important areas of application, and document the simulation model's applicability based on experimental scenarios.

Specifically, we show that the simulation model can be used to investigate the behavior of aggregation models (Sect. 7.1), to assess the performance of aggregation models under circumstances that are particularly hard to investigate empirically (Sect. 7.2), and to explore new hypotheses for further research (Sect. 7.3).

We use the simulation model to conduct experiments by artificially constructing specific scenarios and measuring the corresponding behavior of the aggregation models. For this purpose, we define a reference

setting for the experimental scenarios consisting of seed variables, the number of intervals $|I|$, and the number of simulation runs to create a sample of results; it defines the basic setting that we use for all experiments (unless explicitly specified otherwise), which ensures comparability of the different scenarios. The models can rely on 25 seed variables for each expert. We derive the selectable intervals from the range of the averaged 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 the two remaining intervals are open intervals towards $-\infty$ and $+\infty$, respectively. Again, we assume events to be unweighted averages of cues. We use statistical analysis to determine the minimum number of necessary simulation runs to obtain a sample of sufficient size given a detectable effect size (Faul, Erdfelder, Lang, & Buchner, 2007). If the analyzed behavior is compared between different scenarios, then 10,000 cycles are necessary. If not, then 3,000 cycles are sufficient. The event specifications, the size of the expert crowd, and the individual characteristics are defined per scenario, as they fundamentally define the experiments. This allows us to create a versatile range of artificial scenarios to examine and evaluate the WOC phenomenon, without being restricted to empirical events. Hence, the simulation model's ability as a tool to derive new research findings is shown.

### *7.1. Understanding Aggregation Models in Depth – Expertise Diversity and Seed Events*

One application of the simulation model is to further understand the particulars of aggregation models. As the inner workings of many aggregation models are difficult to understand from the outside, a deeper analysis is required (Clemen & Winkler, 1986). When creating a decision model based on an aggregation model, it is crucial to understand which aggregation model will make the best use of the scenario's characteristics (e.g. Hammitt & Zhang, 2013). For example, in a volatile environment, the CWM's reliance on historic information can constrain its performance, while the UWM might perform reasonably well. Besides the dependence on historic information in the form of seed events, the type and intensity of weighting, or the manner of expert selection are pivotal factors for understanding aggregation models (Table 1).

We create two scenarios, fashioned specifically to illustrate the discrepancies in the aggregation models' weighting. Each scenario consists of ten experts who only differ in the number of cues they have access to. The remaining parameters are set to their default values as previously introduced at the beginning of Sect. 7. The information matrix of the first scenario is a triangular matrix (i.e., expert $n \in \{1, \dots, 10\}$ has access to exactly $n$ out of ten cues) while the information matrix of the second scenario is a diagonal matrix (i.e., each expert has access to a different one of the ten cues). The triangular matrix portrays a heterogeneous distribution of expertise in the crowd, while the diagonal matrix depicts experts that are very similar in terms of expertise. Figure 3 shows the cumulative average weights for both scenarios as a function of the share of experts, sorted by the size of the weight and interpolated. For example, 20% of the experts in the triangular scenario possess almost 70% of the weights when using the CWM. Thus, a steep incline in the curve signals the allocation of substantial weight to only a few experts.

Since the UWM distributes equal weights to all experts independent of the scenario's characteristics, the cumulative weights always proceed linearly. The PWM assesses the experts' historic performance based on the RPS scoring rule and assigns weights accordingly. Under heterogeneous expertise, as in the triangular scenario, 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, leading to a much broader range of experts' average weights. 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, thus leading to a very specific behavior.

Figure 4 shows the corresponding performance scores. As high performance and low variability of performance are desirable, it suggests a clear ranking of aggregation models for the triangular matrix scenario with BEM being best, followed by CWM, CM, PWM, and UWM as worst model. In other words: The more unequal the weighting, the better the performance in this scenario.

