Managing and Aggregating Group Evidence under Quality and Quantity Trade-offs

Zoi Terzopoulou, Patricia Mirabile, Pien Spekreijse

Managing and Aggregating Group Evidence under Quality and Quantity Trade-offs

Zoi Terzopoulou[1], Patricia Mirabile[2], and Pien Spekreijse[3]

Abstract

Trade-offs between quality and quantity arise in an abundance of contexts concerning group decision making. With the starting point being that group members provide more accurate evidence when they are involved with fewer tasks, team managers often encounter the following dilemma: Should they assign their group members with many tasks (attempting to gather more evidence with lower quality), or with fewer tasks (aiming at receiving less, but more high-quality evidence)? Secondly, what is the optimal way to aggregate the collected evidence from a group, which may be contrasting and varying in accuracy? Should more weight be given to the more accurate group members, or to the larger number of those who provide the same answer? This topic is already studied within the mathematical framework of Terzopoulou and Endriss (2019). In this paper we complement it experimentally, by investigating to what extent people's decision-making patterns are in accordance with the optimal ones proposed by the normative model. Our findings suggest that people understand the task at hand and generally opt for optimal choices, especially in conflict-free cases. Still, a tendency towards overvaluing the importance of additional evidence, despite their accuracy, is observed; this translates into choosing options that align with the majority rule in aggregation problems.

Keywords

social choice, evidence quality, evidence quantity, information aggregation, online experiment

[1] GATE, Saint-Etienne School of Economics, Jean Monnet University, France
[2] ILLC, University of Amsterdam, The Netherlands
[3] New10, Amsterdam, The Netherlands
Email: zoi.terzopoulou@cnrs.fr

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

1 Introduction

Consider a large journal where the lead editor is in charge of assigning incoming articles to reviewers. A non-straightforward decision problem—which lies at the heart of this paper—has to be resolved by the editor. On the one hand, assigning plenty of the available reviewers to incoming articles and collecting all of their judgments will provide more information to her about the articles' validity and suitability for publication; on the other hand, asking reviewers to evaluate several articles at the same time will arguably result in less precise judgments on their side, due to multitasking efforts, split-attention effects, and restricted time spent on each single article. A question thus arises: What kind of article assignment increases the chances of an appropriate evaluation to be made by the reviewers and of a good decision to be taken by the editor based on the collected information? Answering this question requires confronting a conflict between the *quality* and the *quantity* of the evidence that must be first gathered and then aggregated.

Analogous trade-offs of quality and quantity appear in numerous social contexts that involve decision making by groups, including the assignment of cases to judges in courts and patients to doctors in hospitals. Given a limited budget, designers of crowd-sourcing experiments also have to choose between consulting fewer, expensive experts, or more, cheaper non-experts. More broadly, in universities, private institutions, and companies, team leaders encounter the option to create either larger working groups with multiple employees that interact on a number of simultaneous tasks, or smaller groups with individuals that are responsible for separate tasks.

Moreover, after assigning tasks to the members of a group, team leaders regularly collect the obtained evidence and need to aggregate it in order to make the best possible collective decision. A trade-off between quality and quantity of the available information is brought to the surface again: In light of contrasting evidence, should priority be granted to the judgments reported by the most accurate group members, or to those held by a larger number of them?

A pool of articles suggests that multitasking, time-pressure, and hasty reasoning have negative effects on performance (Payne et al. 1988; Edland and Svenson 1993; Ariely and Zakay 2001; Wilhelm and Schulze 2002; Adler and Benbunan-Fich 2012). However, little has been said about the way in which team leaders take this fact into account in their managerial decisions. Notably, behavioural evidence concerning decisions made by real people is generally overlooked within the academic literature on group decision making.

A mathematical framework that answers how trade-offs of quality and quantity can be optimally resolved has been previously developed by Terzopoulou and Endriss (2019). Here, we are interested in testing experimentally the extent to which the decisions made by real people correspond to the optimal ones proposed by the normative model. Knowing whether people make good decisions when faced with such trade-offs is of high importance for applications: Managers in positions of power rarely have access to all relevant information that is needed in order to be faithful to the mathematics (for example, about the accuracy of their group members); rather, they rely on intuitions and heuristics. Our work investigates how often these intuitions give rise to correct choices in extreme cases of full information. Learning that people are good in resolving these

trade-offs can create a safety net for more complicated instances that directly rely on human judgment; on the other hand, understanding on where possible errors in resolving the trade-offs concentrate can help set up appropriate training to forge improvement.

The topic of this paper is relevant for the field of *social choice* (Arrow et al. 2002)—including the more recent one of *computational social choice* (Brandt et al. 2016)—that is concerned with the formal analysis of methods that groups (should) use to make decisions as a whole. *Epistemic* social choice specifically investigates the optimal way to aggregate the judgments of different group members in order to discover a *ground truth*, that is, an objective answer to a complex question (such as the treatment to a rare illness). The epistemic approach to social choice was instigated by the famous Condorcet Jury Theorem (de Condorcet 1785), intuitively stating that if we want to learn the answer to a single yes/no question with high probability, it is better to ask as many group members as possible, given that they are more accurate than random guessers. Yet, in his original work, Condorcet did not consider the problems arising when more than one question have to be tackled in parallel. Since then, social choice researchers have explored various related topics, such as discovering a correct ranking for a number of alternatives in a scenario of preference aggregation (Caragiannis and Micha 2017), or learning the correct answer to logically interdependent propositions in judgement aggregation (Hartmann et al. 2010).

The mathematical, philosophical, and computational tools of social choice have been developing mostly independently of empirical and psychological evidence; this can be contrasted with related subdomains of economics that are concerned with individual—as opposed to group—decision making. For instance, in behavioural economics and psychological decision theory it is widely accepted that *biases* and framing effects cause human decision-makers to make sub-optimal decisions (Tversky and Kahneman 1985; Kahneman 2003). Few exceptions exist—some social choice researchers have been interested in examining the descriptive power of established normative models; for example, people's insincere behaviour in voting (Bassi 2015), the actual behaviour of the participants in doodle polls (Zou et al. 2015), and the way in which people strategise when a popular aggregation method (the plurality rule) is used in iterative settings (Meir et al. 2020) have already been subject to investigation. In this paper we support this line of work, testing a normative model of social choice in practice.

Our setting also connects to the *information-acquisition* literature at large, where optimal decisions must be made given uncertain information and limited resources to access that information. For example, consider the following problem taking place in a single round (Azevedo et al. 2020): The manager of a firm has a set of ideas and a number of employees available to test them. The quality of each idea is uncertain, drawn from a prior distribution. To learn about the value of an idea, the manager can run an experiment on a subset of her employees that will produce a noisy signal of the quality of the idea. The question is how to assign the total budget of available employees to each idea and then select which ideas to implement. Although similar in flavour, our setting is different in that the desired information is interconnected; our manager's decision will depend on two given criteria that must be evaluated in equal terms, and then an aggregation task will take place. There is little work on problems where a specific structure of the

information trade-off is posited. Pertinent research domains are also those on *value of information* and *optimal-learning*, dealing with the general problem of sequential choice amongst several actions, where at each stage a decision-maker may stop and take a definite action or continue sampling for more information. There are costs attached to taking inappropriate action and to sampling, and the expected cost against the expected amount of information to be obtained must be balanced (Arrow et al. 1949; Moscarini and Smith 2001). Touching this topic, *bandit* problems are the most basic examples of sequential decision problems with an exploration–exploitation trade-off, i.e., the balance between staying with the option that gave highest payoffs in the past and exploring new options that might give higher payoffs in the future (Thompson 1933)—see Bubeck et al. (2012) for an overview of this stream of literature, which has been developing rapidly, especially in computer science. Our work departs from those problems by focusing on decisions that take place in a single round rather than sequentially, and assuming that the precision of the information signal may change depending on how the manager decides to assign tasks.

For the remainder, let us abstract away from the numerous parameters that play a role in scenarios where quality is in conflict with quantity, and employ a toy example to focus on the aspects that are more pertinent in our investigation. This toy example will be also used in our experiment, to familiarise the participants with our setting and our questions. It goes as follows: The manager of a pie factory needs to decide whether the produced pies satisfy the necessary standards for them to be acceptable for sale. Specifically, the factory imposes two relevant evaluation criteria for each pie: it has to look good enough (that is, the *visual* criterion), and its price has to be calculated correctly (that is, the *numerical* criterion). The manager assigns different tasks to the factory workers, by asking them to check the visual and the numerical criteria for a number of pies—after the workers have expressed their judgments on their assigned tasks, the manager makes the final decision about whether each pie can be for sale, by aggregating the information she received. There are two clear phases in this process: the first is about *task assignment*, and the second is about *aggregation*. In both these types of problems, our goal is to examine whether people's actual decisions agree with the decisions suggested by the relevant mathematical model. Within the context of our experiment, we find evidence that people understand the problem presented to them and generally opt for maximal accuracy. Briefly, as far as task assignment is concerned, alignment between people's choices and the normatively optimal ones is common. In aggregation problems, an agreement between people's choices and the recommendations of the standard majority rule prevails. The latter observation brings out an additional possible interpretation, which is partially observed in task assignment too: namely, that people's choices sometimes over-rely on the importance of lower accuracy judgments.

