The presence of irrelevant alternatives paradoxically increases confidence in perceptual decisions

Nicolás A. Comay[1+], Gabriel Della Bella[1], Pedro Lamberti[2], Mariano Sigman[3,4], Guillermo Solovey[5*], Pablo Barttfeld[1*]

1- Cognitive Science Group. Instituto de Investigaciones Psicológicas (IIPsi, CONICET-UNC), Facultad de Psicología, Universidad Nacional de Córdoba, Córdoba, Argentina.

2- Facultad de Matemática, Astronomía, Física y Computación, Universidad Nacional de Córdoba, Córdoba, Argentina.

3- Laboratorio de Neurociencia, Universidad Torcuato Di Tella, Buenos Aires, Argentina.

4- Facultad de Lenguas y Educación, Universidad Nebrija, Madrid, Spain.

5- Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, UBA-CONICET, Buenos Aires, Argentina

+ corresponding author
** last authors*

Abstract

Confidence in perceptual decisions often reflects the probability of being correct. Hence, we predicted that confidence should be unaffected or be minimally decreased by the presence of irrelevant alternatives. To test this prediction, we designed three experiments. In Experiment 1, participants had to identify the largest geometrical shape among two or three alternatives. In the three-alternative condition, one of the shapes was much smaller than the other two, being a clearly incorrect choice. Counter-intuitively, all else being equal, confidence was higher when the irrelevant alternative was present. We accounted for this effect with a computational model where confidence increases monotonically with the number of irrelevant alternatives, a prediction we confirmed in Experiment 2. In Experiment 3, we evaluated whether this effect replicated in a categorical task, but we did not find supporting evidence. Our findings stimulate the use of multi-alternative decision-making tasks to build a thorough understanding of confidence.

Keywords: *confidence, perceptual decision making, multiple alternatives, bayesian confidence hypothesis, computational modeling, open data*

Statement of relevance

Our decisions are always accompanied by a degree of confidence in being correct. Therefore, a deep understanding of this process is critical since it partakes in everyday decisions and fields such as politics, economics, medicine, etc. In the lab, confidence has been studied mostly using simple two choice tasks. In real life, however, we often judge confidence in decisions among multiple alternatives. Our research shows that confidence may increase as irrelevant alternatives are added, a result that is at odds with the classical view of confidence as the probability of being correct, and calls for more accurate models of confidence.

Introduction

We constantly make decisions under uncertainty, accompanied by a feeling of confidence -a belief or a subjective feeling that our own thoughts, knowledge, performance, percepts or decisions are correct (Grimaldi et al., 2015; Mamassian, 2016; Meyniel et al., 2015). To be useful in our everyday activities, confidence has to reflect the true likelihood of being correct (Ais et al., 2016). However, this correspondence is not perfect, and several biases have been described in the literature (Rahnev & Denison, 2018), including overconfidence in difficult tasks and underconfidence in easy tasks (Baranski & Petrusic, 1994; Gigerenzer et al., 1991), a mismatch between confidence and accuracy on individual trials (Maniscalco & Lau, 2012), and overconfidence due to misperception of stimulus variability (Zylberberg et al., 2014). However, despite these findings, the dominant view of the field is that confidence reflects the probability that the decision made is correct. This definition is known as the Bayesian confidence hypothesis (BCH) (H.-H. Li & Ma, 2020; Meyniel et al., 2015; Sanders et al., 2016) and is inline with the classical assumption —known as independence of irrelevant alternatives (Luce, 1979)— that the subjective value of a stimulus is only dependent on the stimulus properties.

This hypothesis is based mostly on 2-alternative forced choice tasks (2AFC), because they allow a relatively easy and straightforward computational modeling of confidence (Rahnev, 2020) using diverse approaches, such as signal detection theory (Fleming & Lau, 2014), accumulation of evidence (Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2013) and Bayesian modeling (Drugowitsch et al., 2014; Galvin et al., 2003; Kepecs & Mainen, 2012; Mamassian, 2016; Pouget et al., 2016). However, while methodologically useful, 2AFC tasks are a large oversimplification of the decision-making process, often implying more than two (or an undefined number of) options (H.-H. Li & Ma, 2020; Niwa & Ditterich, 2008; Rahnev, 2020). Including a larger number of alternatives may not only reveal that the methodological tools developed for 2AFC tasks are insufficient, but also that the proposed mechanisms are wrong (H.-H. Li & Ma, 2020; Rahnev et al., 2021). Specifically, the context in which we make decisions might alter our subjective interpretation of the facts (Ariely & Wallsten, 1995; Northoff & Mushiake, 2020). For instance, a phenomenon called divisive normalization —originally described as a neuronal adaptation mechanism (Heeger, 1992)— has been proposed in the field of neuroeconomics to account for irrational value assignment. The probability of choosing a stimulus might negatively depend on the value or the number of the alternative options (Louie et al., 2011, 2013). Moreover, it has recently been proposed that the allocation of attention —modulated by the value of stimuli (Gluth et al., 2020) or its location (Winter & Peters, 2021)— might also give rise to context influence on decision-making. Nevertheless, partly due to the prevalence of 2AFC tasks paradigms, the impact of these effects on confidence remains unknown.