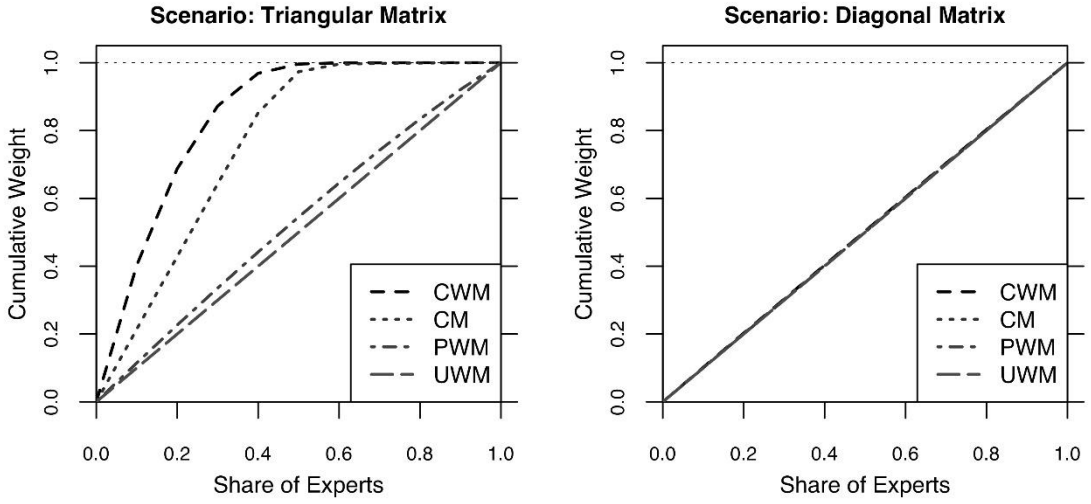


*Figure 3 - Cumulative weighting in two different scenarios*

Expertise and therefore experts' weights are clearly heterogeneous in the triangular scenario. However, in the scenario with a diagonal matrix, on average, all experts show equal performance and no superior expertise can be found. Consequently, on average, the weighting models (PWM, CWM, and CM) compute an equal weighting for all experts (Figure 3) and achieve mean performance scores similar to the UWM (Figure 4). However, while the average performance is comparable, the variance of the performance measures is much higher in scenarios with little to no differentiation in the experts' levels of expertise, like the diagonal scenario. This scenario leads to lower performance of models that select experts (CWM, CM, and BEM) compared to the UWM and PWM.

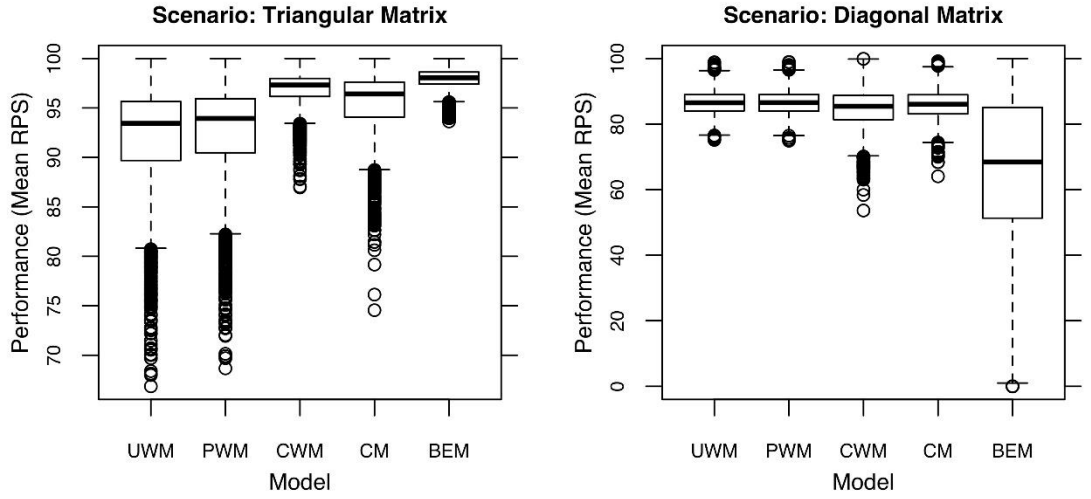


*Figure 4 - Boxplot of RPS values for two different scenarios (note different scaling of scores)*

We hence conclude that the overall performance of weighting models depends heavily on a certain variation in a crowd's expertise. If experts are similar, aggregation models using performance measures will often

perform worse than an equally weighted model since they sometimes falsely introduce a weighting, even though no expert shows superior expertise. Furthermore, the performance of the CM is strongly related to that of the CWM. Both models benefit from diverse crowds and will generally perform better than the UWM if there is special expertise to be found in the crowd. The BEM performs strongly if there are very good experts to be found in the crowd, but performs poorly when experts are not well-informed.

In a second step, we elaborate on the aggregation model's dependence on seed events as an information base. Within our selection of aggregation models, apart from the UWM, all depend on identifying good experts based on their historic performance. It is necessary for the models to have access to historic judgments to apply their specific performance measure. Consequently, the number of observable seed events influences the performance of the aggregation models (Cooke & Goossens, 2008; Eggstaff, Mazzuchi, & Sarkani, 2014) and therefore, information on the necessary quantity of seed variables is interesting. Budescu et al. (2016) find that for every random subset of only 15 out of 105 experts and 40 out of 86 seed events, the CWM performs roughly as well or even better than the maximal Brier Weighted Model score (PWM in our generalization). While this hints at the necessary number of seed variables, a deeper understanding of the coherences, especially concerning the CWM, still needs to be obtained.

We investigate this 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 multiple simulation runs, averaged across all experts. A low standard deviation signals that the aggregation models reliably calculate almost the same weights in every simulation run. We focus on two scenarios, each containing five experts: a diagonal information matrix and a matrix with two fully informed and three uninformed experts (block matrix). To limit computational complexity, we use several powers of 2 as seed amounts: 4, 8, 16, 32, 64, 128, and 256. We assume that higher numbers will seldom occur in a real-world context.