This paper is organised as follows. In Section 2 we briefly summarise the existing formal model that tackles conflicts between quality and quantity from the perspective of social choice theory (Terzopoulou and Endriss 2019); we recall its main results and present the implied predictions regarding human behaviour. In Section 3 we demonstrate our experiment that aims to test the aforementioned model, studying how far or not people's behaviour is from the theoretically predicted one, and we elaborate on our

observations. In Section 4 we offer a general discussion and conclude.

2 Formal model and predictions

In this section we scrutinise the mathematical framework of Terzopoulou and Endriss (2019), developed to study the trade-offs between quality and quantity when managing and aggregating evidence from groups.

The formal model aims at answering epistemic questions: Consider a group of individuals that need to collectively determine the answer to a binary question (e.g., "is a pie ready to be sold?") that directly depends on the evaluation of several independent criteria (e.g., "does the pie look good enough?" and "is the price of the pie calculated properly?"). A correct *yes*/*no* (or *approve*/*reject*) answer on the different criteria exists, and each individual in the group can provide independent evidence about what that answer may be, by expressing a judgment that has a certain probability of being correct. But, most importantly, different individuals may be asked to assess different criteria. A main assumption of the model is that the more criteria an individual tries to assess, the less accurate her judgments are likely to be. How can the group then maximise the probability of discovering the correct answer to the question they are facing?

The model's contribution is twofold: It determines what the optimal rule to aggregate the evidence provided by such a group of individuals is, and finds the optimal way to manage those groups in terms of assigning the right amount of tasks to them.

2.1 Mathematical preliminaries

Let φ and ψ be two independent criteria associated with a correct *yes*/*no* answer. *A priori*, each of the two answers is equally likely to be the correct one. Every individual i in a group $N = \{1, \ldots, n\}$ with $n \geqslant 2$ holds a personal judgment $J_i \subseteq \{\varphi, \overline{\varphi}, \psi, \overline{\psi}\}$. With $\varphi \in J_i$ we mean that individual i judges φ as true, and with $\overline{\varphi} \in J_i$ that individual i judges φ as false. We write $J^{\blacktriangle} \subseteq \{\varphi, \overline{\varphi}, \psi, \overline{\psi}\}$ for the judgment that captures the correct evaluation on the two criteria.

An *aggregation rule* F is a function that maps every reported profile $\boldsymbol{J} = (J_1, \ldots, J_n)$ of all individuals' judgments to collective judgment $F(\boldsymbol{J})$. Intuitively, an aggregation rule is responsible to determine the final answer to our criteria, given the evidence provided by the individuals in the group.

We define N_1^{φ} to be the set of individuals who report a judgment on *one* criterion, φ, and say "yes" to it (analogously for "no", we replace φ with $\overline{\varphi}$). By N_2^{φ} we denote the set of individuals who report a judgment on *both* criteria, φ and ψ, and say "yes" to φ (and analogously for "no"). We also define $n_1^{\varphi} = |N_1^{\varphi}|$ and $n_2^{\varphi} = |N_2^{\varphi}|$ to be the respective sizes of these sets.

We denote by p the probability that individual i's judgment J_i is correct on a criterion when i judges both criteria, and by q the relevant probability when i only judges a single criterion (assuming that the probability of each individual's judgment being correct on a criterion φ is independent (i) of whether φ is true or false and (ii) of what i's judgment on criterion ψ is). We assume that the probabilities p and q are the same for all individuals, but the individuals make their judgments independently of each other. We

shall moreover suppose that all individuals make judgments that are more accurate than a random guess, but not perfect, and that those judging a single criterion are strictly more accurate than those judging both criteria, i.e., that $1/2 < p < q < 1$. Then, $P(\boldsymbol{J})$ denotes the probability of the concrete profile $\boldsymbol{J}$ of judgments to be reported by the individuals.

The *accuracy* $P_\varphi(F)$ of an aggregation rule F regarding the criterion φ is defined as:

$$P_\varphi(F) \quad = \sum_{\substack{\boldsymbol{J} \text{ s.t.} \\ F(\boldsymbol{J}) \text{ and } J^\blacktriangle \text{ agree on } \varphi}} P(\boldsymbol{J})$$

Recall that one of the model's aims is to find the correct answer on the criterion φ with the highest probability of this answer being correct, by aggregating the evidence provided by the individuals in the group— formally, this translates to finding the rule F that maximises the probability $P_\varphi(F)$.

Yet, in order for the individuals in the group to provide evidence regarding the two criteria, the appropriate tasks need to be assigned to them. That is, we need to know which individuals will assess which criterion. Different choices for assigning individuals to criteria yield a correct collective answer with different probability. We are interested in finding the optimal (viz., the most accurate) such assignment.

Let us denote by $n_1 \leqslant \lfloor \frac{n}{2} \rfloor$ the number of individuals that will be asked to report a judgment only on criterion φ. For symmetry reasons, we assume that the same number of agents will be asked to report a judgment only on criterion ψ, and the remaining $n - 2n_1$ agents will be asked to report a judgment on both criteria. Then, $P_{\varphi,n_1}(F^{OTT})$ is the probability of the aggregation rule F^{OTT} producing a correct answer on φ. This is the probability we aim at maximising, by finding the number $\operatorname{argmax}_{0 \leqslant n_1 \leqslant \lfloor \frac{n}{2} \rfloor} P_{\varphi,n_1}(F^{OTT})$. In simple words, we must know how many individuals will be assigned with a single task, and how many will be assigned with two tasks.

2.2 Results

For the proofs of the results stated in this section, the reader is referred to original paper where they are presented (Terzopoulou and Endriss 2019). We only give the information that is needed to evaluate our experimental observations.

It has been shown that the optimal truth-tracking (OTT) rule (i.e., the aggregation rule F that maximises the probability $P_\varphi(F)$) is a weighted-majority rule, assigning to the individuals weights according to how many criteria they assessed—individuals that assessed only one criterion, and are thus providing more accurate judgments, are assigned more weight in the following manner:

$$F^{OTT}(\boldsymbol{J}) = \operatorname*{argmax}_J \sum_{i \in N} w_i \cdot |J \cap J_i|,$$

where $w_i = \log \frac{q}{1-q}$ if $i \in N_1^\varphi$ and $w_i = \log \frac{p}{1-p}$ if $i \in N_2^\varphi$. Observe that the base of the logarithm in the definition of w_i is irrelevant.

Consider for example a group of three individuals with $q = 0.80$ and $p = 0.60$ providing the following judgments:

	φ	ψ
individual 1:	Yes	–
individual 2:	No	Yes
individual 3:	No	Yes

The weight assigned to individual 1, who has a positive judgment on φ and judges only that criterion, equals to $\log 0.80/0.20 = 0.60$. On the other hand, the weight assigned to individuals 2 and 3, who have a negative judgment on φ while they judge both criteria, equals to $\log 0.60/0.40 = 0.17$. Since $0.60 > 0.17 + 0.17 = 0.34$, the OTT rule produces a collective *yes* answer on φ. For ψ, the unanimous answer will also be *yes*.

The same work (Terzopoulou and Endriss 2019) also studied the optimal assignment method to distribute tasks to the individuals in the group, for small groups of size 2, 3, and 4 (for large groups, the mathematical analysis becomes too complex for formal results to be given, but numerical estimations can be provided). Of course, the optimal assignment depends on the specific values p and q of the individual accuracy. Intuitively, if q is much larger than p, then it will be better to assign single tasks to many individuals, because you can expect very high-quality judgments; if q is close to p and you cannot rely on quality, then it may be smarter to increase the quantity of the evidence you get by asking more individuals to assess both tasks. Drawing the line between the quality side and the quantity side is our desideratum.

For groups of only two individuals, it is always optimal to ask each one of them to evaluate one of the two criteria ($n_1 = 1$) rather than asking both individuals to evaluate both criteria ($n_1 = 0$). This shows that for tiny groups quality undeniably beats quantity. This is not an obvious discovery, since it holds even for values of p that are extremely close to q (i.e., within a group of individuals that are almost perfect multitaskers, it is still optimal to *not* have them multitask). We will indeed see later on, in Figure 4, that this is a case where people's actual behaviour largely differs than what the optimal model recommends. Intuitively, having two evaluations does not help because when they agree to the correct answer it will be with probability $p^2 < q$, and when they disagree they will produce a tie that offers no information.

For groups of three individuals, we know that $\operatorname{argmax}_{n_1} P_{\varphi,n_1}(F^{OTT}) = 1$ (that is, the best choice is to assign one group member only with criterion φ, another group member only with criterion ψ, and the third group member with both criteria) if and only if $q \geqslant p^2(3 - 2p)$. For example, if individuals who evaluate both criteria are correct 60% of the time, then you should ask two of them to focus on a single criterion each if and only if their accuracy for doing so is at least 64.8%.

Finally, for groups of four individuals, we have that $\operatorname{argmax}_{n_1} P_{\varphi,n_1}(F^{OTT}) = 1$ if $q < \frac{p^2}{(1-p)^2+p^2}$ and $\operatorname{argmax}_{n_1} P_{\varphi,n_1}(F^{OTT}) = 2$ otherwise. This means that we should either ask two individuals to focus on a single criterion each and two to multitask (if the group consists of good multitaskers), or to split the group in two teams, each assigned with a different criterion (if their multitasking accuracy is low). Notably, we should never ask all individuals to multitask.