Building on these ideas we explored whether a phenomenon already described in the literature for likelihood judgments (Windschitl & Chambers, 2004) line-up identifications (Charman et al., 2011) and associative memory (Hanczakowski et al., 2014) takes place in multi-alternative perceptual decision-making. This effect is defined as "*the tendency to become more confident that a chosen response option is correct if it is surrounded by implausible response options*" (Charman et al., 2011, p. 479) and is called the "dud-alternative effect". Bayesian models of confidence predict that confidence should remain virtually invariant when adding irrelevant alternatives, because these hardly affect the probability of being correct. Moreover, they practically do not affect the difference between the posterior probabilities that the two best options are correct, a recently proposed computation for confidence in multi-alternative tasks (H.-H. Li & Ma, 2020). On the other hand, the dud-alternative effect found in non-perceptual tasks predicts that confidence should paradoxically increase.

We investigated the influence of irrelevant alternatives on confidence in three visual decision making tasks. In Experiment 1 we studied the effect of a third, irrelevant stimulus on confidence and then contrasted our data with predictions from four computational models. In Experiment 2 we replicated the results from Experiment 1 and, based on the models' simulations, extended them up to three irrelevant stimuli. In Experiment 3 we explored how stimuli-dependent our results are, using a task recently introduced to study confidence in a three alternative decision making context (H.-H. Li & Ma, 2020).

Experiment 1

We designed a size discrimination task to evaluate the effect of the presence of a third, weak alternative (i.e. dud alternative) in confidence (Fig. 1). This task allowed us to manipulate the subjective value of the alternatives, making them more or less eligible only by varying their relative size.

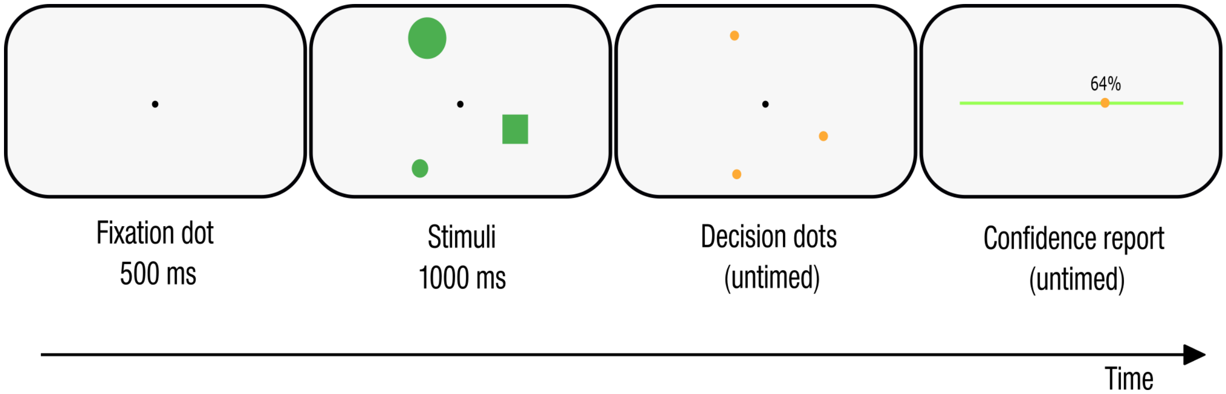

Fig. 1. Size discrimination task. Participants had to decide which was the largest shape. The size of the stimuli varied in every trial, and the relative size of the two largest shapes controlled the difficulty (established in 5 levels). Half of the trials had 2 alternatives and the other half had 3 alternatives, making up a total of 120 trials. The third alternative was very small when present, and varied in size (see Methods). After choosing the largest shape, participants had to report how confident they felt about their decision on a scale from 0% (not confident at all) to 100% (completely confident).

Methods

The task was programmed in JavaScript using the library jQuery. It ran on a JATOS (Lange et al., 2015) server on a DonWeb hosting (donweb.com). The experiment could be run on mobile devices or desktop computers.