Table 3 shows the standard deviation of the CWM's weighting for each scenario and seed amount. We can observe that the standard deviations are generally lower for the block matrix because the uninformed experts are mostly deselected by the CWM, and their weights seldom deviate from zero. A trend in the data is visible: When quadrupling the amount of seed variables, the standard deviations decrease by at least 30%, on average even by 45%. Not surprisingly, more seed events are better for the CWM.

| **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 |
| **Block Matrix** | 0.025 | 0.020 | 0.006 | 0.004 | 0.003 | 0.002 | 0.002 |

*Table 3 - Standard deviation of the CWM's weighting, depending on the amount of seeds available*

Even for as many as 256 seed events, the weights calculated by the CWM in the scenario with the diagonal matrix 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.

In sum, this section focuses on the performance of history-based weighting aggregation models. It highlights drivers for weighting, including the diversity of expertise and the availability of seed events. The results indicate that a certain level of expertise diversity is essential to the good performance of weighting models that more seed events are better, and that diversity trumps the number of seed events.

These findings have not yet been established with empirical data, presumably as in real world judgment scenarios it is difficult to impossible to judge experts' expertise independent of their judgements.

### *7.2. Evaluating Model Performance – Structural Breaks*

The aggregation of multiple judgments is mainly concerned with improving overall judgment accuracy. In general, the literature on WOC, and aggregating judgments in particular, is highly concerned with quantitatively assessing the aggregation model's performance (Clemen, 1989). Based on the nature of WOC, the performance evaluation is foremost conducted empirically (e.g., Budescu et al., 2016; Clemen & Winkler, 1999; Cooke & Goossens, 2008). This is a sound approach for evaluating performance under externally given circumstances. In contrast, when measuring the performance of aggregation models for specific circumstances, this approach reaches its limits. Due to the uniqueness of the circumstances, empirical data is available very scarcely. A general example for such specific circumstances is the evaluation of structural breaks in a time series, namely situations where the underlying characteristics (that describe an industry, for example) change fundamentally and remain in this new state. With ever more quickly changing technological landscapes, fast moving industries and highly interlinked and volatile global financial markets, structural breaks are especially relevant in practice.

Looking at history-based aggregation models, one underlying assumption is that experts who performed well in the past are more likely to perform well in the future as they are deliberately assigned a higher weight. With structural breaks, this hypothesis might not always be true. Consider a market with an emerging technology where experts are called to provide judgments on the leading technology. Experts in that market might be clustered in two groups: 1 and 2. Experts in Group 1 might bet on the success of the existing technology, while experts in Group 2 bet on the emerging technology's success. While the emerging technology is still a niche product, experts favoring the old technology will deliver more accurate judgments. Yet as soon as the emerging technology has a breakthrough, it rapidly gains market share and eventually replaces the existing technology. Since this usually happens in a short period and experts tend to stick to their judgments, experts from Group 1 will now deliver weak judgments, while Group 2 delivers accurate ones. One example for such a structural break was the rise of digital photography (Lucas & Goh, 2009). Transforming this in terms of the underlying simulation model, we include two experts: 1 and 2; each expert represents one group. Prior to the structural break, Expert 1 is knowledgeable by observing one cue that has the same mean as the event. Expert 2 sees a cue that is far from the mean of the event. While both experts have access to the same number of cues, these cues differ in their current forecasting potentials. Thus, Expert 1's judgment is, on average, closer to the event realization than that of Expert 2 since Expert 2 bets on a technology that has not yet achieved substantial market penetration. After the structural break (the breakthrough), the characteristics of the two available cues switch, and the first cue is far from the event mean, while the second cue has the same mean as the event. Therefore, Experts 1 and 2 switch their degree of expertise. We acknowledge that studying only two experts is an extremely small expert panel. However, it simplifies the analysis and the same qualitative patterns emerge in larger, equally structured crowds.

Since history-based aggregation models use the existing seeds to weight the experts, the point in time of the structural break impacts the aggregated judgments. We thus evaluate the aggregation models´ performance depending on the point in time of the structural break. Figure 5 shows the mean RPS of the aggregation models as a function of the structural break's point in time. The model's performance is measured on their judgments for period $t = 0$.

The first structural break (at $t = -25$) is equivalent to the information switch before the first period included in the simulation model's history; as such, the models only observe the experts providing their post-structural-break estimates. Thus, the models behave as if there were no structural break. On the other side of the

spectrum (structural break at $t = -1$), the models only observe one historic period, which occurs after the structural break.