2.3 Predictions

Based on the formal model, we derive two types of empirical predictions: First, we wish to check whether people make decisions that *precisely* agree with those that are prescribed by the normative model. For example, we saw in Section 2.2 that individuals in teams of two should never be asked to multitask—do people assign tasks in accordance to that fact? Because of the model's complexity, we do not expect this to hold in general, but at least in the simpler cases when the trade-off between quantity and quality is small. Second, we expect to confirm that people's decisions are influenced by the "right" parameters, that is, the parameters that the normative model suggests. For example, if the performance rate of the multitaskers increases, then the probability that people choose to make their team members multitask should increase as well.

Our empirical predictions concern the two core problems addressed by the theory, (i) task assignment and (ii) aggregation, where the following hypotheses are relevant.

H1. People are always more likely to choose the normatively correct option.
H2. People are more likely to choose the normatively correct option as the probability of that option being correct increases.

Besides H1 and H2, we also examine simple heuristics that people may adopt specifically in the context of task assignment. In particular, we investigate whether they will tend to prefer options that favour monotasking (H3), or options that favour multitasking (H4).

H3. People are more likely to choose the option that assigns the maximum number of team members to single tasks.
H4. People are more likely to choose the option that makes the maximum number of team members multitask.

In addition, we examine simple heuristics that people may adopt in the context of aggregation. In particular, we investigate whether they will tend to prefer options options that follow the opinion of the majority of team members (H5), or options that are in line with the information provided by the most accurate team members (H6).

H5. People are more likely to choose the option that agrees with the majority of team members.
H6. People are more likely to choose the option that agrees with the most accurate team members.

3 Experiment

In our experiment, participants performed the duties of the fictional manager from the *pie factory* toy example described in the Introduction. They were told to imagine the following story: their (i.e., the managers') ultimate responsibility is to approve or reject pies on the basis of two different quality criteria (a visual and a numerical criterion, which are described in more detail below), but without performing the pie quality checks themselves. Instead, the manager's role is to manage teams of workers who perform

the pie quality check tasks and then to aggregate those workers' judgments in order to make an *approve* or *reject* decision (the participants were told nothing about the *a priori* probabilities of a pie being acceptable or not, but there is no reason to believe that they assigned some unequal prior to them). Every team of workers is supposed to be responsible for one pie, and must collectively decide whether the pie satisfies each one of the two quality criteria. Given a certain pie, every worker in the team must perform at least one of the pie quality check tasks, i.e., must indicate whether the pie meets the relevant quality criterion. Workers have a specified accuracy track record when performing the two pie quality check tasks at the same time, that is when *multitasking*, and when performing only one of these tasks, that is, when *monotasking*. To be in agreement with the assumptions of the theoretical model, the accuracy of the evidence when each group members multitask (p) or monotask (q) was shown to be the same, with q always strictly larger than p.

Managing a team means deciding whether team members should be asked to only complete one of the tasks (to monotask), or instead to complete both tasks (to multitask). In other words, the manager is met with the alternative of having her fictional workers complete more tasks, and therefore collecting more judgments from workers who are less accurate, or having workers complete fewer tasks and therefore collecting fewer judgments but from workers who are more accurate.

Participants in this experimental setting completed two separate phases of the evidence management and aggregation process: the *task assignment* phase and the *aggregation* phase. This allowed us to examine how the evidence management and the aggregation decisions of the participants were impacted by the consideration of quality versus quantity trade-offs. For each phase of the task assignment and aggregation process, we sought to answer two questions. First, does the OTT model provide a better *qualitative* description of people's decisions when compared with alternative heuristics (corresponding to H1 in Section 2.3)? Second, how well can people's decisions be modelled by the OTT model's *quantitative* predictions regarding the optimal choice (corresponding to H2 in Section 2.3)?

3.1 Method

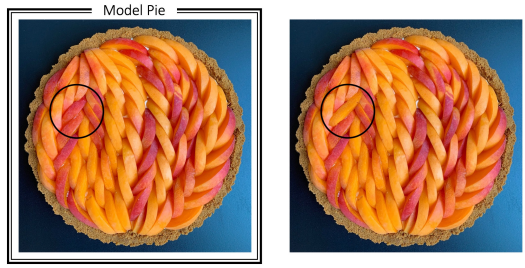
The planned sample size, predictions, statistical models, and priors were preregistered through the Open Science Platform. Materials, experimental scripts, analyses and data are freely available at `https://osf.io/cpqfz/`. At the time when this experiment was designed, the independent ethics committee of the University of Amsterdam was informed, and oral approval was granted. No official statement was necessary due to the non-sensitive nature of this work. The experiment was conducted on April 13, 2021, and no access to information that could identify individual participants during or after data collection was available.

Participants. We conducted a simulation-based power analysis to determine an appropriate sample size for the models in our planned investigations. This is a two-step procedure. First, we simulated data sets according to the data generation process implied by the OTT model and assuming a weak correlation of 0.45 log odds (or 0.61

in probability units) between the probability of a response being correct according to the OTT model and the probability that a participant would choose that response. Second, we confirmed that our statistical models were able to consistently detect the effects of interest for a large number of simulated data sets. We found that a sample of 600 participants would be sufficient to reliably detect those effects in 100 percent of cases and to allow for good convergence of the models. After accounting for an expected drop out rate of 15 percent, we recruited 690 participants via the Prolific.co platform. The data from 91 participants was lost due to experimenter error and 13 participants were excluded for failing two attention checks or for providing duplicate response sets. This left us with 586 participants (229 women, 314 men, 43 other; M_{age} = 23.0, SD_{age} = 8.6) who received a flat payment of 0.86£ as compensation for their participation. This payment—considering that the average time for completing the experiment was around 10 minutes—corresponds to around 5£ per hour, the minimum wage at the UK the year that the experiment took place. Note that providing a flat payment is the standard method in psychology (Croson 2005; Hertwig and Ortmann 2001), in contrast with other disciplines such as experimental economics. Indeed, such experiments are very common (it has been reported that that in a sample of 106 empirical studies published in psychology journals, fewer than three percent provided performance-based incentives (Hertwig and Ortmann 2003)) and present several advantages (Tversky and Kahneman 1989; Voslinsky and Azar 2021).

Procedure. The experiment was run fully online and consisted of three main parts: an *instructions* part, a *task assignment* part and an *aggregation* part.

Instructions part. After providing informed consent, participants were introduced to the experimental setting we have described above. To help make the setting more concrete, participants were given the opportunity to try out the pie quality check tasks themselves, both in a monotasking and in a multitasking capacity. Specifically, participants completed six trials of the pie quality check tasks. To verify that the pie satisfied the visual criterion, participants completed the visual check task, in which they had to compare two images, one of a model pie and one of the pie to be checked, and decide whether they were identical or not (we used lightly edited pie pictures from the `lokokitchen.com` website for this task): See Figure 1(a) for an example. To verify whether the pie satisfied the numerical criterion, participants completed the numerical check task, in which they had to confirm that the result of an addition with large decimal numbers was correct: see Figure 1(b) for an example. After completing those six trials, participants were told that their (purported) accuracy had been 75 percent when monotasking and 50 percent when multitasking. This served to underline that the tasks were sufficiently challenging, that workers with similar accuracy whom the participants would be asked to manage in the following parts of the experiments were competent, and that multitasking realistically decreases someone's accuracy (this type of deception, frequent in experimental psychology, was not expected to—and indeed did not, in any provable way—affect the subsequent responses of the participants). Participants were informed that this step was purely meant to familiarise them with the setting, and that it did not play a role for their decisions to follow.



(a) The visual check task.

1 kg Flower	$ 39.72
1 kg Fruit	$ 172.87
5 kg Sugar	$ 40.61
	——————— +
Total	$ 300.10

(b) The numerical check task.

Figure 1. Pie quality checks.

Task assignment part. Next, participants completed the task assignment part, which consisted of one practice trial and ten task assignment trials. For an example of a task assignment trial, see Figure 2. To be able to test the theoretical model, for each trial we only provided a certain number of relevant choices. For example, with a team of size three, there was the option to have 1 person multitasking and 2 people monotasking, as well as the option to have all three multitasking. There was no option to have 2 people multitasking and 1 monotasking because then one of the two criteria would receive a higher amount of evidence than the other, which is not accounted for in theory.