Participants: 101 participants took part in Experiment 1. The study was approved by the ethical committee of the Institute of Psychological Research (IIPsi - CONICET - Córdoba, Argentina). Participants should read and accept an informed consent prior to the realization of the experiment. All participants reported no psychiatric, illegal drug consumption or neurological history. Instructions were written on screen prior to the experiment. 34% of the sample performed the experiment on a mobile device, whereas the rest performed it on a computer.

Experimental design: the experiment consisted of a decision making task where participants had to decide which of two or three geometric shapes was the largest. It had 120 trials, 60 with two alternatives and 60 with three alternatives. The shapes were circles and squares, and the two major shapes were always a circle and a square (to make the comparison harder). On half of the trials a circle was the correct option, and on the other half a square. The third alternative, when present, was randomly a square or a circle, and it was much smaller than the other 2 shapes in order to make it easily discarded (in fact, it was only chosen on 7 out of 11751 trials). The size of the largest shape was sampled from a normal distribution with a mean value that depended on screen size (7500px^2 for a screen with a width size >= 600px; 6500px^2 for a screen with a width size < 600px and >= 400; 5500px^2 for a screen with a width size < 400px) and a standard deviation of one fourth of the mean value. The second largest shape area was set to the 0.7, 0.8, 0.9, 0.93 or 0.95 of the area of the stimulus 1, and the third alternative' area (if present), was set to the 0.07, 0.14, 0.21, 0.28, 0.35 or 0.42 of the area of the stimulus 1. Consequently, there were 60 trials with 2 alternatives (of which there were 12 trials per each size of the second largest alternative) and 60 trials with 3 alternatives (of which there were 2 trials for each combination of size of the third and second largest alternative) (Table 1). Stimuli were randomly displayed in 3 equispaced positions (120 degrees) located in a circular array of a radius that varied with the screen size. The stimuli-array was randomly rotated so the positions changed across trials. Shapes were shown for 1 s, and the participants could respond by clicking the chosen shape or, once disappeared, by clicking a small circular dot marking the position. After that, participants had to report their confidence level on the decision on a scale from 0% (not confidence at all) to 100% (completely confident). Participants had to move a dot that was initially hidden (until they moved the cursor) and appeared in the cursor's position.

Table 1. Task structure.

Third alternative size (area3/area1)	Second alternative size (area2/area1)				
	0.7	0.8	0.9	0.93	0.95
Absent	12 trials	12 trials	12 trials	12 trials	12 trials
0.07	2 trials	2 trials	2 trials	2 trials	2 trials
0.14	2 trials	2 trials	2 trials	2 trials	2 trials
0.21	2 trials	2 trials	2 trials	2 trials	2 trials
0.28	2 trials	2 trials	2 trials	2 trials	2 trials
0.35	2 trials	2 trials	2 trials	2 trials	2 trials
0.42	2 trials	2 trials	2 trials	2 trials	2 trials

Data analysis: We excluded 2 participants from subsequent analysis due to low overall performance (8 and 66% overall accuracy; the median performance for the sample was 85% with an interquartile range (Q3-Q1) of 6.7%). The definitive sample consisted of 99 participants. Our sample size was based on a pilot study with 20 participants in which we observed our main findings. We excluded trials with response times (RT) shorter than 200 ms and larger than 10 s. We conducted repeated measures ANOVAs and Tukey HSD post hoc tests (both with the software Statistica). Our predefined p-value for statistical significance was 0.05. We defined task difficulty as the ratio between the area of the two largest shapes. The ANOVA for incorrect trials was done with the highest levels of difficulty (0.9, 0.93, 0.95) because there were not enough incorrect trials in easier conditions (0.7 and 0.8).

Results

Task difficulty had a significant impact on the response times of both the decision task (type 1 task) and the confidence report (type 2 task). Higher difficulty was associated with larger RT ($F_{4, 392} = 130.95$, $p < 0.000001$, $\eta_p^2 = 0.57$ - type 1 task - Fig. 2a; $F_{4, 392} = 19.36$, $p < 0.000001$, $\eta_p^2 = 0.16$ - type 2 task - Fig. 2b). The presence of a dud alternative in half of the trials did not affect RT in none of the tasks ($F_{1, 98} = 0.79$, $p = 0.37$, $\eta_p^2 = 0.008$ - type 1 task - Fig. 2a; $F_{1, 98} = 0.002$, $p = 0.96$, $\eta_p^2 = 0.00002$ - type 2 task - Fig. 2b).

