Theory Dec. (2018) 85:99–116
https://doi.org/10.1007/s11238-017-9643-7

# The wisdom of collective grading and the effects of epistemic and semantic diversity

**Aidan Lyon**[1] · **Michael Morreau**[2]

Published online: 4 December 2017


**Abstract** A computer simulation is used to study collective judgements that an expert panel reaches on the basis of qualitative probability judgements contributed by individual members. The simulated panel displays a strong and robust crowd wisdom effect. The panel's performance is better when members contribute precise probability estimates instead of qualitative judgements, but not by much. Surprisingly, it does not always hurt for panel members to interpret the probability expressions differently. Indeed, coordinating their understandings can be much worse.

**Keywords** Qualitative probability · Scores and grades · Wisdom of crowds · Computer modeling

## 1 Introduction

It is well known that collective judgements can be better than individual ones. Say you want to know how many jelly beans there are in a jar. Often you will do better to take the average of your friends' guesses rather than just rely on your own judgement; some will guess too high, others too low, and the errors will cancel out. Unless you're a highly skilled jelly-bean estimator, the average is likely to be more accurate than your own judgement, or that of anyone else in the group. This effect is known as the *wisdom of crowds* or *collective wisdom* (Surowiecki 2004; Page 2008).

The jelly beans are of course just a toy example. Collective wisdom has been documented in many different tasks, from Francis Galton's classic example of guessing

✉ Michael Morreau
michael.morreau@uit.no

1 University of Maryland at College Park, College Park, MD 20742, USA

2 UiT-The Arctic University of Norway, 9037 Tromsø, Norway

the weight of an ox at a county fair (Galton 1907) to forecasting geopolitical events (Mellers et al. 2014) and medical diagnosis (Wolf et al. 2015). The kind of judgement varies from case to case (Lyon and Pacuit 2013). In Galton's example, the individual inputs are point estimates ("The ox weighs 200 kg"). With geopolitical events they might be subjective probabilities ("I'm 80% confident that the Democrats will win the next election"); with medical diagnosis, categorical judgements ("This patient has Lyme disease"). The aggregation method also differs: you can take the *median* of point estimates, the *average* of probabilities and a *super majority* of diagnoses. Collective wisdom arises in many different contexts but just *how* it arises varies from case to case (Lyon forthcoming).

This article is about collective wisdom in groups whose members express themselves using scores and grades. These are coarse-grained expressions such as the numerical scores 1–6 used by the Arts and Humanities Research Council (AHRC) in Britain for evaluating research proposals. Other examples are qualitative probability expressions such as *probable*, *tossup* and *unlikely*, and the letter grades used in academic evaluation around the world. Characteristically, scores and grades come with an ordering from "top" to "bottom": a 6 in the sense of the AHRC is better than a 5; a *probable* event is more likely than an *unlikely* one; a *C* is a higher grade than a *F*, and so on.

Scoring and grading are common in juries, committees and panels. Naturally they have attracted the attention of theorists of social choice. Much work in this field has an "axiomatic" focus, being concerned with the formal properties of various aggregation methods (List 2013). Accordingly, one main goal of work on collective grading has been to characterize particular methods for aggregating grades in terms of the axioms they satisfy (see for instance Aleskerov et al. 2007; Gaertner and Xu 2012). Another goal has been to improve on evaluation procedures currently in use by introducing new and better aggregation methods. Thus, Balinski and Laraki (2007, 2011) propose the method of *majority judgement*, a generalization of taking medians which they recommend not only to expert panels judging wines and performance in sports and the arts, but also as a replacement for traditional voting methods in political elections. The idea of grading candidates in elections is also found in the earlier idea of approval voting (Brams and Fishburn 1978; Alcantud and Laruelle 2014). In a different line of enquiry, collective choice procedures are evaluated accordingly as they "track the truth" about the worth of the alternatives under consideration, in some choice-independent sense (Beisbart et al. 2005; Beisbart and Bovens 2007; Beisbart and Hartmann 2010). Thus, Pivato (2016) states conditions under which approval voting, in particular, may be expected to maximize utilitarian social welfare.

Wide ranging though it is, none of this literature on scoring and grading in economics and political theory takes up the matter of collective wisdom, and how its emergence depends on details of people's interpretations of the scores and grades in which they express their inputs. That is our topic here.

One prominent feature of natural languages is their *contextuality*: different people interpret words differently, and the same people interpret them differently on different occasions. Languages of scores and grades are to a large extent continuous with natural languages, and they too are contextual. We may expect that, as a result, groups of graders will often be *semantically diverse*. That is, they will include people with

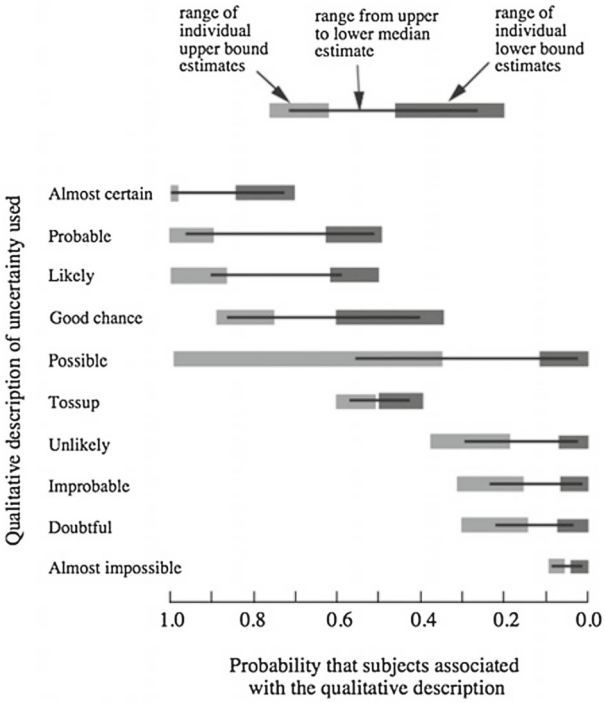


**Fig. 1** People can mean very different things by qualitative judgements of probability such as *probable*, *possible*, and *doubtful*. (Figure from Wallsten et al. 1986, redrawn by Morgan 2014)

different interpretations of the grades. In the case of probability grades, in particular, great semantic diversity has in fact been documented among doctors and their patients (Ohnishi et al. 2002), members of a science panel (Wardekker et al. 2008) and students of business and the social sciences (Fig. 1, Wallsten et al. 1986; Morgan 2014).