In each trial, participants were presented with a new team of workers and were informed of the workers' accuracy when performing the pie quality check tasks both in a monotasking (q) and a multitasking (p) capacity, with $q > p$, and with all workers in a given team having the same monotasking and multitasking accuracies. Depending on the team size, participants had either two or three task assignment configurations to choose from in a forced-choice paradigm (see Table 1 for a visual representation of those options). For teams of size two, participants could either choose to have both workers monotask (Option 1), and therefore collect one high accuracy judgment for each pie quality check, or to have both workers multitask (Option 2), and therefore collect two lower accuracy judgments for each pie quality check. For teams of size three, participants could either choose to have two workers monotask and one worker multitask (Option 1), that is, to collect two judgments per pie quality check with half of those being high

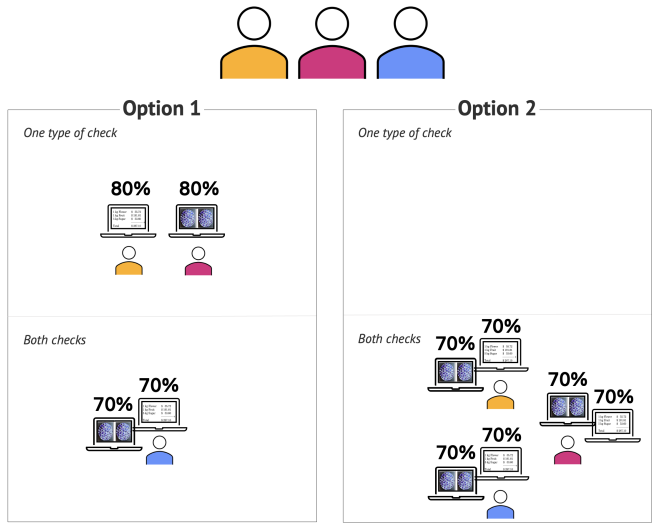


Figure 2. Task assignment trial for a team of size 3, with $q = 0.80$ and $p = 0.70$.

accuracy judgments, or to have all three workers multitask (Option 2), that is to collect three lower accuracy judgments per pie quality check. Finally, for teams of size four, participants had three options: have all workers monotask (Option 1) and collect two high accuracy judgments for each pie quality check; have two workers monotask and two workers multitask (Option 2), therefore collecting again two judgments per pie quality check but with half of those judgments having lower accuracy; have all workers multitask (Option 3) and collect four lower accuracy judgments for each pie quality check. Note that choosing Option 1 meant collecting fewer overall judgments but maximizing the number of high accuracy judgments, while Options 2 and 3 decreased the number of high accuracy judgments while increasing the total number of collected judgments.

Team size, monotasking and multitasking accuracies for each trial are reported in Table 1. The specific values for p and q were selected so that the different hypotheses to be tested could be distinguishable (that is, the corresponding assignment methods would give sufficient contrasting recommendations). In addition, we aimed at choosing accuracy values that induce different optimal responses for groups of varying size. We know from the theoretical results (last paragraph of Section 2.2) that if for given accuracy the monotasking option is optimal for groups of size 4, then the monotasking option will also be optimal for groups of size 3. Thus, we focused on accuracy values that propose monotasking in groups of size 3 but not necessarily in groups of size 4; we selected two scenarios that showcase markedly different accuracy values—both in terms of their absolute figures and the relative discrepancies within each case: these are the trials 80/70 and 95/80. Additionally, trial 85/80 was considered for groups of size 3 to observe the participants' reaction compared to the 95/80 case (where only the value of q was altered). Given the unique nature of groups of size 2, where monotasking always emerges as the

Table 1. Team size, task accuracies and assignment configuration options per trial for the task assignment phase.

Trial	Team size	q	p	Option 1	Option 2	Option 3
2-95/51	2	.95	.51	V & N	VN & VN	—
2-80/70	2	.80	.70	V & N	VN & VN	—
2-95/94	2	.95	.94	V & N	VN & VN	—
3-80/70	3	.80	.70	V, N & VN	VN, VN & VN	—
3-85/80	3	.85	.80	V, N & VN	VN, VN & VN	—
3-95/80	3	.95	.80	V, N & VN	VN, VN & VN	—
3-95/51	3	.95	.51	V, N & VN	VN, VN & VN	—
4-80/70	4	.80	.70	V, V, N & N	V, N, VN & VN	VN $\times$4
4-85/80	4	.85	.80	V, V, N & N	V, N, VN & VN	VN $\times$4
4-95/80	4	.95	.80	V, V, N & N	V, N, VN & VN	VN $\times$4

Note: q corresponds to the workers' accuracy when monotasking and p to the workers' accuracy when multitasking. All team workers in a given team had the same monotasking and multitasking accuracies. In the configuration options, V and N indicate that a worker would be assigned to monotask (on the visual or numerical task, respectively) and VN indicates that a worker would be assigned to multitask.

optimal option, we incorporated only one of the aforementioned "conflicting" scenarios for them—the 80/70 trial. We then introduced two extreme trials to probe the participants' inclination towards monotasking under varying degrees of obviousness: the 95/51 trial, where choosing monotasking is intuitive, and the 95/94 trial, where such a choice is unintuitive. Moreover, the straightforward 95/51 trial was also included for groups of size 3, serving as further attention test and validation that participants understand the task at hand. Trials were presented to participants in one of four randomised orders.

Aggregation part. Participants next completed the aggregation part, which consisted again of one practice trial and of ten judgment aggregation trials. For an example of an aggregation trial, see Figure 3. Participants were shown judgments collected from teams of workers of varying sizes and task accuracies who had performed the pie quality check tasks in one of the task assignment configurations described in the task assignment part above. Each trial corresponded to a different team of workers and participants were asked, in a forced-choice paradigm, to either approve or reject each of the two pie quality checks for the pie under consideration. Importantly, participants were not able to verify the pie quality checks themselves but instead had to rely on the judgments provided by the team workers to make their decisions. Note also that, although most of the teams shown in this part had identical team sizes and accuracies to those used in the task assignment part, they were not necessarily shown in the same task assignment configurations as those chosen by participants in the task assignment part.

Team size, monotasking and multitasking accuracies, and collected judgments for each trial are reported in Table 2. The specific values for p and q were selected so that the different hypotheses to be tested could be distinguishable (that is, the corresponding assignment methods would give sufficiently contrasting recommendations). We tested

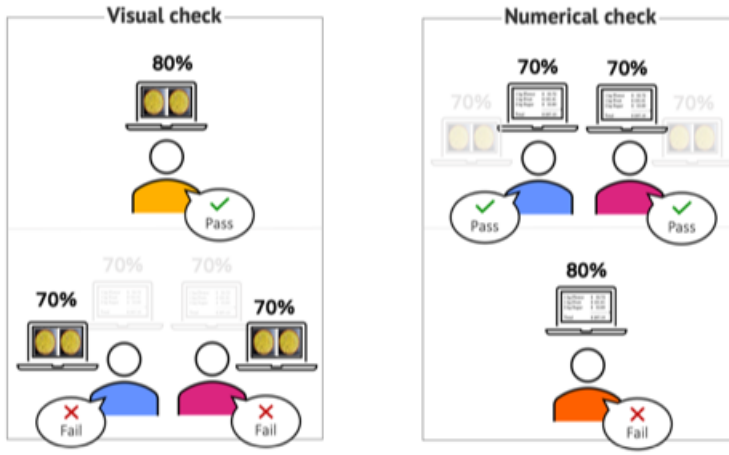


Figure 3. Aggregation trial for a team of size 4, with $q = 0.80$ and $p = 0.70$.

Table 2. Team size, task accuracies and collected judgments per trial and task for the aggregation phase.

			Visual task			Numerical task		
Trial	q	p	Case	Monot.	Multit.	Case	Monot.	Multit.
1	.80	—	80A	A	—	80A	A	—
2	.95	—	95A	A	—	95A	R	—
3	.80	.70	80A-70R	A	R	70A-80R	R	R
4	—	.80	80AA-80R	—	AAR	80AA-80R	–	AAR
5	.80	.70	80A-70RR	A	RR	70AA-80R	R	AA
6	.85	.80	80AA-85R	R	AA	85A-80RR	A	RR
7	.95	—	95A-95R	AR	—	95RR	RR	—
8	.85	.70	85A-70RR	A	RR	70AA-85R	R	AA
9	.95	.80	80AA-95R	R	AA	95A-80RR	A	RR
10	.95	.75	75AA-95R	R	AA	75AA-95R	R	AA

Note: A and R indicate, respectively, an "Approve" and "Reject" judgment. In the analyses, we collapse trials and tasks into cases, which correspond to unique decision contexts.

cases with the same accuracy combinations as in the task-assignment part that participants had already seen (80/70, 85/80, and 95/80), building conflict in the answers—for instance, a participant would have to choose whether to agree with a positive judgment of 80% accuracy or two negative judgments of 70% accuracy. Then, we

also included two more conflicting trials, with varying differences between the presented accuracy: the 85/70 trial (which serves as a comparison point with the 95/80 trial by including higher accuracies of the same distance) and the 75/95 trial (which includes accuracies of larger distance than other trials). To avoid overloading participants with challenging questions and promote familiarity with the task, the rest of the aggregation part contained trials devoid of conflict for someone with basic understanding: two trials with a single judgment (which the optimal rule copies), one trial with the same number of supporters for diverging responses (in which the optimal rule chooses the response of higher accuracy), one trial with responses of the same accuracy (in which the optimal rule chooses the response with the most supporters), and one trial with a tie. Trials were presented to participants in one of four randomised orders.

Final part. The experiment concluded with a short questionnaire in which participants were asked to confirm they had completed the experiment without any distractions, and participants were then provided with a debriefing sheet about our research goals.

3.2 Results

We are now ready to present our experimental results.