Participants' performance, as expected, decreased as task difficulty increased irrespective of the number of alternatives ($F_{4, 392} = 374.31$, $p < 0.000001$, $\eta_p^2 = 0.79$) (Fig. 2c). We found a marginally significant effect of the inclusion of the third alternative on performance, restricted to the middle difficulty ($F_{1, 98} = 4.04$, $p = 0.047$, $\eta_p^2 = 0.03$) (Fig. 2c). We didn't find a significant interaction between the number of alternatives and the task difficulty ($F_{4, 392} = 1.79$, $p = 0.12$, $\eta_p^2 = 0.017$) (Fig. 2c).

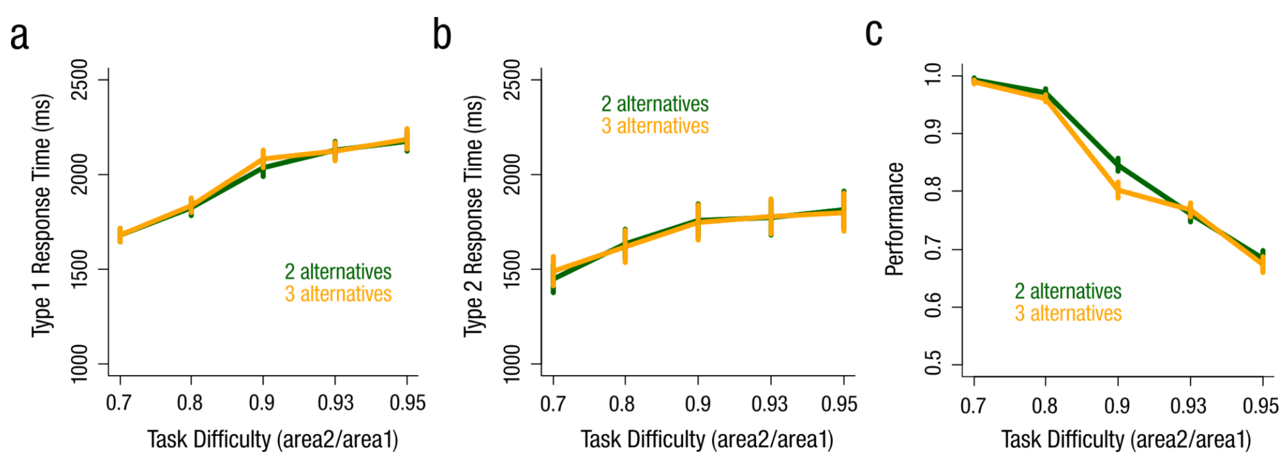

Fig. 2. The dud alternative did not have an effect on response times neither in (a) the decision task nor in (b) the confidence report. The dud alternative only decreased task performance in difficulty 0.9 (c). In all figures vertical bars refer to s.e.m. and "area2" and "area1" refer to the area of the second largest shape and the largest shape, respectively.

Confidence, just like performance and as expected, decreased with increasing levels of task difficulty ($F_{4, 392} = 233.47$, $p < 0.000001$, $\eta_p^2 = 0.7$) (Fig. 3a). Importantly, we found a significant increment of confidence in the three-alternative condition ($F_{1, 98} = 33.25$, $p < 0.000001$, $\eta_p^2 = 0.25$) (Fig. 3a and 3b). Overall, 72 out of the 99 participants had an increased confidence in the three-alternative condition (Fig. 3b). We also find a significant interaction between the difficulty of the task and the presence of a weak alternative ($F_{4, 392} = 6.85$, $p = 0.00002$, $\eta_p^2 = 0.06$) (Fig. 3a).