Interpersonal differences in perspectives and cognitive style are known to improve the judgement and decisions of groups (Surowiecki 2004; Page 2008; Tetlock 2005; Nielsen 2011). This *epistemic diversity* is not the same thing as diversity in age, gender, cultural and linguistic background, or other features that determine people's social identities, but the two go together.

Epistemically and socially diverse groups will often also be semantically diverse, though, and resulting misunderstandings might be expected to pull in the opposite direction. Certainly there is support in the literature for the idea that differences in people's understanding of scores and grades must create trouble. Thus, Hubbard and Evans (2010) cite "variability of verbal labels" as a problem in qualitative assessment of risk. Balinski and Laraki (2011) state as a requirement for using their method of majority judgement that the members of the group must share a "common language" of grades. Morreau (2016) argues that interpersonal differences in grading thresholds can make collective judgements "unsound," in a technical sense to be elaborated soon.

We consider the possibilities for collective wisdom in a quite common kind of task: selecting the top $k$ of some given $n$ ordered items. A committee or panel can approach

this task as follows. First, each member grades each of the items under consideration. The vocabulary of grades they use is fixed in advance, but different members might interpret them differently. For each item a collective grade is determined on the basis of the grades assigned to it by all members of the group; then all of the items are ranked according to their collectively assigned grades, and $k$ items are chosen from the top of the group's ranking. The *epistemic performance* of the group—its capacity to track the truth about which $k$ to choose—is measured by calculating how often, in a large number of trials, the $k$ chosen events really are to be found at the top of the actual ordering of the $n$ items. In this way we can study how features of grading languages and their interpretation by group members—such as how many labels there are, and whether everyone has the same understanding of them—affect epistemic performance.

Many factors may be expected to affect epistemic performance in addition to similarities and differences in members' understandings of grades. These include the size of the group and the individual expertise of its members. To study the sometimes complex interactions between these and other factors we approach our topic using, in addition to analytic methods, a computer simulation.

Our findings are surprising. One main conclusion is that epistemic performance can be very good even with a lot of semantic diversity. Indeed, other relevant factors being equal, having a common interpretation of grade expressions can seriously *depress* the performance of the group. Differences in understanding can under realistic conditions actually sharpen up the epistemic edge of groups.

There seem to be far-reaching consequences for the design and training of juries, committees and expert panels throughout society. It might for instance be counterproductive for members of a hiring committee to discuss and agree among themselves, say, what it is for a candidate to be *excellent*, or merely *good* or *fair*. Better for them to skip the coordination and work with whatever diverse understandings they happen to bring to the group—better, even, to *increase* semantic differences among members artificially, perhaps by exploiting the familiar contextuality of scores and grades. These differences tend to create disagreements that are, in part, verbal. Even so, they can help the group to find out which of the candidates are better than which.

We proceed as follows. In Sect. 2, we first discuss some of the main features of grades that suit them well to individual and collective judgement. Then we show how semantic diversity results in violations of Morreau's condition of soundness, but counter that this condition is much too strong to impose as a general requirement on acceptable procedures for collective grading. This prepares the ground for Sect. 3, where we set up the model of collective grading behind our simulation of risk panels. In Sect. 4, we present observations of the simulated panels. We comment as well on the effects of varying certain parameters in the simulation, and on several elaborations of the basic model. Section 5 sums up.

## 2 Pros and cons of grading

Allowing panel members to use qualitative language helps them to make probability judgements that are both confident and timely. We often cannot say *exactly* how probable events are. Sometimes that's because relevant information is unavailable, or

because expertise is lacking, and sometimes it's just because there's no time to find out. Even so, we might be able to say that there's a *good chance* of the one event occurring, while the other is a *tossup*. The reason is that these qualitative expressions are coarse grained, each covering a *range* of precise probabilities. A panel member might be reasonably confident that the probability of some event falls within a given range though unable to come up with an exact number.

But qualitative language does not only help *individuals* to make judgements. It also enables groups of people to speak, as it were, with a single voice. To see why, suppose that instead of eliciting qualitative judgements we ask panel members simply to *rank* events by their probabilities. That is another possibility when, for whatever reason, precise estimates cannot be had. Then there is a well-known theoretical obstacle to aggregating the several individual inputs into a single collective ranking. The Marquis of Condorcet's (1785) "paradox of voting" tells us that, depending on what the individual rankings happen to be, majority decisions about which events are more probable than which cannot be relied on to produce an ordering of all the events. Arrow's (1951) "impossibility" theorem tells us that the problem here is not some quirk of the particular aggregation method, pairwise majority voting, since no procedure *whatsoever* for deriving a single "social" ranking from several individual rankings meets a short list of seemingly mild conditions.

When group members grade things instead of ranking them, on the other hand, there are different ways to proceed. Suppose for each item under consideration we somehow gather together all the individually assigned grades into a collective grade. Then a collective ranking of all the events can be read off from their collective grades. There are different ways of determining the collective grades. One is to take medians. The aggregation method used in our simulated expert panel in Sect. 3 is based on this idea, and Balinski and Laraki's "majority judgement" is a sophisticated elaboration of median taking. So it is that grading enables groups of people to reach judgement, first individually and then collectively.

That grading opens up possibilities that are unavailable when panel members instead are invited to contribute weak orderings might seem surprising. After all, assignments of grades correspond directly to weak orderings of the graded items. However, because the set of grades is finite and the same for all members of the panel, the weak orders arrived at by grading are of a special sort: there is a limit to the number of equivalence classes, the same one for all inputs. This amounts to a "domain restriction", in the sense of social choice theory. In addition, the grades provide a way of bringing into correspondence with one another the equivalence classes of different people's weak orderings. This means that there's more information in grades than in the weak orderings of Arrow's framework.

Aggregating probability grades is under some special circumstances a good way to get information about probabilities from the members of a panel. When panel members have different interpretations of the applicable expressions, though, it can be quite misleading. Collective judgements based on the probability grades that panel members submit can fail to track the private opinions on which those grades are based. Examples adapted from Morreau (2016) illustrate.