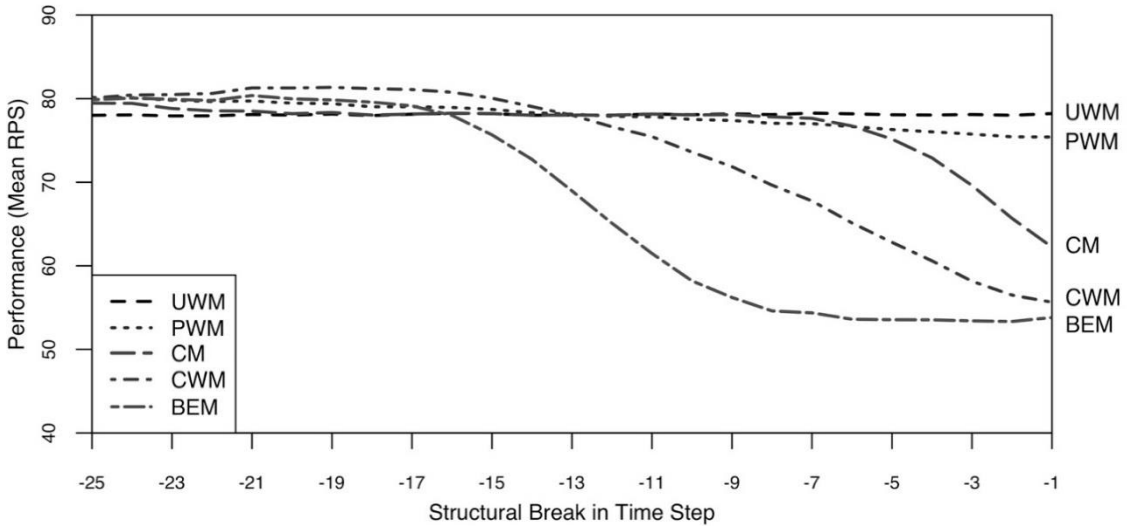


*Figure 5 – Impact of structural breaks on the performance of aggregation models*

In the graph, we can see that by design of the UWM, the performance of the history-independent UWM remains constant over time. As expected, on average, the history-dependent models show a decrease in performance the later the structural break takes place. Additionally, it can be seen that in the beginning the models are relatively close to each other, while their performance is strongly dispersed for later structural breaks. Among the history-based models, the extent to which their performance declines and the point where the decrease becomes substantial differ. The BEM is the first model to show a substantial drop in performance, followed by the CWM and CM. The PWM performs rather close to the UWM and is characterized by a constant, slight decrease. The same order holds for the absolute lowest performance the models reach in case of a structural break in period $t = -1$. Comparing the models to the UWM as a benchmark, the performance increase of history-based models, in case of an early structural break, is far lower than the performance decrease in case of a late structural break. Furthermore, in the beginning, the history-based models seem to be very close to each other, and the CWM can hold its performance advantage against the UWM longest.

The overall performance decrease of history-based models can be tied to their implied weighting. The later the structural break takes place, the more pre-structural break information is included in the calculation of weights. Since the performance of the experts switches after the structural break, the included information is flawed, and the models are allocating above average weight to an expert, who performs worse than experts with a lower weight after the break. This conclusively leads to a decrease in performance. The extent to which the models react to the structural break thus depends on the strength of their implied weighting. By their specific logic of creating weights, the intensity of the weighting differs substantially. The BEM, for example, puts all the weight on one expert and hence creates the most extreme weighting, which in this case leads to a rapid decline in performance in case of late structural breaks, and asymptotically reaches the lower boundary. Moreover, the experimental simulation brought an unexpected behavior of the CWM to light. Looking at periods t = -25 to t = -19, an increase in the mean RPS can be seen. This behavior is unexpected since the model can access the most representative data on the experts if the structural break takes place in period zero; thus, one would also expect performance to be highest under those circumstances. The later the structural

break takes place, the more that flawed information (the good historic performance of an expert who will be bad in the future) is incorporated into the calculations of the weights. Taking a closer look at the underlying weights, it becomes clear that the flawed information leads to a less extreme weighting, which results in a more moderate judgment and increases the average performance. Thus, it appears that in periods t = -25 to t = -19, the CWM suffers from overfitting and benefits from adding flawed information.

Again, these findings have not yet been established empirically, presumably as structural breaks only occur seldom, especially in combination with judgement data available from a crowd of experts.

### *7.3. Exploring New Hypotheses*

Experimentation is a core use case for the application of simulation models. Effective experimentation leads to the discovery and elaboration of new theories (Davis et al., 2007). Simulation methods enable experimentation across a wide range of conditions simply by changing the software code. By varying assumptions and values in our model, we identified two new hypotheses that need further exploration and understanding. These hypotheses focus on the optimal composition and characterization of expert crowds. First, we address a specific issue of the CWM, which can lead to a flawed assessment of expert performance and therefore impair the CWM's overall 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 an expert's performance relative to the crowd's performance by comparing the crowd's performance with the expert to the performance of the same crowd without that expert, using the unweighted mean as aggregation. Therefore, a reasonably good expert can still be deselected if the crowd without him is performing better than with him included. On the other hand, an uninformed expert can increase the crowd's performance by balancing a slight bias held by the majority of experts and might therefore be selected by the CWM. Imagine a scenario consisting of five experts and three cues $\{c_{t,1}, c_{t,2}, c_{t,3}\}$. Let one expert have access to only $c_{t,1}$ and the other four experts can be reasonably well-informed, but very similar to one another (access to $c_{t,2}$ and $c_{t,3}$ each). Thus, the first expert has access to a cue that nobody else has. In a scenario like this, we expect the CWM to distribute a bulk of the weight among the four well-informed experts, but still select the first expert because of his access to a rare cue. However, a simulation of this scenario results in CWM weights and model performance scores, as depicted in Figure 6.