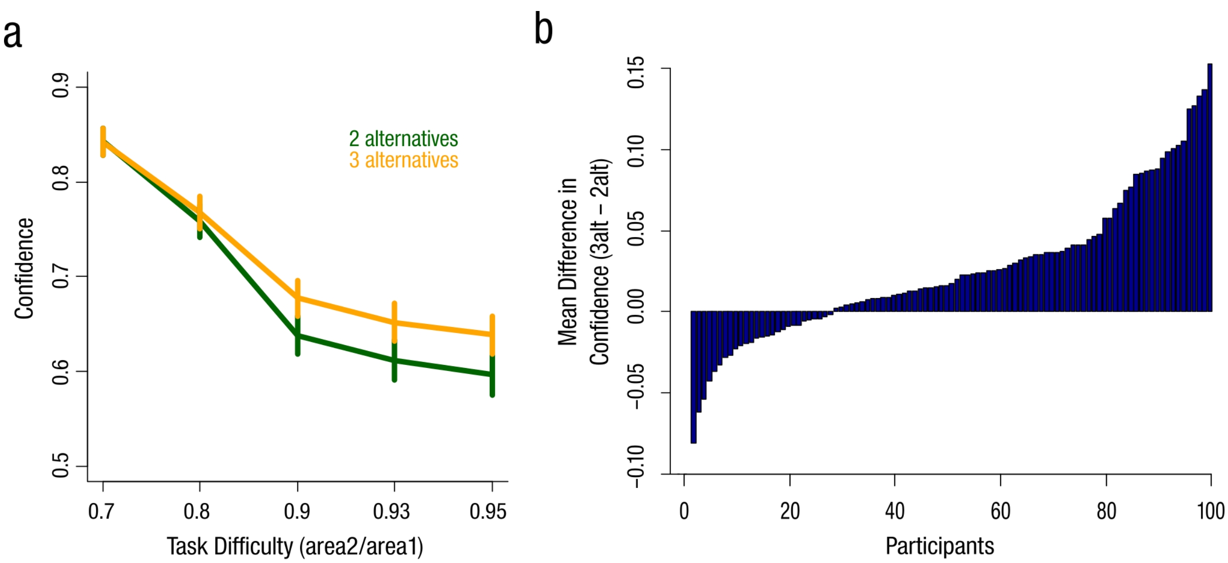

Fig. 3. Confidence (a) diminishes as a function of task difficulty but decreases less when a dud alternative is present (specifically at high levels of difficulty). Difference in confidence (2 vs 3 alternatives) per participant (b). Each line in (b) represents the change in confidence for each participant (considering all difficulty levels for each condition). A positive value means that confidence was higher when the dud alternative was present, whereas a negative value means that, conversely, confidence decreased in that condition.

The effect of increased confidence is present both in correct and incorrect trials ($F_{1, 98} = 26.65$, $p = 0.000001$, $\eta_p^2 = 0.21$ - correct trials; $F_{1, 61} = 22.2$, $p = 0.00001$, $\eta_p^2 = 0.27$ - incorrect trials) (Fig. 4). Moreover, for incorrect trials the mean increment in confidence is higher (Fig. 4c).

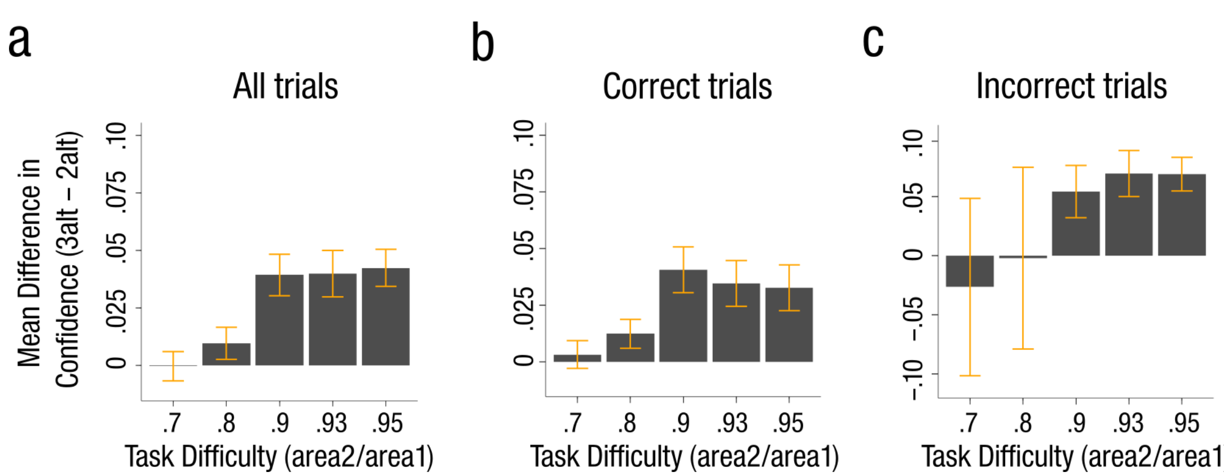

Fig. 4. Confidence increases when the third alternative is present. All panels depict the confidence difference between 3-alternative and 2-alternative conditions. The figure shows confidence difference

for (a) all trials, (b) correct trials and (c) incorrect trials. Confidence difference seems to be higher for incorrect trials (c). Large s.e.m. in (c) at low difficulties (difficulties 0.7 and 0.8) is due to low numbers of incorrect responses.