Consider a panel of two experts that is to rank some given events by their probability. The panel approaches its task as follows. First, each member publicly assigns to each of

the events under consideration one of three available probability grades: *good chance*, *tossup* and *unlikely*.[1] They do this on the basis of precise individual estimates of the probabilities.[2] Next, a collective grade is determined for any given event using the following method of *splitting the difference*. If both panel members assigned the same grade then that is the collective grade of this event as well. If one member counted it a *good chance* and the other an *unlikely*, its collective grade is the intermediate *tossup*. Otherwise, the event receives one of two special grades reserved for compromising. These are $g \cdot t$, intermediate between *good chance* and *tossup*, and $t \cdot u$, between *tossup* and *unlikely*. Whenever one grader says *good chance* and the other *tossup* the collective grade is $g \cdot t$; *tossup* and *unlikely* come to $t \cdot u$. Finally, the panel's ranking $\succeq$ is read off from the collective grades: $x \succeq y$ if the collective grade of $x$ is at least that of $y$, perhaps better. $\succeq$ is a collective ranking based on the panel members' public descriptions of the events in qualitative terms.

Additionally, let $x \geq y$ mean that the average of the individual graders' precise estimates of the probability of $x$ is at least as great as the average of their estimates for $y$. This second relation $\geq$ compares events by how probable the panel members collectively think the events are, by some reckoning that gives equal weight to their opinions.

Finding out what the panel members think was the point of eliciting the grades from them. So, it is to be hoped that the ranking $\succeq$, determined from the elicited grades by splitting the difference, is true to $x \geq y$, which tracks what the panel members actually think. What's wanted, more precisely, is that for any $x$ and $y$:

If $x \succ y$ then $x > y$.[3]

Morreau (2016) introduces a technical notion of *soundness* as a criterion for evaluating grade-aggregation rules. Intuitively, a rule is sound if rankings determined on the basis of collectively assigned grades must always agree with the result of aggregating the underlying estimates of the individual graders. We can be sure that $x > y$ whenever $x \succ y$ if the method of splitting the difference is sound with respect to averaging precise probability estimates.[4]

Our first example shows that aggregating grades can indeed be a reliable way to get at what the panel members think. Suppose both members share the following interpretation of the applicable probability expressions. The common threshold for *good chance* is 0.8, on the probability scale 0-1. That is, each grader counts an event as a *good chance* if he thinks its probability is greater than 0.8. An event counts as

[1] It's simpler with just three grades but this example generalizes naturally to any other finite number.

[2] Assume for the sake of the example that these are not public, perhaps because they are not even consciously accessible to the panel members themselves.

[3] Here, $x \succ y$ just means that $x \geq y$ but not $y \geq x$; $>$ is defined similarly in terms of $\geq$. Notice that we should not hope that, in addition, whenever $x > y$ also $x \succ y$. The reason is just that grades are coarse-grained. For instance, two events $x$ and $y$ can receive the grade *unlikely*, both individually and collectively, even though both panel members think that one of them, say $x$, is a little more likely than the other. In this case $x > y$ but not $x \succ y$.

[4] Compare Morreau (2016), Definition 10 clause (b).

a *tossup* if its probability is greater than 0.4 but no greater than 0.6, and an *unlikely* event, as far as the two graders are concerned, is one whose probability is up to 0.2.[5]

With this common interpretation of the applicable probability expressions, if $x \succ y$ then $x > y$. To see that it is so, suppose $x \succ y$. That is, suppose $x$'s collective grade is higher than $y$'s. Say $x$ has a collective $g \cdot t$ and $y$ a collective *tossup* (the reasoning is similar for any other pairs of grades resulting in $x \succ y$ so we consider only this case). Inspection of the aggregation procedure tells us that one panel member counted $x$ a *good chance* and the other a *tossup*. Additionally, either both members counted $y$ a *tossup*, or else one of them counted $y$ a *good chance* and the other an *unlikely*. Given the common interpretation of the three grades, the collective estimate of the probability of $x$ is the average of some number above 0.8 and another above 0.4. It has to be above 0.6. The collective probability estimate for $y$ on the other hand is either the average of two numbers no greater than 0.6, or else the average of one number no greater than 1.0 and another no greater than 0.2. Either way, it can't be above 0.6. Whatever the numbers happen exactly to be, then, $x > y$.

The next example shows how semantic diversity can result in unsoundness.

For realism, suppose the events in question are threats. Mr. Queasy has a low tolerance for these. His threshold for saying there's a *good chance* of any given threat materializing is just 0.35. ("Better safe than sorry!") Ms. Breezy on the other hand has a much higher threshold: for her, a *good chance* means the probability is at least 0.6. An event counts as *unlikely* all the way up to probability 0.38, as far as she is concerned. ("Chill, monkey buddy!") Now Queasy thinks the probabilities of some particular threats, call them $x$ and $y$, are around 0.36 and 0.62, respectively. His input to the panel is therefore that there is a *good chance* of each. Breezy for her part privately puts them around 0.61 and 0.37, respectively. While agreeing with Queasy that there is a *good chance* of $x$, she says $y$ is *unlikely*. Aggregating as before, $x$'s two *good chance*'s come to a collective *good chance*, while $y$'s *good chance* and an *unlikely* come to a collective *tossup*, so: $x \succ y$. Averaging the individual probability estimations, though, it's just the other way around: $y > x$.

| | $x$ | $y$ |
|---|---|---|
| Queasy | 0.36, good chance | 0.62, good chance |
| Breezy | 0.61, good chance | 0.37, unlikely |

This example shows that when there is semantic diversity within a group, collective qualitative judgements can misrepresent what the members really think. That surely must tend to defeat an important purpose of evaluating in groups: getting better judgements by pooling input from people with different perspectives. Notice that the two estimates of $y$'s probability (0.37 and 0.62) dominate those of $x$ (0.36 and 0.61). Arguably Queasy and Breezy collectively count $y$ more probable than $x$ on *any* way of reckoning their collective estimation that treats them as equals—not just averaging, as in the example. Notice, also, that interpretations as radically different as theirs are

[5] We assume that, in the nature of the particular case, the panel will be able to make do with just these three grades—say because it is expected apriori that the events under consideration fall into one of these categories.

not unusual and have been documented even within groups that are semantically and culturally quite homogeneous. We adapted Queasy and Breezy's interpretations of *good chance* and *unlikely* from Fig. 1.

Morreau (2016) studies the semantic diversity through the lens of social choice theory, by extending Arrow's (1951) framework to collective grading problems. When there is extreme uncertainty about the extent of semantic diversity, he shows, collective grading is, in a precise sense, *vacuous*: under seemingly mild conditions it cannot be counted on to track the members' collective opinion on *any* acceptable way of reckoning this at all, whether averaging or any other. Morreau makes this point by showing that while collective grading can meet analogues of all conditions of Arrow's "impossibility" theorem, when too little is known about people's interpretations it is not possible to meet in addition to them a further condition of *soundness*. The example of Queasy and Breezy is a case in which the grade aggregation procedure of Sect. 2 is not sound, in Morreau's sense, with respect to averaging individual estimates.