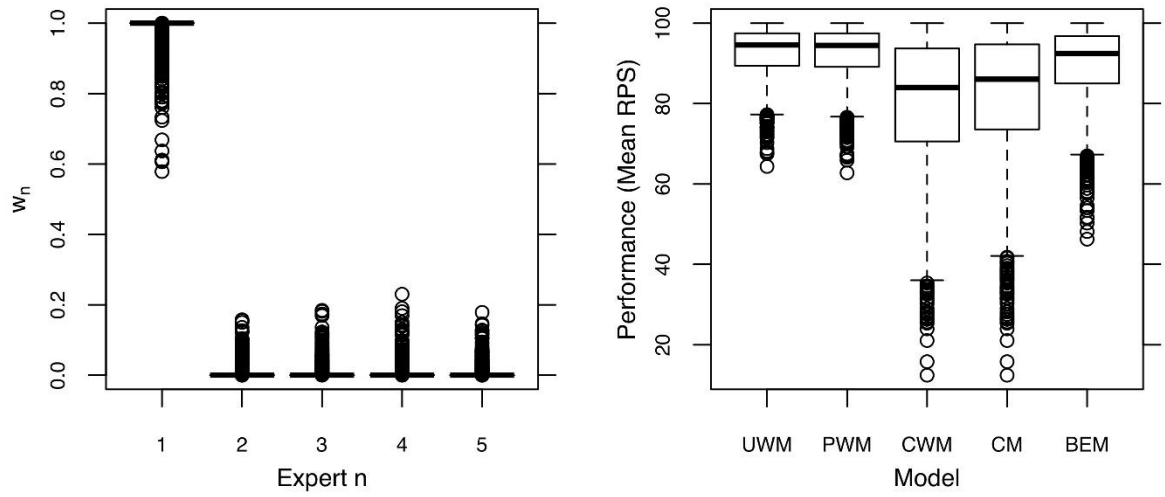


*Figure 6 – Weighting phenomenon of the CWM*

In 84% of simulation runs, the CWM allocates a weight of 1 to the first expert, while all well-informed experts are weighted with 0. Consequently, we can see that the performance of the CWM and the CM is, on average, considerably lower than that of the other examined models, as it mostly only uses the uninformed expert's judgment (Figure 6).

We hypothesize that this happens for two reasons. First of all, the contribution score of Expert 1 to the crowd is extremely positive as he adds information about a rare cue to the crowd's information pool. Therefore, from the CWM's perspective, selecting him and allocating a relatively high weight seems reasonable. Secondly, the contributions of the other four experts, as calculated by the CWM in the formulation by Budescu and Chen (2015), are mostly negative. Since they are very 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 and therefore to a weight of 0. Thus, the effect appears due to the isolated perspective of the CWM on a single expert's performance relative to the crowd. It will become stronger the more similarly experts are characterized and the stronger these groups of similar experts are in a crowd. Simultaneously, when experts are characterized diversely and independently, the effect will disappear.

In a second experiment, we want to gain insight into the experts' individual uncertainty $\sigma_n$ and how it affects the performance of expert judgments. We focus on the optimal individual uncertainty (i.e., the uncertainty value that maximizes an expert's individual judgment performance). First, we want to answer the question: Which factors influence the value of the optimal uncertainty and how strong is the impact on expert performance if we deviate from the optimal value? We build a scenario consisting of 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 $\mu_n$ and the number of available cues as described by $\sum_{j=1}^{|J|} \alpha_{n,t,j}$. Subsequently, we let the expert deliberately deviate from this calculated optimal uncertainty value to see how strong the impact of the uncertainty is on the expert performance. The optimal uncertainty values for each parameter constellation as well as the performance scores are shown in Figure 7. $\mu_n$ is defined within reasonable borders.

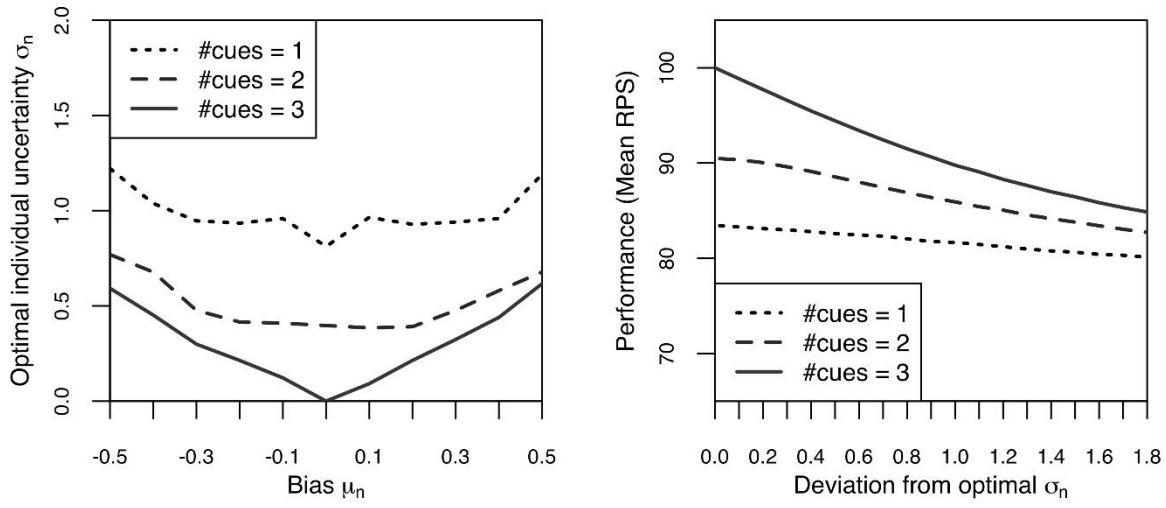


*Figure 7 – Optimal individual uncertainty and its implications*

Optimal uncertainty values 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 (as described in