Analytic approach. To test our predictions regarding the descriptive adequacy and the predictive ability of the OTT model, we fit Bayesian regression models with the `R` package `brms` (Bürkner 2018) and the probabilistic programming language Stan (Carpenter et al. 2017), which uses Markov Chain Monte Carlo algorithms. A Bayesian analysis estimates model parameters as probability distributions, with the joint probability distribution of the data, y, and a given parameter, θ, being computed via the prior probability of θ and the probability $p(y \mid \theta)$:

$$p(y, \theta) \quad = \quad p(y \mid \theta) \times p(\theta)$$

This result is derived from Bayes' Rule, which serves to calculate the posterior probability, $p(\theta \mid y)$, as follows:

$$p(\theta \mid y) \;\propto\; p(y \mid \theta) \times p(\theta) \;=\; p(y, \theta)$$

This posterior probability distribution can be interpreted as indicating the relative plausibility of possible values of the parameter θ, conditional on the prior probability of that parameter, the probability distribution of the responses (or likelihood function), and the data itself.

Because we sought to examine whether participants' responses matched the responses predicted by the OTT model or by one of the alternative heuristic rules or assignment methods, we recoded participants' responses into a series of binary dependent variables which indicated whether a given response had been successfully predicted by each of the competing models or heuristics. As a consequence, we chose to model the probability of a participant's response being successfully predicted by an assignment method or decision rule as arising from a Bernoulli distribution, with our models estimating the logit-transformed probability of an answer being successfully predicted. The logit-transformation converts a probability p (which is, by definition, restricted to the 0 to

1 range) into a log odds ratio by taking the logarithm of the ratio between p and $1 - p$. A log odds ratio of 0 means that p and $1 - p$ are independent, a positive log odds ratio means that p is higher than $1 - p$, and a negative log odds ratio means that p is lower than $1 - p$.

We also specified prior distributions over the possible effects each parameter could have on the probability that a response would be successfully predicted. Specifying these priors is recommended because it allows regularization of parameter estimates (see eg., Bürkner 2018; McElreath 2020). For all models reported in this experiment, we specified weakly informative priors that indicated extreme estimated effects as unlikely while remaining agnostic to the direction of these effects. Finally, because we used a repeated measures design where participants rated multiple items, we also included a (hierarchical) mixed-effects structure to our models, which estimated how group-level (or random) effects deviated from population-level (or main) effects while accounting for possible correlations in responses provided by the same participant.

For all models reported in this paper, MCMC diagnostics indicated sufficient mixing of the chains, sufficiently high bulk and tail effective sample size values and an $\widehat{R}$ convergence diagnostic of 1.00 for all parameters, which is within the recommended value range (Vehtari et al. 2021).

Task assignment part. Recall Table 1 for the different trials of task assignment in our experiment. The choices of the participants are displayed in Figure 4. In all trials except 2-95/94, 4-80/70, and 4-85/80, participants were more likely to choose the option predicted by the OTT assignment method, indicated by ($\star$) on the figure; while in the three cases that constitute an exception participants chose the OTT option less than 50 percent of the time, in trials 4-80/70 and 4-85/80 the OTT option was still selected more often than each one of the alternative options. On the other hand, participants were more likely to choose the option predicted by the monotasking assignment method (i.e., Option 1) in just over half of the trials, and to choose the option predicted by the multitasking assignment method (i.e., Option 3 for teams of size 4 and Option 2 for smaller teams) only in two trials.

Note however that participant choices matched the multitasking option 30 to 40 percent of the time in trials 2-80/70, 3-80/70, 3-95/80, and 4-85/80, even if that was not the optimal one. Also, 80 percent of participants made a non-optimal multitasking decision in trial 2-95/94. These observations hint towards a possible tendency of people to overestimate the importance of additional judgments in cases of conflict. Since this tendency is not always present (see that 40 percent of participants did not choose the optimal multitasking option in trial 3-85/80), future trials could help clarify its persistence, especially on further borderline cases concerning for instance accuracies 80/75 or 75/70 that we have not yet examined.

To compare the three decision rules for the task assignment part, we first fit a Bayesian multivariate mixed-effects logistic regression model, which estimated the overall probability that responses would correspond to the correct option according to each one of the assignment methods: the OTT, Monotasking, and Multitasking dependent variables indicate whether a response follows each assignment method. This model also

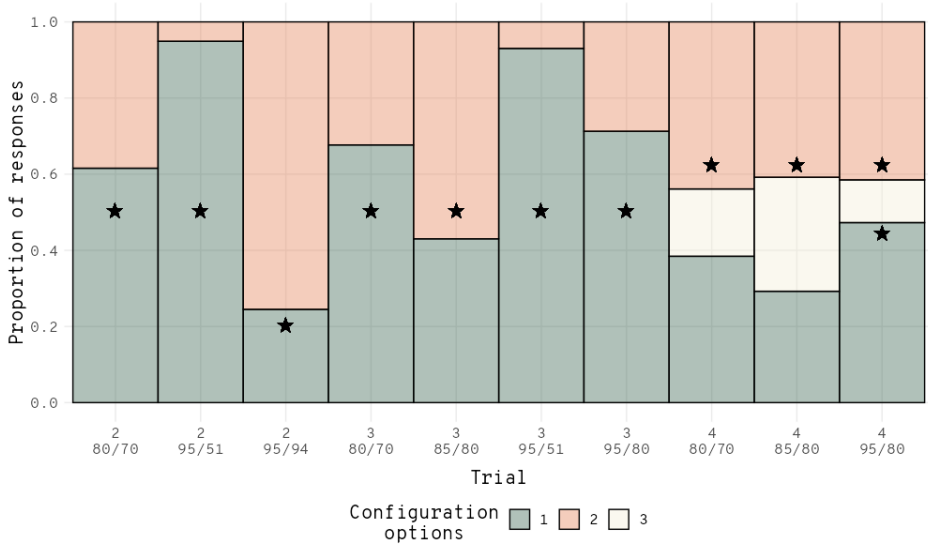


Figure 4. Proportion of participants' task assignment decisions for each trial. ($\star$) indicates the response(s) predicted by the OTT assignment method. The monotasking assignment method predicts that option 1 will be chosen for all cases and the multitasking assignment method predicts that option 2, or 3 when available, will be chosen for all cases.

included a mixed-effects hierarchical structure with participants as a grouping factor:

Model 1: {`OTT, Monotasking, Multitasking`} $\sim$ `1 + (1|Participant)`

According to model 1, the probability that responses would match the Multitasking assignment method was below chance ($b = -0.99$, [-1.10:-0.89] 95 % CI), while the probability that they would match the OTT and Monotasking assignment methods was above chance ($b = 0.60$, [0.54:0.65] 95 % CI and $b = 0.34$, [0.25:0.44] 95 % CI respectively). The difference in probability estimates between the OTT and the Monotasking assignment method was meaningfully different from zero ($b_{diff} = 0.25$, [0.14:0.37] 95 % CI). Probability estimates for the OTT assignment method were higher than estimates for the Monotasking and Multitasking assignment methods (Figure 5).

For a more detailed comparison per trial, we next fit a second Bayesian multivariate mixed-effects logistic regression model, which estimated for each trial the probability that responses would correspond to the correct option according to each one of the assignment methods (with the same dependent variables as in Model 1). This model also included Trial as a categorical predictor as well as a mixed-effects hierarchical structure with participants as a grouping factor:

Model 2: {`OTT, Monotasking, Multitasking`} $\sim$ `0 + Trial + (1|Participant),`

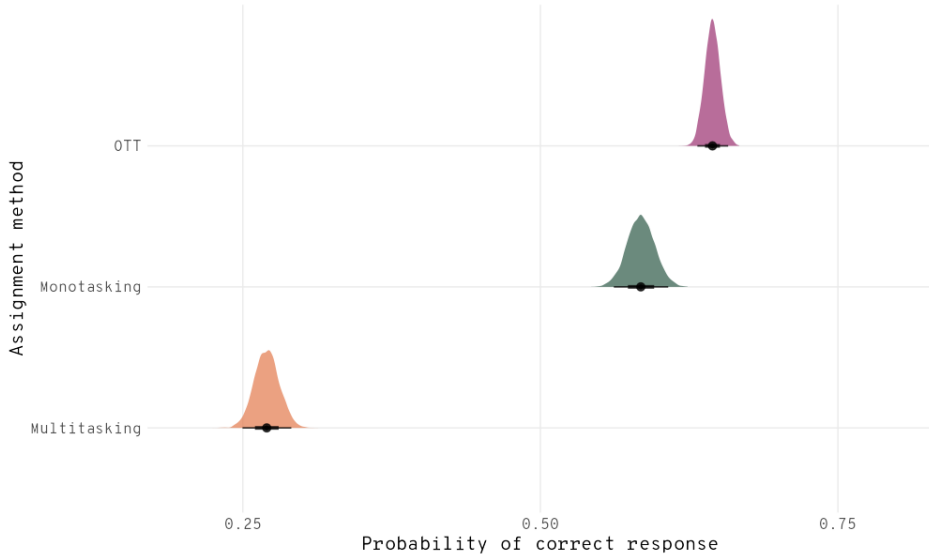


Figure 5. Posterior probability distributions (with mean and 95 percent CI) of the probability estimates that responses will match each one the three assignment methods in the task assignment phase.

where the `0 +` syntax indicates that no separate intercept was estimated for this model. Results for Model 2 are displayed in Figure 6 and indicate that, in almost all trials, the probability that a response would correspond to the correct response according to the OTT assignment method was either higher than the probability that it would correspond to the correct response according to the competing assignment methods, or was equally as high in the trials where one of the competing assignment methods predicted the same option as the OTT assignment method to be the correct one.