But Morreau's soundness requirement is very demanding. Often it will not be critical to make the very best possible use of information distributed among panel members. It's important to do a good job when evaluating project proposals for instance, but not at all costs. In the remainder we impose no requirement of soundness. Grading thresholds will be allowed to vary among different members of the group, even quite radically, so mismatches like the one we saw with Queasy and Breezy will occur. They might depress epistemic performance to some extent, but provided they are not too common, perhaps not very much. To see what sort of performance may be expected from semantically diverse panels we turn now from the analytic method familiar from social-choice theory to a different approach, and one that has proven useful for studying complex social phenomena: computer simulation.

The next section introduces a model of an expert panel whose task is to select the most probable $k$ of $n$ events. Section 4 reports observations of the simulated panel under different settings of parameters that may be expected to affect its performance. These include, crucially, the extent of semantic diversity in the group.

## 3 The model and simulation

We model an expert panel with a simple but important kind of task: selecting the most probable $k$ of $n$ events. The panel approaches this task in the way set out in the introduction. A language of probability grades is fixed. Each panel member assesses all the events and assigns to each one a probability grade. The different grades assigned to any given event by all the panel members then determine its collective grade. The collective grades amount to a collective (weak) ordering of the events. Finally, $k$ events are chosen from the top of the collective ordering. The performance of the panel is a matter of how likely it is to choose some $k$ events that are in fact among the most probable events.

To simulate a panel of this sort we need to fix some further assumptions.

First, the grades. We assume throughout that these are *absolute*. This means that, once we have fixed an interpretation for some label, whether or not it is correct to apply this label to any given item is independent of which other items happen to be

under consideration along with it.[6] Absolute grading has certain advantages when working in groups. For one thing, it makes it possible to divide up the task, letting some people work on some of the items, others on others, because their results can simply be merged. Balinski and Laraki (2011, p. 185) argue that absolute grades are needed when we come to aggregate inputs from different voters in a political election.

Second, the events. We assume there are $n = 100$ of them, and that the number to be selected, $k$, is 10. For simplicity we assume to begin with that the distribution of their probabilities is approximately Gaussian, with a mean of 50% and a standard deviation of 30%, and truncated to the probability scale. For now this assumption is part of our "base model" but we relax it in the next section, where we consider a distribution derived from real data.

Next, the panel members. We assume that they assign grades on the basis of their (i) individual estimates of the probabilities of the events and their (ii) individual interpretations of the grades, which specify in absolute terms upper and lower thresholds for applicability. For example, a panel member who understands the top label *A* (or "likely"[7]) to mean at least 80% will award an *A* to any event he estimates to have a probability greater than 80%. These assumptions are compatible with our stipulation that grades are absolute. Different panel members can have different interpretations of the grade labels. That is semantic diversity. But once a label has been interpreted as one or the other interval of probability—once it has come to express a grade—its applicability to an event is determined entirely by the probability of this event, and that is independent of whichever other events happen also to be under consideration.

The estimates in (i) are assumed to be completely precise and somewhat noisy. Both assumptions contain idealizations. Concerning precision, it would be quite unrealistic to suppose that people can actually produce precise estimates when asked for them. That is not our assumption, though. Rather, the individual precise estimates are theoretical entities of our model; by analogy with the physical fiction of "center of mass," we can think of them as the locations of imprecise estimates. The noise in the precise estimates is for each member of the simulated panel assumed to be Gaussian with a mean of 0% and a standard deviation of 10% (i.e., 1/3 of the standard deviation of the event distribution). Thus, we assume in effect that all panel members have the same level of expertise and that all are unbiased, in that the expectation of their estimates of the risk of any given event is its true risk.[8]

Consider the graders now as a *collective*—as the panel they together make up. We assume that the collective assigns a grade to each event. The collective grade of an event is the median of the grades assigned to it by the individual members of the panel.[9] These collective grades determine a ranking of all the events, and this is used

[6] The alternative is *relative* grading, as for instance when each panel member first ranks all of the events he's been given to consider as best he can from most to least probable, then assigns the top grade to the top 5%, say, the next grade to the next 10%, and so on down the ranking.

[7] We use letter grades for convenience because the order is known to all.

[8] We set aside philosophical qualms about the notion of an event's true risk.

[9] The median grade is just the middlemost one, when they're all put in order from top to bottom; the related notion of "splitting the difference," introduced in Sect. 2, plays no further role here. With an even number of panel members, a random selection is made from the two middlemost grades.

to select 10 events in the following manner. First, if there are not more than 10 events that receive the top grade then all of these events are selected. Then, if by selecting in addition all events that get the second-best grade not more than 10 events are selected, then all of these are selected as well. This is repeated until more than 10 events would be selected by selecting all events that got the next grade down. When that happens, it means that there are ties. We select at random from the tied events to make up the required 10.

We need some way of assessing the performances of our panel members, both as individuals and as a collective. Since the grading task at hand is to select the 10 most likely events, we measure individual performance by the percentage of 10 most likely events that are selected in the manner just described. The performance of the risk panel as a whole is measured similarly.

We have explained the basics of the model: the nature of the events, experts, and risk panel. We now turn to the main focus of our study: the different ways in which panel members can interpret the applicable grades, and the consequences for their individual and collective performance. We assume, now, that when the panel takes up its task, nothing whatsoever is known about the distribution of the probabilities of the events under consideration. For all anyone can tell there might be a lot of very probable events, or a lot of very improbable ones, or any mixture of these. In keeping with our assumption that grades are absolute, this has consequences for which interpretations can be recommended. Since the 10 most probable events might be at the upper end of the probability scale 0–100%, or at the lower end, or anywhere in between, the most suitable interpretations would appear to be ones that spread out the thresholds between grades over the whole scale. Accordingly three cases seem particularly relevant, which we call *Symmetric Consensus*, *Random Interpretations* and *Random Consensus*. We explain them in turn.