Proposition 1). When deviating in one or both parameter dimensions (bias, cue access), the optimal uncertainty values become higher. In scenarios where an expert's stochastic judgment is far from the true value of the event realization through bias or missing cues, the expert benefits from more variance in the error term. This can be explained by the nature of the error term: The error term scatters in both directions. Thus, it can cancel out the deviation from the real value with a certain probability. 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 judgment performance becomes stronger, the better an expert is otherwise calibrated.

As before, these effects have not yet been demonstrated with empirical data, presumably as these settings do not occur that purely and it is difficult to impossible to disentangle justified judgment form idiosyncratic error in real-life settings.

## 8. Discussion and Conclusion

Simulation is an important toolkit in WOC research as data availability is a limiting factor. In our study, we propose a novel model to simulate expert density judgments, with the aim of shedding light on expert judgment aggregation models and the WOC phenomenon in general. To do so, we first deduct propositions on the WOC effect from the literature and design a model to simulate WOC scenarios. After completing all validation and verification steps (Conceptual Model Validation, Computerized Model Verification, and Operational Model Validation), we conclude that the conceptual model and its implementation are valid representations of the real-world problem entity. With its help, we then gain exemplary new insights into the field of WOC.

This paper contributes to WOC research in three major respects. First, we compile relevant literature on the subject into propositions that serve as axioms of the WOC effect. With their help, it is possible to reach a deeper understanding of WOC and its characteristics and influence factors. The propositions are also designed to act as a foundation for further WOC research and can be utilized as validation criteria for simulation models with a similar background.

Second, the conceptual simulation model is a novel representation of experts providing specifically density judgments. While major institutions like the ECB and other central banks are using density judgments as forecasting input, there is currently no fundamental simulation model available for this form of judgment. Our model sets itself apart from the density judgment model by Hammitt and Zhang (2013) who have only incorporated two experts with very special characteristics. 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 conceptual model to produce new findings in the field of WOC. For example, the model supports the iterative specification and testing of new aggregation models under a wide variety of potential circumstances. Furthermore, it enables researchers who want to understand, assess and compare existing aggregation models as it breaches boundaries imposed by relying only on empirical data.

Third, we conduct experiments to assess the operational validity of the model and hence derive new insights. The findings from those experiments are conclusive and build a deeper understanding of the overall judgment and aggregation process. We list existing aggregation models, both established (e.g. UWM and PWM) and relatively newly designed (e.g. CWM and CM), and identify their drivers for weighting and performance,

such as the diversity of expertise or the availability of seed events. When comparing their performance in a broad range of scenarios, some strengths and weaknesses in special situations (e.g., structural breaks) become noticeable. The degree to which aggregation models are influenced by structural breaks varies substantially. The more extreme an aggregation model weights the crowd members (assigning high weights to individual experts), the higher the performance decrease in case of a structural break. Additionally, the observed scenario implies a higher damage for weighting-based models in case of a recent structural break, in comparison to the benefits in the case of a very distant structural break. This analysis demonstrates that simulation of aggregation scenarios and models can trigger unexpected findings (e.g., potential overfitting by the CWM) and suggest routes for model improvements. We also conduct experiments to create new hypotheses on select WOC elements or aggregation models. For example, we demonstrate the CWM's difficulties in scenarios with many similar experts. Under those conditions, multiple 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 expert performance. Depending on an expert's quality (number of observed cues and bias), the ideal individual uncertainty differs. The less informed an expert is, the higher the ideal individual uncertainty. It is important to note that we have also tested whether optimal individual uncertainty maximizes crowd performance. A first experiment has yielded the result that depending on the aggregation model in use, there is evidence both for and against this statement. However, further research should be conducted to thoroughly answer the question.

To apply aggregation models in practice, we have shown that the choice of aggregation model depends highly on factors derived from the underlying scenario. These factors include expert characteristics such as individual uncertainty, crowd characteristics (such as diversity of expertise), and event characteristics (such as the availability of seed events or the probability of a structural break). These considerations, in combination with our simulation model, can help practitioners choose the right aggregation model for a specific WOC scenario. Furthermore, we have highlighted the potential risk of weighting-based algorithms with the example of structural breaks. 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 problem entity and therefore simplifies certain aspects of it. For the experiments, we chose to have the individual uncertainty and the cues be normally distributed. Furthermore, our model assumes that all cues are equally important and that an expert either has full access to a cue or none at all. The simulation of density judgments via multiple drawings from normal distributions implies high computational complexity, which also limits the intricacy of the model. This is particularly relevant when looking at rather computationally expensive aggregation models such as the CWM. In most of our experiments, we use extreme scenarios that are unlikely to occur in the real world and might limit the explanatory power of the results.