Finally, in order to investigate how well the OTT assignment method could predict participants' decisions, we fit a Bayesian mixed-effects logistic regression model with regressed the probability that a response would follow the OTT assignment method on the probability (in log odds units) of that response being the correct one according to the OTT model. In addition, this model had a mixed-effects hierarchical structure with participants as a grouping factor:

Model 3: $\texttt{OTT} \sim logit(\texttt{OTT probability}) + (1|\texttt{Participant})$,

where the logit function is defined as $logit(p) = log(\frac{p}{1-p})$ for $p \in (0, 1)$. This model found a positive relationship ($b = 0.48$, [0.40:0.56] 95 % CI) between the probability that a response would be correct according to the OTT model and the probability that a participant's response would follow that assignment method, indicating that participants became more likely to choose the correct option according to the OTT assignment method as the probability that the given response would be correct according to the OTT

Figure 6. Mean (with 95 percent CI) estimates per trial of the probability that participants' responses correspond to each one of the decision rules in the task assignment phase.

model increased (see Figure 7). Notably, however, even when the OTT model estimated a high 95 percent probability of a response being correct, the predicted probability that participants would choose that option was only around 70 percent.

In future work, an extension of this experiment would be intriguing, including additional trials with new combinations of accuracy values in order to shed more light to the link between *accuracy* and *identifiability* of the OTT options. Such additional trials could *a priori* affect the results presented in Figure 7: for example, by setting $q = 80\%$ and $p = 51\%$ people may still largely opt for the OTT answer, although the probability of that answer being correct is not perfect (which would produce points at the upper left corner of Figure 7). A more systematic study of the relationship between the probability of the OTT answer being correct and the probability that participants choose this answer would include diverse trials where the difference of accuracies $q - p$ is fixed but the accuracy of the OTT answer varies, such as the trials 2-70/(51,55,60,65), 2-80/(60,65,70,75), and 2-90/(70,75,80,85). It is unclear what observations one could expect in these trials since no comparative conclusions can be obtained from our current experiment. Until follow-up work has been conducted, we must stress that Figure 7 should only be interpreted as an illustration of preliminary observations on this topic, strongly connected to the selected trials we have performed.

Aggregation part. The responses of the participants concerning aggregation decisions are presented in Figure 8 (recall the trials included in this part of our experiment, from Table 2). There are some cases in which the three decision rules, and especially the majority rule, are unable to discriminate between the approve and reject decision and

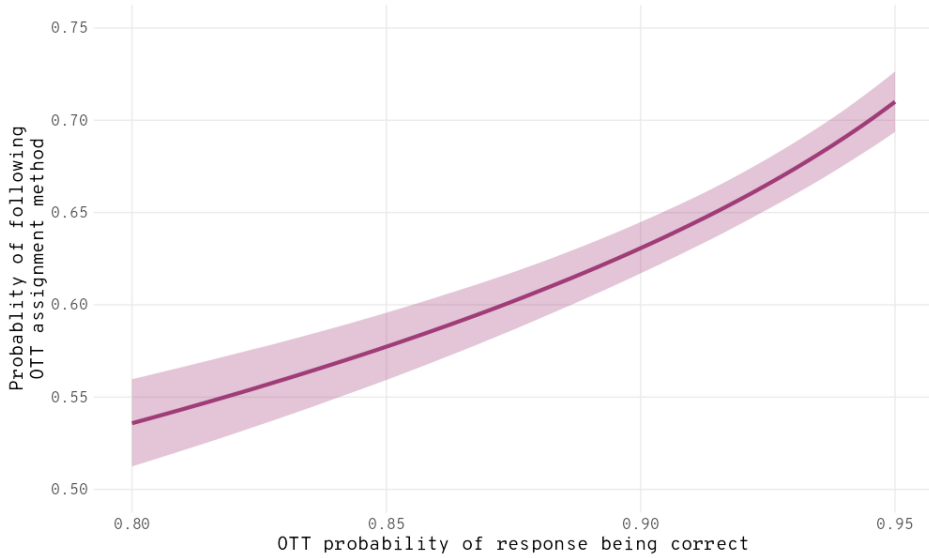


Figure 7. Conditional effects plot (with 95 percent compatibility interval) for the relationship between the probability of an option being correct estimated by the OTT model and the probability that a participant's response will match that option.

make an explicit recommendation. Focusing only on those cases where the decision rules are discriminating, we can see that the OTT rule, indicated again by (⋆), matches participants' responses in three out of four trials, while the majority rule (indicated by ⊛) matches those responses in over four out of five trials, and the accuracy rule (indicated by ⧫) only matches them in half the cases.

Note that the simpler trials of the aggregation part indicate that participants understood the task at hand: when there was no conflict (e.g., in trials such as 70Y-80N or 95N), the optimal answer was chosen most often.

Analogously to the task assignment phase, we detect a discernible inclination among participants to disproportionately value the contribution of supplementary judgments of low accuracy. Consider, for example, the trial 85Y-70NN, where 70 percent of the participants chose the negative answer that had two supporters, although the optimal answer was the 'yes' proposed by the one high-accuracy judgment. Given our limitation to only observe participant responses to trials without insight into the cognitive mechanisms behind these responses, distinguishing between the following two plausible interpretations is challenging. The first interpretation suggests that participants' reasoning aligns with the OTT rule but applies excessively high weights to judgments of lower accuracy. The second interpretation proposes that participants internally address the aggregation task via a majority-based approach.

To test our original hypothesis, we compared the three decision rules for the aggregation part by fitting a Bayesian multivariate mixed-effects logistic regression

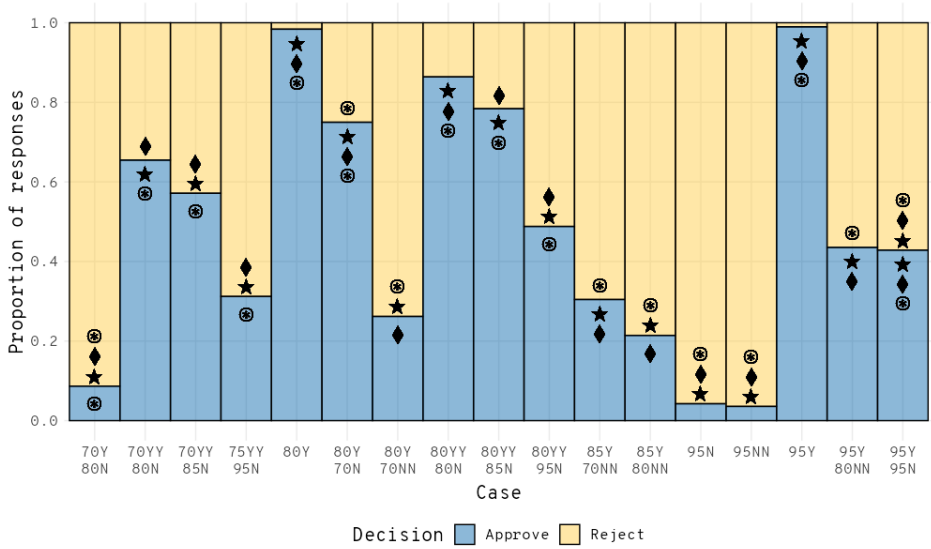


Figure 8. Proportion of participants' aggregation decisions for each case. ($\star$) indicates response(s) predicted by the OTT rule, ($\blacklozenge$) those predicted by the accuracy rule and ($\circledast$) those predicted by the majority rule.

model, which estimated the overall probability that responses would correspond to the correct option according to each one of the decision rules (the OTT, Accuracy and Majority dependent variables indicate whether a response is correct or not according to each rule). Note that Models 4 and 5 constituted a deviation from our preregistered analyses, in which we had planned to compare the accuracy of the three decision rules by their respective Areas Under the Curve. With this approach, we found that the OTT rule ($AUC = 0.76$) and the Majority rule ($AUC = 0.77$) performed approximately equally well, while the Accuracy rule performed worse ($AUC = 0.66$). We choose to report Models 4 and 5 instead because they allow for a more direct comparison between the two parts of the experiment. This model also included a mixed-effects hierarchical structure with participants as a grouping factor:

Model 4: $\{$`OTT, Accuracy, Majority`$\}$ $\sim$ `1 + (1|Participant)`

Model 4 estimates indicated that the probability that responses would correspond to a decision rule was above chance for all three rules: ($b = 1.21$, [1.15:1.26] 95 % CI for the OTT rule, $b = 0.75$, [0.69:0.81] 95 % CI for the Accuracy rule, and $b = 1.43$, [1.34:1.51] 95 % CI for the Majority rule), with the difference in probability between the OTT and Majority rules estimates ($b_{diff} = -0.22$, [-0.32:-0.12] 95 % CI) indicating that the responses corresponding to the OTT rule were meaningfully less likely than responses corresponding to the Majority rule. As shown in Figure 9, probability estimates for the OTT and Majority rules were higher than the estimates for the Accuracy rule. Note