There is a *Symmetric Consensus* if all of the graders interpret the grades in *exactly* the same way and it is this: the grades map on to the 0–100% probability scale in a *symmetric* fashion. For example, if the available grades are $A$, $B$, $C$, $D$, and $E$ then, in case of Symmetric Consensus, each grader interprets $E$ as 0–20%, $D$ as 20 –40%, $C$ as 40–60%, $B$ as 60–80% and $A$ as 80–100%. Symmetric Consensus models the case in which members of the group are successfully trained to the point that they have one and the same rather special interpretation of the grading expressions, while abstracting away the details of the training method used. Symmetric Consensus might on occasion be achievable. In general, though, it must be regarded as, at best, a regulative ideal: something to aim for even if it cannot quite be achieved.

In case of *Random Interpretations* people's interpretations of the grades could be anything. In the simulation, each time the expert panel goes about the task of grading some events, an interpretation of the applicable grades is chosen at random for each panel member, independently for each one. With five grades for instance an interpretation is generated by selecting some four threshold points from the 0–100% scale. This allows panel members some pretty "crazy" interpretations. For example, one possible interpretation is: $E$: 0–2%, $D$: 2–4%, $C$: 4–55%, $B$: 55–57%, and $A$: 57–100%. Because the threshold points are generated according to uniform distribution, though, on average they will be spread out uniformly across the 0–100% scale. Random Interpretations models the case in which there could be lot of semantic diversity within

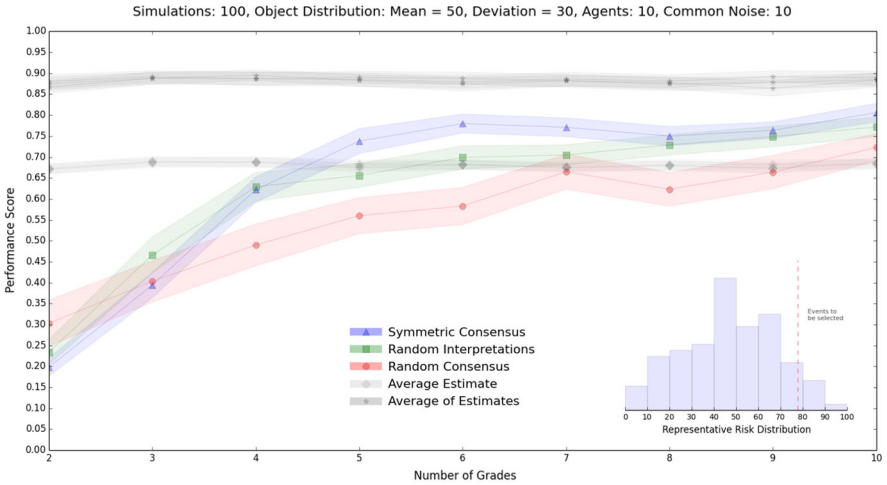


**Fig. 2** The basic model. Points with connecting lines indicate mean performances across simulation runs. Shaded regions depict 95% confidence intervals about the means

the group of experts while abstracting away from the origin of this diversity: whether the differing cultural backgrounds of members of the panel, their various ages and levels of experience, or what have you.

Finally, there is a *Random Consensus* when all panel members have the same interpretation of the grades but it could be anything: each time the panel goes about its task a common interpretation is chosen at random, as described above. Random Consensus models the case in which each member of the panel starts off with their own interpretation; then, realizing they might have different ones, and thinking that must be bad, they have a group discussion that results in their settling on a common interpretation. It is well known that such group discussions can be flawed in all sorts of ways (cf. Sunstein 2006). Presumably it won't help our panel's performance if one dominant member convinces the others to adopt a badly skewed interpretation that, making many distinctions at the bottom of the scale but few at the top, might not be so good for discriminating among the most likely events. Random Consensus models the results of such discussions while abstracting away the back and forth.

## 4 Simulation results

With all of these details in place, we can now simulate this grading task to see what happens in the three cases explained above. Note that we have not fixed the number of grades the risk experts use, since we want to see what happens when we vary this. Figure 2 shows performances of the different risk panel collectives with from 2 to 10 grades (blue for the Symmetric Consensus group, green for the Random Interpretation group, and red for the Random Consensus group).

There are also plots for the performance of the average of the graders' precise estimates (the dark grey line at the top of the figure), and the average performance of the graders' precise estimates (the light grey line underneath it). The performance of the

average is a benchmark because averaging precise estimates is the classic "wisdom of crowds" aggregation method (cf. Page 2008; Lyon forthcoming). Notice that it forms an upper bound to all the performance plots. This level of performance is an ideal that groups can aspire to, even if they can only achieve it when precise quantitative information is available. The average performance of the graders' precise estimates represents what you should expect if you chose an individual grader at random, somehow found out their precise estimates, and used these to make the selection.

Random Consensus is clearly the worst of our three cases. Symmetric Consensus and Random Interpretations both do systematically better and have roughly the same performance levels. Notice that with three grades it is best if the experts have interpretations chosen at random. With just two grades, both Random Interpretations and Random Consensus outperform Symmetric Consensus.

These observations are surprising. It has seemed intuitively obvious that interpersonal differences of interpretation must lower the quality of collective judgements, by creating equivocation and misunderstandings. The results also seem to be at odds with claims of Balinski and Laraki (2011) and of Morreau (2016) about the importance of common languages and shared interpretations. Here it seems we may expect almost identical performance from one group that's completely "on the same page" and another whose interpretations could be all over the place![10] Soon we will illustrate the robustness of the results, by varying relevant parameters including the size of the panel, the expertise of its members, and the distribution of event probabilities. First, though, let's try to understand why some of the results are as they are. We do not now have full explanations, but can suggest where these will be found.

Grading languages like the ones studied here, with from just 2 to 10 terms, don't make many distinctions. Their expressive resources are stretched in this task, with 100 events to distinguish between. Allowing different people to have different interpretations of the same terms, though, will tend to increase the total number of distinctions drawn by different members of the group. Somehow the aggregation procedure exploits these differences so as to arrive at a finer-grained ranking of the 100 items than is possible with a consensus interpretation.

The generally poor performance of Random Consensus in comparison with Symmetric Consensus is, we think, a result of differences in the suitability, for the panel's particular task, of different interpretations of grading languages. The symmetric interpretation, we suggest, is one of those that tends to give pretty good epistemic performance, even if it is not optimal. There are many other interpretations, some of which don't make distinctions where these are needed. The result in Fig. 2 suggests to us that the symmetric interpretation is better than most of these alternatives.