Computational models

In order to evaluate whether the predictions of different models of confidence could fit our empirical results of Experiment 1 —and to propose some venues for future models— we carried out simulations for 4 models: the Difference model (H.-H. Li & Ma, 2020), the Max model (H.-H. Li & Ma, 2020; Meyniel et al., 2015; Sanders et al., 2016), the Contrast model (Charman et al., 2011; Hanczakowski et al., 2014; Windschitl & Chambers, 2004) and the Average-residual model (Charman et al., 2011; Hanczakowski et al., 2014; Windschitl & Chambers, 2004).

Methods

Computational modeling. All models were developed under a Bayesian modeling approach (Ma, 2019). For all models we started by setting the generative model. Same stimuli as in Exp. 1 were used. Then we followed by simulating the internal responses on each trial. We assumed that observers make a noisy measurement x_i of each stimulus S_i. This measurement was modeled using a Gaussian distribution with mean on the real value of the stimuli and a standard deviation σ. Then, ideal observers compute the posterior probability that each stimuli is the largest, formally: $p_i = p(S_i = max(S_1, S_2, S_3) \mid (x_1, x_2, x_3, s))$. . As some variability in behaviour is due to a "late noise" at the level of the decision variable (H.-H. Li & Ma, 2020), we assumed that observers do not maintain the exact posterior probability but a noisy version of it, q_i, modeled as a random sample of a Dirichlet distribution centered on the true posterior probabilities with a spread controlled by a parameter α (H.-H. Li & Ma, 2020). We assumed participants choose the stimulus with the highest $q(S_i \mid x_1, x_2, x_3, s)$. For readability, from now on we will refer to these noisy posterior probabilities as simply the posterior probabilities.

We simulated confidence levels according to four different models (Fig. 5a). For the Difference model (Diff), confidence level is the result of the difference between the two highest posterior probabilities. In the Max model confidence level is the larger posterior probability. This model corresponds to the Bayesian Confidence hypothesis, stating that confidence reflects the probability that the decision is correct. The Contrast model and the Average-residual model are computational implementations of the two mechanisms proposed in the dud alternative literature to account for the effect (Charman et al., 2011; Windschitl & Chambers, 2004). The Contrast model states that confidence is obtained by a series of pairwise comparisons between the chosen option and the remaining. Following this, we modeled confidence by the sum of the differences between the largest posterior probability and all other posterior probabilities. The Average-residual model proposes that confidence reflects the evidence favoring the chosen option minus the average evidence of the remaining options. Therefore, we modeled confidence as the difference between the alternative with the highest posterior probability and the mean of the remaining posterior probabilities. Formally:

Diff model: $confidence = q_1 - q_2$

Max model: $confidence = q_1$

Average-residual model: $confidence = q_1 - \frac{(q_2+q_3)}{2}$

Contrast Model: $confidence = (q_1 - q_2) + (q_1 - q_3)$

Finally, we also performed simulations for the Average-residual model in a context with 0, 1, 2 or 3 dud-alternatives.

Model fitting. We fitted models' free parameters (i.e. sensory noise, σ, and decision noise, α) to the confidence data of each participant. For that, we first obtained confidence levels reported on each trial for each condition (i.e. five levels of difficulty and two conditions for the number of alternatives) per participant. Then we simulated confidence for 100 x the number of trials for each condition and took the mean of them. Finally, with those simulated numbers, we fitted the model using the least squares method (Farrell & Lewandowsky, 2018). That is, we chose a combination of σ and α that minimized the summed squared error between the model predictions and the data. Using the resulting parameters values we simulated confidence according to each model, rescaling the simulations to the minimum and maximum confidence reported by each subject to keep simulated confidence within the range of our data. To evaluate the ability of our method to detect the true parameters for each subject, we performed parameter recovery by first simulating data using two arbitrary values of σ (0.15) and α (15) and then fitting models to that fake data with the mentioned procedure. We addressed this statistically by performing two one-sample t-tests for recovered sigmas and alphas to see if the obtained distribution of parameters differed significantly from a distribution with mean on the true parameters value. Neither the distribution of estimated σ nor the distribution of estimated α were significantly different from the true value of the parameter (σ: $t_{19} = 0.49$, $p = 0.63$; α: $t_{19} = -0.34$, $p = 0.74$). This suggests that our method successfully recovers the true parameters underlying each participant data. We used the mean squared error (MSE) to arbitrate between models.

Results

For both the Max model (MSE = 0.012) and the Diff model (MSE 0.012) the presence of a third irrelevant option diminishes confidence (Fig. 5a). In contrast, for both the Contrast model (MSE = 0.037) and the Average-residual model (MSE = 0.009) confidence increases with the presence of an irrelevant option (Fig. 5b). However, the Contrast model shows a constant and very high confidence increase, whereas the Average-residual model accounts for the pattern shown in our data: the difference in confidence between three and two alternatives increases with the difficulty. Furthermore, if we extend simulations to a context with more dud alternatives, the Average-residual model predicts that confidence should increase even more (Fig. 5d).