Future work should use the simulation model for further experimentation to broaden our knowledge base of the WOC phenomenon and aggregation models. 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 parametrized with common expert and crowd characteristics.

## 9. References


Ashton, A. H., & Ashton, R. H. (1985). Aggregating Subjective Forecasts: Some Empirical Results. *Management Science*, *31*(12), 1499–1508.

Bates, J. M., & Granger, C. W. J. (1969). The Combination of Forecasts. *Operations Research Society*, *20*(4), 451–468.

Bickel, J. E. (2007). Some Comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules. *Decision Analysis*, *4*(2), 49–65.

Bowles, C., Friz, R., Genre, V., Kenny, G., Meyler, A., & Rautanen, T. (2007). The ECB Survey of Professional Forecasters (SPF): A Review After Eight Years' Experience. *ECB Occasional Papers*, (59), 35.

Brailsford, T. J., & Faff, R. W. (1996). An evaluation of forecasting techniques. *Journal of Banking and Finance*, *20*, 419–438.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgement: A critical examination. *Organizational Behavior and Human Decision Processes*, *65*(3), 212–219.

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

Bröcker, J., & Smith, L. A. (2007). Scoring Probabilistic Forecasts: The Importance of Being Proper. *Weather and Forecasting*, *22*(April), 382–388.

Budescu, D. V., & Chen, E. (2015). Identifying Expertise to Extract the Wisdom of Crowds. *Management Science*, *61*(February).

Budescu, D. V., Chen, E., Lakshmikanth, S. K., Mellers, B. A., & Tetlock, P. E. (2016). Validating the Contribution-Weighted Model: Robustness and Cost-Benefit Analyses. *Decision Analysis*, (April).

Budescu, D. V. (2006). Confidence in aggregation of opinions from multiple sources. In *Information Sampling and Adaptive Cognition* (pp. 327–352). Fiedler K, Juslin P.

Carbone, R., & Armstrong, J. S. (1982). Evaluation of extrapolative forecasting methods: Results of a survey of academicians and practitioners. *Journal of Forecasting*, *1*(2), 215–217.

Chen, E., Budescu, D. V, Lakshmikanth, S. K., Mellers, B. A., & Tetlock, P. E. (2016). Validating the Contribution-Weighted Model: Robustness and Cost-Benefit Analyses. *Decision Analysis*, 1–26.

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

Clemen, R. T., & Winkler, R. L. (1986). Combining Economic Forecasts. *Journal of Business & Economic Statistics*, *4*(1), 39–46.

Clemen, R. T., & Winkler, R. L. (1999). Combining Probability Distributiond from experts in Risk Analysis. *Risk Analysis*, *19*(2), 155–156.

Cooke, R. M., & Goossens, L. L. H. J. (2008). TU Delft expert judgment data base. *Reliability Engineering and System Safety*, *93*(5), 657–674.

Dalrymple, D. J. (1975). Sales forecasting methods and accuracy. *Business Horizons*, *18*(6), 69–73.

Dana, J., Broomell, S. B., Budescu, D. V., & Davis-Stober, C. P. (2015). The composition of optimally wise crowds. *Decision Analysis*, *12*(3), 130–143.

Davis-Stober, C. P., Broomell, S. B., Budescu, D. V., & Dana, J. (2014). When is a crowd wise? *Decision*, *1*(2), 79–101.

Davis, J. P., Eusebgardt, K. M., & Binghaman, C. B. (2007). Developing Theory Through Simulation Methods. *Academy of Management Review*, *32*(2), 480–499.

de Menezes, L. M., W. Bunn, D., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, *120*(1), 190–204.

Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2014). The effect of the number of seed variables on the performance of Cooke's classical model. *Reliability Engineering and System Safety*, *121*, 72–82.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*(1), 158–172.

Epstein, E. S. (1969). A Scoring System for Probability Forecasts of Ranked Categories. *Journal of Applied Meteorology*.

European Central Bank. (2017). *ECB Survey of Professional Forecasters*. Retrieved from https://www.ecb.europa.eu/stats/prices/indic/forecast/shared/files/reports/spfreport2017_Q1.en.pdf?00bc8eef81aba4266fcece27ec87d64b on Sep 29, 2017, ISSN 2363-3670, EU catalogue No QB-BR-17-001-EN-N.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–91.

Fildes, R., & Hastings, R. (1994). The Organization and Improvement of Market Forecasting. *The Journal of the Operational Research Society*, *45*(1), 1–16.

Fischer, G. W. (1981). When Oracles FaiI-A Comparison of Four Procedures for Aggregating Subjective Probability Forecasts. *Organizational Behavior and Human Performance*, *110*, 96–110.

Flandoli, F., Giorgi, E., Aspinall, W. P., & Neri, A. (2011). Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. *Reliability Engineering and System Safety*, *96*(10), 1292–1310.

Galton, F. (1907). Vox populi - The wisdom of crowds. *Nature*, *75*, 450–451.