Why does Random Interpretations show better performance when there are three grades? And why does Random Consensus do relatively well when there are just two? Apparently, when the number of grades is very small in relation to the number of items to choose among, fixing a common interpretation once and for all severely depresses collective performance. Randomizing interpretations gives different mem-

[10] With random interpretations there is no interpersonal constraint on interpretations. The only real requirement is that each individual's interpretation is coherent, in the sense that higher grades go with higher probability intervals.

bers different interpretations "at the same time"—which is to say on any given run of the simulation or, intuitively speaking, on any given occasion on which the panel performs its task. Random Consensus forces a common interpretation at any given time, but the group settles on different ones on different occasions; sometimes, at least, the common interpretation will be one that makes a distinction where that is needed, among the most probable events. With just two grades, on the other hand, the symmetric interpretation makes no distinctions at all above 50%; since about half of the 100 events are more probable than that, the result is very poor performance from the Symmetric Consensus panel. Apparently, with a very small number of grades the differences of interpretation that come with Random Interpretations and Random Consensus are enough to bring performance above the very low level obtained with the Symmetric Consensus.[11]

Another interesting point we note, this time without explanation, is that the epistemic advantage in using more grades diminishes rapidly above about five grades.

Our findings are robust under a range changes of parameters in the simulation that correspond to differences among real panels of experts. One such change is to increase the noise in the panel members' estimations. Noisier estimations may be expected, for instance, from panel members with less individual expertise, or who must work more quickly or with poorer information. We model increased noise by increasing the standard deviation of the panel members' estimation process. The qualitatively similar results then obtained are illustrated in Fig. 3. As the experts get noisier, the performance of the average of their estimates decreases and this acts as an upper bound on the performances of the collective grades.

What happens if we change the size of the panel? Not much. If we decrease the number of graders the performances of the collective selections (including the ones based on the precise estimates) all decrease slightly. With more graders the exact opposite happens. This is to be expected: with a smaller group there's less scope for collective wisdom.[12]

So far we have been varying the basic parameters of the model (except for the event probability-distribution parameters). Before varying some structural assumptions we highlight an interesting feature of all results so far: it's hard to do much better than five grades under Random Interpretations. This can have important practical consequences. Our results suggest that if you want to set up a fairly good panel at minimal expense, use about five grades and let panel members understand them however they like (assuming you expect their interpretations to vary a lot). You won't need a very large panel, either: 5–10 members will do nicely.

These recommendations should, of course, be taken with care—if they are taken at all—since they rest on assumptions that will sometimes not be found realistic. For

[11] Luc Bovens (private communication) reports an analysis of binary grading that supports our observation that Random Consensus outperforms Symmetric Consensus in this case. His results suggest that things will be the same with a uniform distribution of the probabilities of the 100 events, instead of the Gaussian of Fig. 2, but that when many more events than 10 are to be selected the Symmetric Consensus will be better. Analyses of this sort are valuable. They help to explain the results and can suggest new hypotheses to test in further simulations.

[12] We omit graphs for these results to save space.

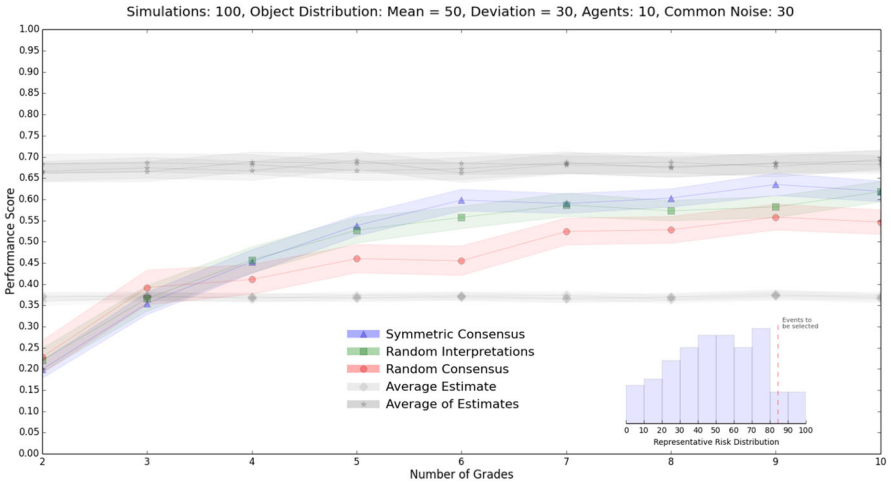


**Fig. 3** The basic model again, but the graders have less expertise

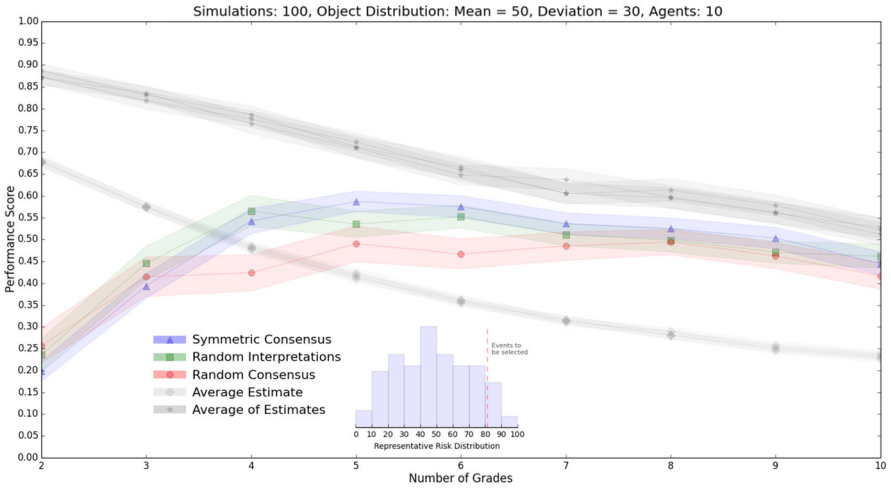


**Fig. 4** The basic model with a trade-off between the experts' level of expertise and the number of grades: the standard deviation of the experts' estimation process is $5n$, where $n$ is the number of grades

example, we assume that the number of available grades has no bearing on the quality of the experts' judgement. One might expect, though, that having to choose among a large number of grades places a cognitive burden on the experts, so let's see what happens when this decreases the accuracy of their estimates. Figure 4 concerns the case in which the noise of the experts is $5n$, where $n$ is the number of grades; so if there are just 2 grades to choose between the standard deviation of the estimates is 10%; with 10 grades it is 50%. Interestingly, the results are quite similar to those obtained with the basic model. The picture is much the same if the cognitive burden is assumed even greater (e.g., instead of $5n$ we use $10n$).