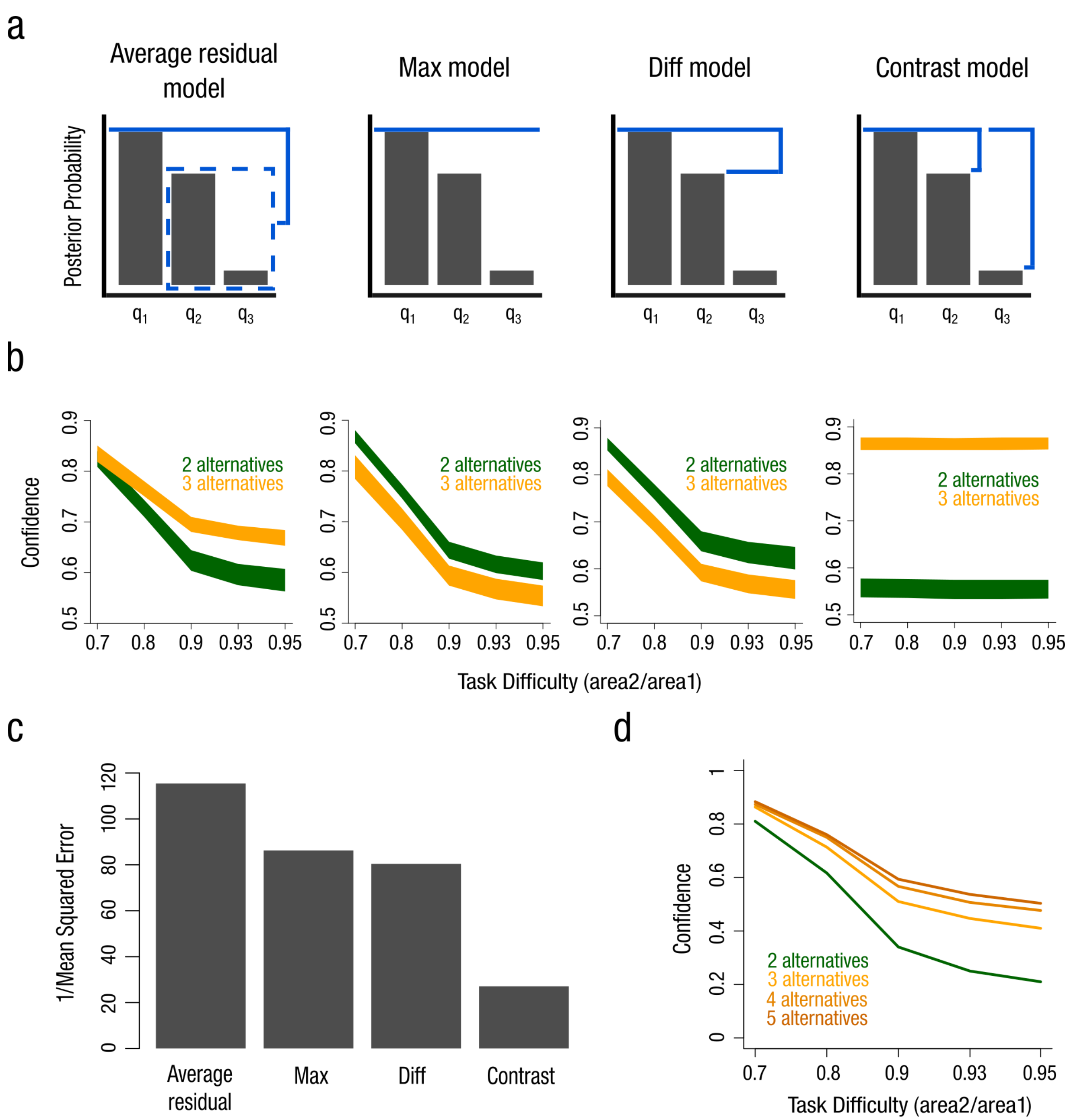

Fig. 5. Sketches of confidence computation for each model (a). In the Average residual Model, confidence is obtained by the difference between the highest posterior probability and the mean of the remaining posterior probabilities. Confidence reflects the probability of being correct in the Max Model. The Difference Model states that confidence reflects the difference between the two best posterior probabilities. Finally, in the Contrast Model confidence is the result of the sum of the differences between the highest posterior probability and each of the remaining posterior probabilities. Panel (b) depicts model fits and panel (c) the inverse of the mean squared error for each model. Both the max model and the Diff model (b) show that confidence should be lower in the 3-alternative condition. Conversely, the Average-residual model and the Contrast model (b) show that confidence should be higher. However, only the Average-residual model replicates the pattern found in our experiment. Shaded regions in (b) are the standard error of the mean of model fits. The Average-residual model (d) predicts that confidence should increase with the number of dud alternatives, as these alternatives

reduce even more the average of the remaining options.