Genest, C., & Zidek, J. V. (1986). Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statistical Science*, *1*(1), 147–148.

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Hammitt, J. K., & Zhang, Y. (2013). Combining Experts' Judgments: Comparison of Algorithmic Methods Using Synthetic Data. *Risk Analysis*, *33*(1), 109–120.

Harling, J. (1958). Simulation Techniques in Operations Research - A Review. *Operations Research*, *6*(3), 307–319.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*(2), 494–508.

Herzog, S. M., & Hertwig, R. (2011). The wisdom of ignorant crowds : Predicting sport outcomes by mere recognition. *Judgment and Decision Making*, *6*(1), 58–72.

Hogarth, M., & Makridakis, S. (1981). Forecasting and Planning: An Evaluation. *Management Science*, *27*(2), 115–138.

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, *21*(1), 40–46.

Hurley, W. J., & Lior, D. U. (2002). Combining expert judgment: On the performance of trimmed mean vote

aggregation procedures in the presence of strategic voting. *European Journal of Operational Research*, *140*(1), 142–147.

Jouini, M. N., & Clemen, R. T. (1996). Copula Models for Aggregating Expert Opinions. *Operations Research*, *44*(3), 444–457.

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

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: a meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404–426.

Keuschnigg, M., & Ganser, C. (2017). Crowd wisdom relies on agents' ability in small groups with a voting aggregation rule. *Management Science*, *63*(3).

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

Kleijnen, J. P. C. (1995). Verification and validation of simulation models. *European Journal of Operational Research*, *82*, 145–162.

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2011). The Social Psychology of the Wisdom of Crowds. In *Social Psychology and Decision Making* (Ed.: J.I. Krueger), pp. 227–242, New York.

Law, A. M., & Kelton, D. W. (1991). *Simulation Modeling & Analysis*. (M.-H. I. E. I. E. Series, Ed.) (Second Edi). Singapore.

Lawrence, M., Goodwin, P., O'Connor, M., & Oenkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*(3), 493–518.

Lee, M. D., Zhang, S., & Shi, J. (2011). The wisdom of the crowd playing The Price Is Right. *Memory & Cognition*, *39*(5), 914–923.

Leimeister, J. M. (2010). Collective Intelligence. *Business & Information Systems Engineering*, *2*(4), 245–248.

Lucas, H. C., & Goh, J. M. (2009). Disruptive technology: How Kodak missed the digital photography revolution. *Journal of Strategic Information Systems*, *18*(1), 46–55.

Mahajan, V., & Wind, Y. (1988). New product forecasting models. *International Journal of Forecasting*, *4*(3), 341–358.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–99.

McKenzie, C. R. M., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior and Human Decision Processes*, *107*(2), 179–191.

Murphy, A. H. (1970). the Ranked Probability Score and the Probability Score: a Comparison. *Monthly Weather Review*, *98*(December), 917–924.

Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science and Engineering*, *9*(3), 10–20.

Petrovic, D., Roy, R., & Petrovic, R. (1998). Modelling and simulation of a supply chain in an uncertain environment. *European Journal of Operational Research*, *109*(2), 299–309.

Sanders, N. R. (1997). The status of forecasting in manufacturing firms. *Production and Inventory Management Journal*, *38*(2), 32–35.

Sargent, R. G. (1987). An overview of verification and validation of simulation models. *Proceedings of the*

*19th Conference on Winter Simulation - WSC '87*, 33–39.

Sargent, R. G. (2005). Verification and Validation of Simulation Models. *Proceedings of the 37th Winter Simulation Conference (WSC'05)*, 130–143.

Scott, J. (1981). The probability of bankruptcy. A comparison of empirical predictions and theoretical models. *Journal of Banking and Finance*, *5*(3), 317–344.

Slack, N., Chambers, S., & Johnston, R. (2007). *Operations Management* (5th ed.). Essex: Pearson Education Limited.

Surowiecki, J. (2005). *The Wisdom of Crowds*. New York: Anchor Books.

Tay, A. S., & Wallis, K. F. (2000). Density Forecasting: A Survey. *Journal of Forecasting*, *19*, 235–254.

Teigen, K. H., & Jørgensen, M. (2005). When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology*, *19*(4), 455–475.

Urban, G. L., Weinberg, B. D., & Hauser, J. R. (1996). Premarket Forecasting of Really-New Products. *Journal of Marketing*, *60*(1), 47–60.

Van Wesep, E. D. (2016). The Quality of Expertise. *Management Science*, *62*(10), 2937–2951.

Wagner, C., & Suh, A. (2014). The wisdom of crowds: Impact of collective size and expertise transfer on collective performance. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 594–603.

Wagner, C., & Vinaimont, T. (2010). Evaluating the wisdom of crowds. *Proceedings of Issues in Information Systems*, *XI*(1), 724–732.

Winkler, R. L., & Makridakis, S. (1983). The Combination of Forecasts. *Journal of the Royal Statistical Society*, *146*(2), 150–157.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688.

Yates, J. F., McDaniel, L. S., & Brown, E. S. (1991). Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise. *Organizational Behavior and Human Decision Processes*, *49*(1), 60–79.