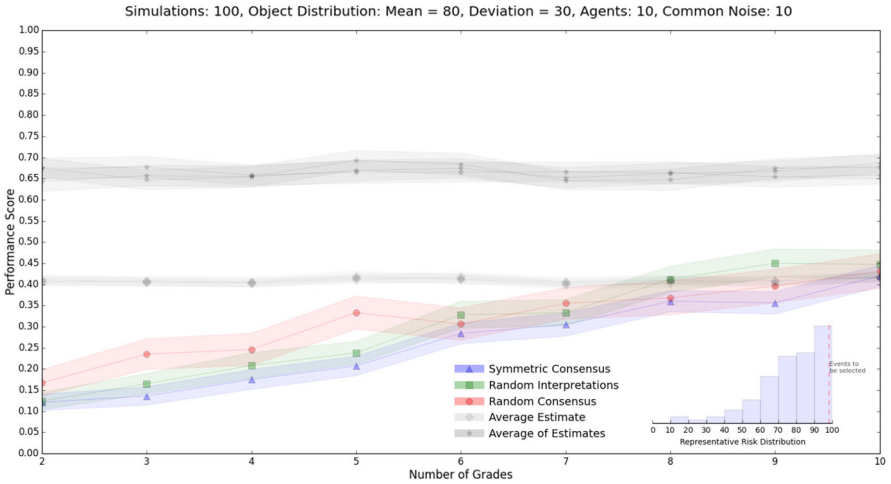


**Fig. 5** The basic model but with mean event probability set to 80%

The trade-off between the number of grades and estimate accuracy might sometimes go in the other direction instead. That could happen if the graders, knowing they can only use a small number of grades, put less effort into coming up with accurate precise estimates of the event probabilities. Conversely, if they know they are expected to use a lot of grades, then they might try harder. The results of these sorts of simulations are not all that interesting. To get anything remotely interesting, we need to assume that there is a lot of noise (i.e., std $\approx$ 50%) when the experts use two grades and that it approaches zero as we increase the number of grades. This has the qualitative effect of increasing the slope of all the curves above those of the base model.

An important assumption so far is that the distribution of event probabilities is Gaussian, with a mean of 50% and a standard deviation of 30%. What if most of the events are very likely? Changing the mean of the Gaussian to 80% (and keeping it truncated at 100%) we see similar results, except that now performance is not as good and improves comparatively slowly as the number of grades increases. Symmetric Consensus, Random Interpretations and Random Consensus perform similarly, and comparatively poorly, over the whole range of grades (see Fig. 5). With a more extreme change, shifting the mean to 90%, say, and shrinking the standard deviation down to 10% (from 30%), we get comparable results. Apparently, when most events are very likely the panel's choice procedure is overwhelmed. It is not capable of making enough distinctions to separate the most likely of all from the rest, whether panel members share a common interpretation or not, and especially when the number of available grades is small.

Moving things in the other direction, say by setting the mean to 20% (while keeping the standard deviation at 30%), the picture is much as it is with the original model. The results are more volatile, though, since the collective grading task is to choose the 10 most likely events and they are all in the long tail of the distribution.[13]

[13] Following up on one reviewer's suggestion we repeated the simulations using beta distributions instead of truncated Gaussians but found no discernible change in the results.

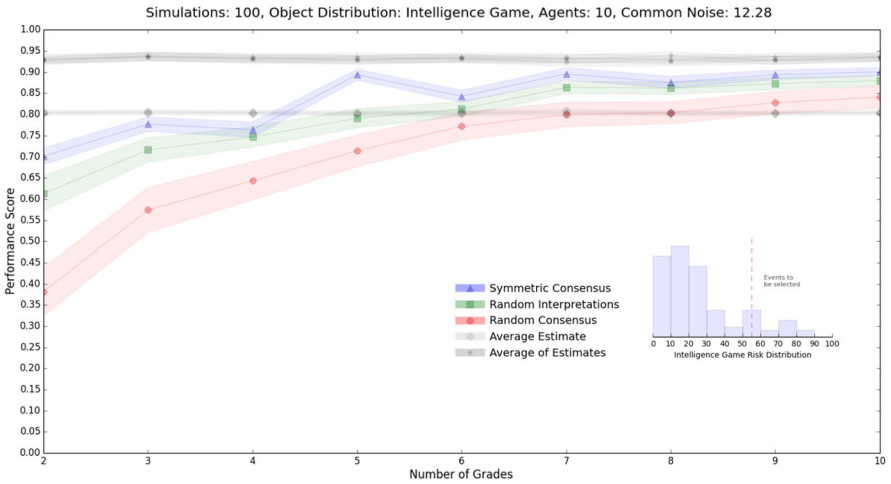


**Fig. 6** The basic model but with the distribution of events and expert noise derived from the *Intelligence Game* risk analysis experiment, (Wintle et al. 2012)

We also derived a real distribution of event risks from the *Intelligence Game* risk analysis experiment, Wintle et al. (2012). The Intelligence Game project involved hundreds of participants forecasting geopolitical events using subjective probabilities. To estimate a "true risk" for each event, we took the average of the probabilities assigned to it by the 100 best calibrated participants in the experiment.[14] By focusing on the best calibrated participants (measured across all of the events that were forecasted), it is hoped that we measured some stable probabilistic feature of the events in question. By looking at this distribution, we found it to be approximately a Pareto distribution, a distribution that commonly appears in nature (see Fig. 6). We were also able to calculate the average noise of the participants about these "true risks", assuming it to be Gaussian with a fixed variance: 12.28% probability. Using these as inputs to the simulation model, we found similar results (see Fig. 6), except that in this case the Symmetric Consensus always did as well as or better than the Random Interpretations (on average).

We've by no means exhausted modifications of the basic model. However, we think the main result of the simulation has become clear: under a wide range of conditions, an expert panel may be expected to perform almost as well when the different experts on the panel have different interpretations of the available grades, chosen at random, as when they all share one and the same symmetric interpretation. If on the other hand the experts share a common interpretation, but it is any old one, chosen at random, collective performance tends to be quite a bit worse.

There is an important lesson here. Training, discussion and other efforts to improve group performance by standardizing everybody's understanding of evaluative language are not always good. They can be counterproductive if the result is convergence

[14] Someone is well calibrated to the extent that their confidence in their predictions agrees with how many of these come true: of the predictions in which they're 90% confident about 90% should come true, and so on.

on an interpretation that is not itself especially good. That could happen, say, if the experts in the panel start off with a wide range of interpretations and, perhaps as a result of group discussion, the panel can settle on any one of those.