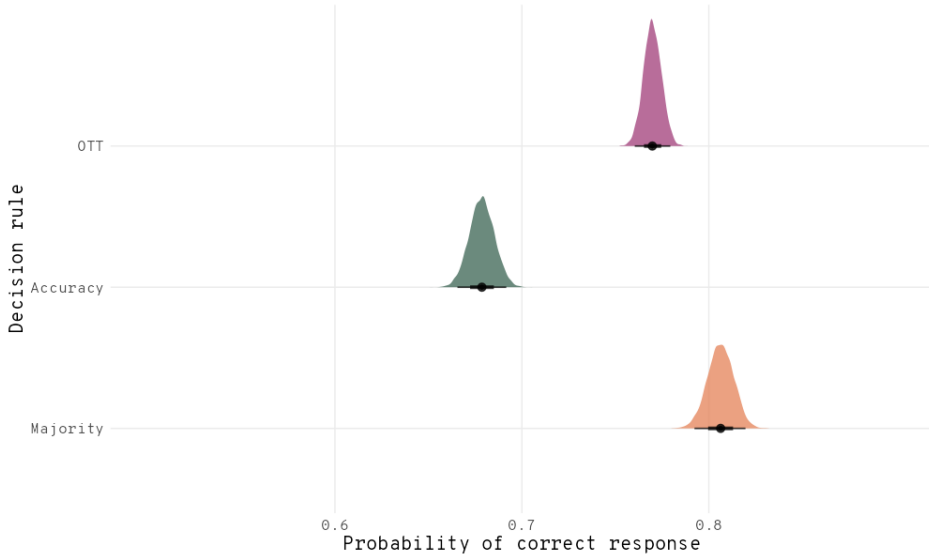


Figure 9. Posterior probability distributions (with mean and 95 percent CI) of the probability estimates that responses will follow each decision rule in the aggregation phase.

however that one limitation of this model, and of Model 5 below, is that in the cases where a decision rule did not make a discriminating recommendation between the Approve and Reject decision, both choices were rated as correct under that rule. Because the Majority rule did not make discriminating recommendations in three cases, versus only one case for the OTT and Accuracy rules, this approach was somewhat biased in favour of the Majority rule.

Model 5: {OTT, Accuracy, Majority} ~ 0 + Case + (1|Participant)

Model 5 found that, in almost all cases, the estimated probability that participant responses would correspond to the Majority rule was either the highest or equally as high in the cases where the competing rules made identical predictions for the correct response (see Figure 10). The OTT rule did second best, with the probability that participant responses would correspond to the responses it predicted being either the highest or equally as high in more than half the cases and the Accuracy rule did the worst, with its estimated probability of matching participant responses in less than half the cases.

We concluded our analysis by examining, again, how well the OTT rule could predict participants' responses. Instead of modelling separately 'Approve' and 'Reject' decisions, we computed the normalised probability (in log odds units) that the 'Approve' decision would be the correct one according to the OTT model and used it to predict, in a Bayesian mixed-effects logistic regression model the probability that participants would make the 'Approve' decision. In the cases where this normalised probability is

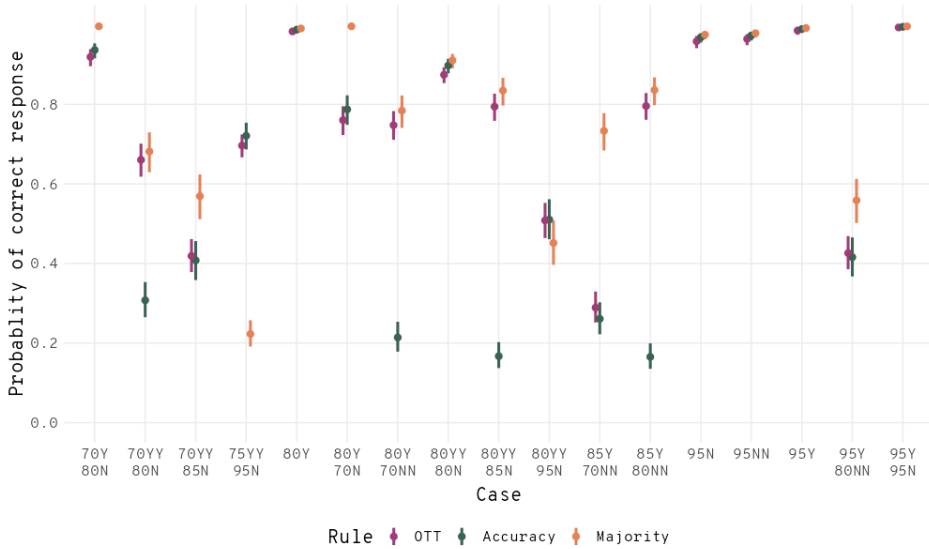


Figure 10. Mean (with 95 percent CI) estimates per case of the probability that participants' responses correspond to each one of the decision rules in the aggregation phase.

below 0.50, the OTT model predicts the 'Reject' decision to be correct instead and we would expect participants to become more likely to choose the 'Reject' decision, and therefore less likely to choose the 'Approve' decision. This model, which also included a mixed-effects hierarchical structure, with participants as a grouping factor, was defined as follows:

Model 6: $\texttt{OTT} \sim logit(\texttt{OTT probability}) + (1|\texttt{Participant})$

Model 6 revealed a steep positive relationship ($b = 1.92$, [1.49:2.27] 95 % CI) between the probability estimated by the OTT model that the 'Approve' decision would be correct and the probability that participants would choose that option. Notably, Figure 11 shows that the OTT probability of 'Approve' being correct tended to overestimate the probability that participants would choose that option on the range from 0.00 to 0.50, and to underestimate it on the range from 0.50 to 1.00. This indicates that when the probability (according to the OTT model) that the 'Approve' decision would be correct was below chance, participants were less likely to choose it than was predicted by the OTT model; and when the probability (according to the OTT model) of the 'Approve' decision being correct was above chance, participants were more likely to choose it than was predicted by the OTT model. If we recall that the probability of the 'Approve' decision being correct is the complement of the probability of the 'Reject' decision being correct, we see that overall, participants made a given decision more often than what the OTT model suggests.

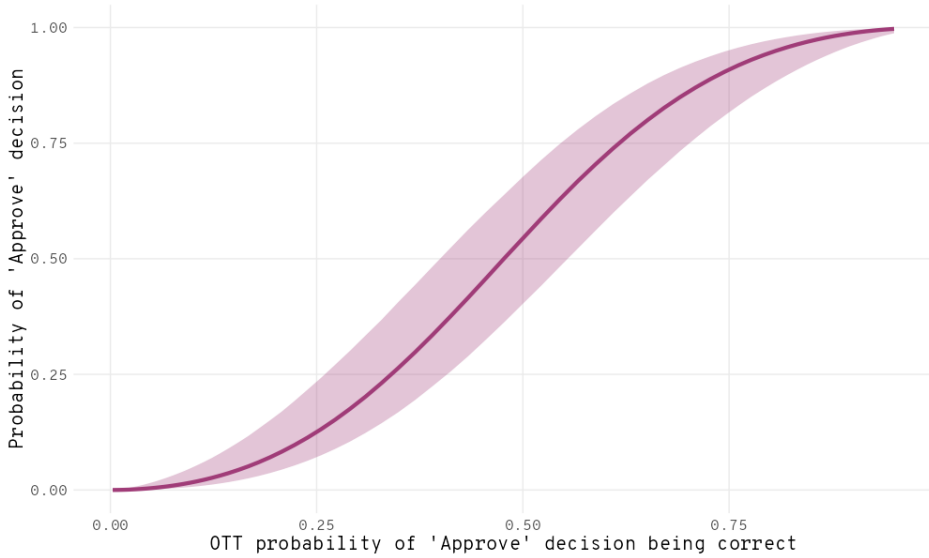


Figure 11. Conditional effects plot (with 95 percent compatibility interval) for the relationship between the probability of an 'Approve' decision being correct according to the OTT rule and the probability that a participant will choose the 'Approve' response.

In this aggregation part, further experimentation could shed more light to interesting unexplored trials such as those with a single judgment of low accuracy (e.g., 55Y or 60YY): whether people's reactions would vary in response to the currently utilised single-judgment trials with high accuracy, and whether the graph of Figure 11 would be affected, is an open question.

Individual differences Because we noted a curious pattern between the two phases of the experiment, where participant answers seemed to more frequently match the OTT suggestions in the task assignment part but to instead favour majority recommendations in the aggregation part, we chose to further explore our results by examining individual differences in responses. We conducted this exploratory analysis by extracting from Models 1 and 4 the group-level estimated probabilities for all the decision rules for each participant. These group-level effects take the form of group-level deviates, which the statistical model assumes to be normally distributed around the population-level effect. Because Bayesian inference works with samples from the posterior distribution of the estimated parameters, we are able to directly compute the estimated probability of each decision rule for each participant by adding, for each sample in the posterior distribution, the estimated population-level probability of each decision with the corresponding estimated group-level deviate. We classified participants as aligning with the responses of a given rule when the mean probability that their answers would be correct according to that rule was strictly higher than the mean probabilities that their answers would be correct according to either of the two other rules.