## 5 Summary

Committees and expert panels commonly rank items of various kinds on the basis of scores or grades contributed by individual members. People interpret scores and grades differently, though, and this might be expected to reduce the group's capacity to track the facts on which good decision making depends. This article investigates the consequences of diversity in grading thresholds for a group's capacity, in particular, to rank events by their probability and choose the most probable ones. It approaches this matter using the methodology of multi-agent computer modeling.

The main finding is that having members with different grading thresholds can boost the epistemic performance of a group. It does so under a range of realistic assumptions about the number of members, levels of their individual expertise, and distributions of the items under consideration. We note that we have tested our finding, though, only for the case in which grades are given absolute interpretations, and using grading languages that appear suitable for the case in which nothing is known in advance about the actual distribution of the events under consideration. While we have no special reason to expect that matters are different with relative grades, or with advance knowledge of distributions, further simulation work is needed to establish that diversity in thresholds boosts epistemic performance in these cases as well.

Our conclusions might be found surprising. Differing interpretations of scores and grades amount after all to a kind of equivocation, and that is something it seems better to avoid when working in groups. We do not offer a full explanation for our main finding but point in the direction in which we expect that one will be found. It is inherent in scoring and grading that relevantly different cases sometimes get lumped together. When a group can make any distinctions that its individual members do, though, a group whose members have different thresholds makes more distinctions than does a group with the same thresholds, and it is able to tell more different cases apart. This, we suggest, is how diversity of thresholds contributes to the performance of groups in ranking tasks.

Our results so far come from observing simulated groups only. Further empirical work is needed to see whether they hold up for real juries, committees and expert panels.

**Acknowledgements** We thank for helpful comments Luc Bovens, Wlodek Rabinowicz and an anonymous reviewer, as well as audiences at Bristol and York Universities, the London School of Economics and the Munich Center for Mathematical Philosophy.

## References

Alcantud, J. C. R., & Laruelle, A. (2014). Dis&approval voting: A characterization. *Social Choice and Welfare*, *43*(1), 1–10.

Aleskerov, F., Yakuba, V., & Yuzbashev, D. (2007). A 'threshold aggregation' of three-graded rankings. *Mathematical Social Sciences*, *53*(1), 106–110.

Arrow, K. J. (1951). *Social choice and individual values*. New York: Wiley (2nd ed. 1963).
Balinski, M., & Laraki, R. (2007). A theory of measuring, electing, and ranking. *Proceedings of the National Academy of Science United States of America*, *104*(21), 8720–8725.
Balinski, M., & Laraki, R. (2011). *Majority judgement*. Cambridge: MIT Press.
Beisbart, C., & Bovens, L. (2007). Welfarist evaluations of decision rules for boards of representatives. *Social Choice and Welfare*, *29*(4), 581–608.
Beisbart, C., Bovens, L., & Hartmann, S. (2005). A utilitarian assessment of alternative decision rules in the council of ministers. *European Union Politics*, *6*(4), 395–419.
Beisbart, C., & Hartmann, S. (2010). Welfarist evaluations of decision rules under interstate utility dependencies. *Social Choice and Welfare*, *34*(2), 315–344.
Brams, S. J., & Fishburn, P. C. (1978). Approval voting. *American Political Science Review*, *72*(3), 831–847.
Condorcet, J.-A.-N. d. C. (1785). *Essai sur l'application de l'analyse a la probabilite des decisions rendues a la pluralite des voix [microform]/par M. le Marquis de Condorcet*. Imprimerie royale Paris.
Gaertner, W., & Xu, Y. (2012). A general scoring rule. *Mathematical Social Sciences*, *63*(3), 193–196.
Galton, F. (1907). Vox Populi. *Nature*, *75*, 450–1.
Hubbard, D., & Evans, D. (2010). Problems with scoring methods and ordinal scales in risk assessment. *IBM Journal of Research and Development*, *54*, 246–255.
List, C. (2013). Social choice theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2013 ed.).
Lyon, A. (forthcoming). Collective wisdom. *Journal of Philosophy*.
Lyon, A., & Pacuit, E. (2013). The Wisdom of crowds: Methods of human judgement aggregation. In: Michelucci P (Ed.), *Springer handbook of human computation*. (pp. 599–614). Springer.
Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., et al. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115.
Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Science United States of America*, *111*(20), 7176–7184.
Morreau, M. (2016). Grading in groups. *Economics and Philosophy*, *32*(2), 323–352.
Nielsen, M. (2011). *Reinventing discovery: The new era of networked science*. Princeton: Princeton University Press.
Ohnishi, M., Fukui, T., Matsui, K., Hira, K., Shinozuka, M., Ezaki, H., et al. (2002). Interpretation of and preference for probability expressions among Japanese patients and physicians. *Family Practice*, *19*(1), 7–11.
Page, S. (2008). *The difference*. Princeton: Princeton University Press.
Pivato, M. (2016). Asymptotic utilitarianism in scoring rules. *Social Choice and Welfare*, *47*(2), 431–458.
Sunstein, C. R. (2006). Deliberating groups versus prediction markets (or hayek's challenge to habermas). *Episteme*, *3*(03), 192–213.
Surowiecki, J. (2004). *The Wisdom of Crowds*. New York: Doubleday.
Tetlock, P. (2005). *Expert political judgment: How good is it? How can we know?*. Princeton: Princeton University Press.
Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, *155*(4), 348–365.
Wardekker, J. A., van der Sluijs, J. P., Janssen, P. H. M., Kloprogge, P., & Petersen, A. C. (2008). Uncertainty communication in environmental assessments: views from the Dutch science-policy interface. *Environmental Science and Policy*, *11*, 627–641.
Wintle, B., Mascaro, S., Fidler, F., McBride, M., Burgman, M., Flander, L., Saw, G., Twardy, C., Lyon, A., & Manning, B. (2012). The intelligence game: Assessing Delphi Groups and structured question formats. In *Proceedings of the 5th Australian Security and Intelligence Conference*. http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1025&context=asi.
Wolf, M., Krause, J., Carney, P. A., Bogart, A., & Kurvers, R. H. (2015). Collective intelligence meets medical decision-making: The collective outperforms the best radiologist. *PloS One*, *10*(8), e0134269.