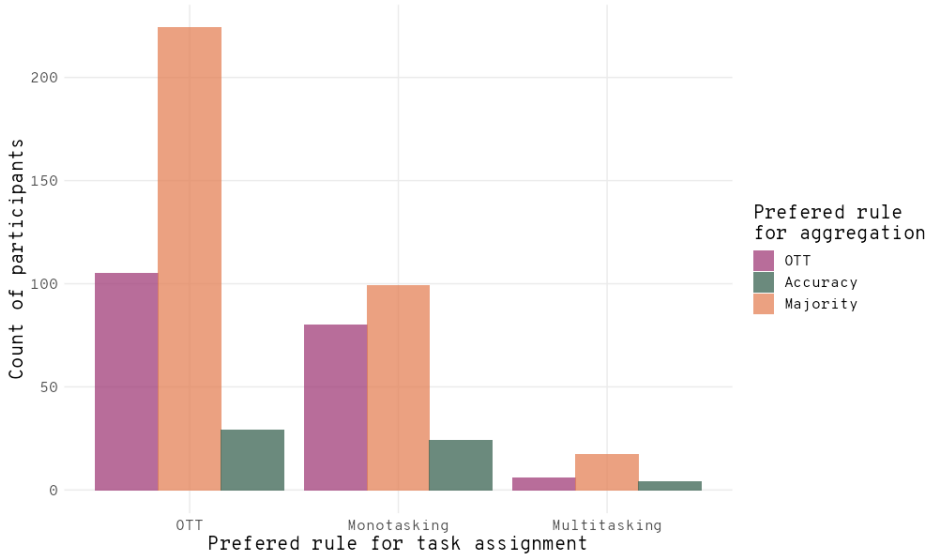


Figure 12. Counts of participants according to whether they were estimated, by Models 1 and 4, as favouring one of the three competing decision rules for the task assignment and aggregation parts.

Figure 12 displays the results from this analysis. It indicates that, for the task assignment part, the responses of a large number of participants were best described as following the OTT rule. However, when considering which decision rule best described participant responses in the aggregation part, the Majority rule was the clear winner, even for participants who were classified in the task assignment phase as favouring a Monotasking rule, which maximized the quality over the quantity of the judgments.

4 Discussion

We have investigated how people behave in contexts where evidence management and aggregation decisions need to be made. If collecting a large quantity of evidence means that the evidence quality is compromised, team leaders have to tackle an apparent trade-off between the number of judgments they collect from their team members on the one hand, and the accuracy of these judgments on the other hand. Work by Terzopoulou and Endriss (2019), summarised in Section 2, has shown that in many cases the trade-off cannot be resolved in an easy fashion—a complex mathematical analysis is required to know exactly where the line should be drawn when choosing whether to favour quality or quantity. In practice, it is not always sensible to expect people's choices to perfectly match the optimal ones; to counteract complexity, decision makers may employ shortcuts or heuristics. Our main question thus arises: how good are people's choices with respect to the ideal choices proposed by the normative model?

Throughout our experiment, it is evident that participants grasp the task's objective: they strive to enhance decision accuracy, meaning that they attempt to follow the OTT's recommendations. They clearly succeed when quantity and quality do not conflict or when resolving such conflicts seems straightforward. This finding is encouraging, as people frequently participate in related activities from diverse perspectives and roles. The absence of strong biases suggests that people are able to make accurate decisions, thereby improving collective performance. Furthermore, possessing a formal model that accurately elucidates human behaviour is crucial, especially in times when the pursuit of explainable AI that mimics human decision-making is of paramount importance.

When significant trade-offs between quantity and quality emerge, participant responses vary. In the task assignment phase, many participants frequently made optimal choices (beyond merely trivial scenarios), with the likelihood of selecting the normatively correct option rising as this option becomes more apt to yield a correct recommendation. In the aggregation phase, participants often favoured options endorsed by the majority rule, even if the majority's judgments lacked high accuracy. This brings to the surface a plausible interpretation: that people tend to overvalue the addition of new judgments, regardless of those judgments' accuracy. During an aggregation process, such a tendency may often lead to responses that are practically indistinguishable from the ones induced by the majority rule. Moreover, observations from the task assignment phase offer further support for this interpretation: while most people's choices appeared to adhere to the OTT options, a substantial portion still opted to increase the amount of judgments gathered.

Several factors could motivate people to prioritise additional judgments beyond an optimal level, neglecting quality in light of quantity. One potential influence may be the societal norm that equates majority opinion with the democratic ideal, especially in consensus-seeking situations. This cultural predisposition may lead people to perceive a higher number of endorsements for an answer as inherently beneficial, encouraging them to default to this approach when no alternative decision-making strategy is apparent.

Lastly, it is critical to acknowledge that our experiment, like all studies in the social sciences, is not without its flaws and offers insights that are context-bound and subject to its limitations. Importantly, we should stress that the selection of questions posed to participants likely played a principal role in the reported results. Furthermore, we can only observe participant decisions and verify their alignment with the recommendations of a given rule, but lack the means to directly examine the cognitive processes behind these decisions. So, any assertions about whether participant choices are guided by the specified rule, a different one, or something else, are only conjectures. In other words, our study focuses on documenting the outcomes of decisions rather than the underlying motivations or thought processes. This critical distinction paves the path for future research into the cognitive mechanisms potentially influencing decision-making in scenarios similar to the one investigated here. Our work is intended as an initial foray into a broader research trajectory. Empirical testing would be valuable for many other formal models within the literature of social choice, where the potential divergence between normative assumptions and real people's behaviour is often understudied.

Acknowledgements

We would like to acknowledge the support of the Dutch Research Council (NWO) through project 612.001.652 ("Customisable Collective Choice"). We are grateful to the two anonymous reviewers of Rationality and Society, for their insightful comments that significantly improved this paper. We also thank Ulle Endriss as well as the participants of several seminars where earlier versions of this work has been presented, for their constructive feedback.

References

Adler RF and Benbunan-Fich R (2012) Juggling on a high wire: Multitasking effects on performance. *International Journal of Human-Computer Studies* 70(2): 156–168.

Ariely D and Zakay D (2001) A timely account of the role of duration in decision making. *Acta Psychologica* 108(2): 187–207.

Arrow K, Sen A and Suzumura K (eds.) (2002) *Handbook of Social Choice and Welfare*. North-Holland.

Arrow KJ, Blackwell D and Girshick MA (1949) Bayes and minimax solutions of sequential decision problems. *Econometrica, Journal of the Econometric Society* : 213–244.

Azevedo EM, Deng A, Montiel Olea JL, Rao J and Weyl EG (2020) A/B testing with fat tails. *Journal of Political Economy* 128(12): 4614–000.

Bassi A (2015) Voting systems and strategic manipulation: An experimental study. *Journal of Theoretical Politics* 27(1): 58–85.

Brandt F, Conitzer V, Endriss U, Lang J and Procaccia AD (eds.) (2016) *Handbook of Computational Social Choice*. Cambridge University Press.

Bubeck S, Cesa-Bianchi N et al. (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* 5(1): 1–122.

Bürkner PC (2018) Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal* 10(1): 395–411.

Caragiannis I and Micha E (2017) Learning a ground truth ranking using noisy approval votes. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*.

Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P and Riddell A (2017) Stan: A probabilistic programming language. *Journal of Statistical Software, Articles* 76(1): 1–32.

de Condorcet M (1785) *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale.

Croson R (2005) The method of experimental economics. *International Negotiation* 10(1): 131–148.

Edland A and Svenson O (1993) Judgment and decision making under time pressure. In: *Time Pressure and Stress in Human Judgment and Decision Making*. Springer, pp. 27–40.

Hartmann S, Pigozzi G and Sprenger J (2010) Reliable methods of judgement aggregation. *Journal of Logic and Computation* 20(2): 603–617.

Hertwig R and Ortmann A (2001) Experimental practices in economics: A methodological

challenge for psychologists? *Behavioral and Brain Sciences* 24(3): 383–403.

Hertwig R and Ortmann A (2003) Economists' and psychologists' experimental practices: How they differ, why they differ, and how they could converge. In: *The Psychology of Economic Decisions*. Oxford University Press, pp. 253–272.

Kahneman D (2003) Maps of bounded rationality: Psychology for behavioral economics. *American economic review* 93(5): 1449–1475.

McElreath R (2020) *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Meir R, Gal K and Tal M (2020) Strategic voting in the lab: Compromise and leader bias behavior. *Autonomous Agents and Multi-Agent Systems* 34(1): 1–37.

Moscarini G and Smith L (2001) The optimal level of experimentation. *Econometrica* 69(6): 1629–1644.

Payne JW, Bettman JR and Johnson EJ (1988) Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14(3): 534–552.

Terzopoulou Z and Endriss U (2019) Optimal truth-tracking rules for the aggregation of incomplete judgments. In: *Proceedings of the 12th International Symposium on Algorithmic Game Theory (SAGT)*.

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3-4): 285–294.

Tversky A and Kahneman D (1985) The framing of decisions and the psychology of choice. In: *Behavioral Decision Making*. Springer, pp. 25–41.

Tversky A and Kahneman D (1989) Rational choice and the framing of decisions. In: *Multiple Criteria Decision Making and Risk Analysis Using Microcomputers*. pp. 81–126.

Vehtari A, Gelman A, Simpson D, Carpenter B and Bürkner PC (2021) Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis* 16(2): 667–718.

Voslinsky A and Azar OH (2021) Incentives in experimental economics. *Journal of Behavioral and Experimental Economics* 93: 101706.

Wilhelm O and Schulze R (2002) The relation of speeded and unspeeded reasoning with mental speed. *Intelligence* 30(6): 537–554.

Zou J, Meir R and Parkes D (2015) Strategic voting behavior in doodle polls. In: *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*.