Experiment 2

The Average-residual model predicts that confidence should increase monotonically with the number of dud-alternatives, because the presence of more irrelevant alternatives decreases the average of posterior probabilities of non-chosen alternatives (Fig. 5d). In this second experiment we not only aimed to replicate the results of Experiment 1 but to see whether confidence increases with the number of dud alternatives (Fig. 6).

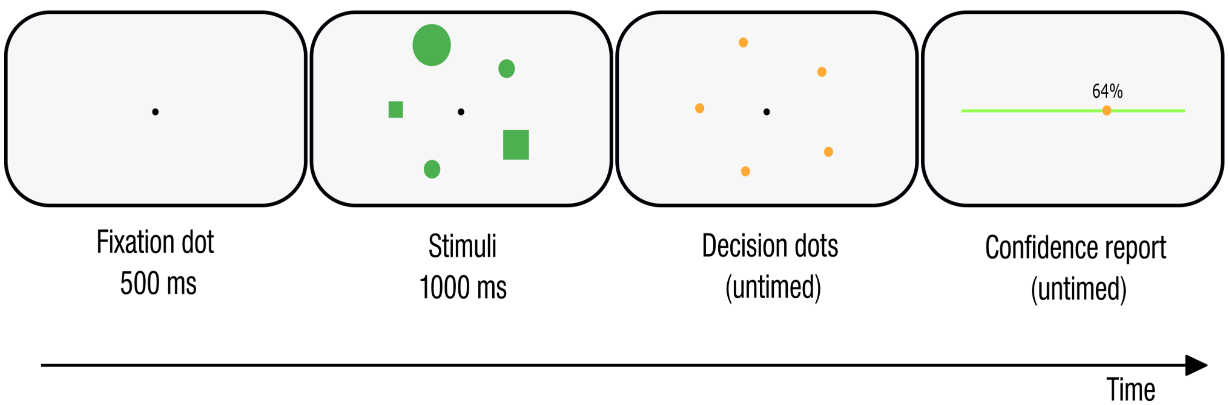

Fig. 6. Size discrimination task. Participants had to decide which one of the shapes was the largest. Stimuli size varied in every trial, and the relative sizes of the two largest shapes controlled the difficulty (established in 5 levels). There were 4 conditions: 2, 3, 4 and 5 alternatives, each one with 120 trials. After that decision, participants had to report their confidence level on a scale from 0% (not confident at all) to 100% (completely confident). Only two alternatives competed for the correct answer, the others (when present) were relatively small and easily discarded.

Methods

As in Experiment 1, the task was programmed in JavaScript using the library jQuery and conducted online on a JATOS (Lange et al., 2015) server. Experiment 2 involved more experimental conditions and, thus, more trials. To guarantee participants' attention during the task, the experimenter made a video call and stayed online for the duration of the experiment.

Participants: 18 participants took part in this experiment. The study was approved by the ethical committee of the Institute of Psychological Research (IIPsi - CONICET - Córdoba, Argentina). Participants should read and accept an informed consent prior to the realization of the experiment. All participants reported no psychiatric, illegal drug consumption or neurological history. The tasks' instructions were written on screen prior to the experiments. All participants performed the task on a computer. Each participant received a payment of roughly 4 US dollars.

Experimental design: the design of Experiment 2 was similar to that of Experiment 1 but involving a larger number of conditions: 480 trials including 120 trials of 2, 3, 4 or 5 alternatives. As in Experiment 1, only two alternatives were large enough to compete for the correct answer; the others were much smaller in order to make them easily discarded (irrelevant alternatives). In trials where the irrelevant alternatives were present, they all had the same size, and each one was randomly a square or a circle. These dud alternatives also varied in size (as in Experiment 1). Participants had to report their confidence level in the decision in the same way as in Experiment 1.

Data analysis: no subject was excluded. We followed the exact same procedure as in Experiment 1. We excluded the easiest level of the task from the performance analysis because one category (the one with 5 alternatives) had perfect performance. The dissociation between correct and incorrect trials carried out in Experiment 1 was not possible because there were not enough incorrect trials.

Results

Type 1 RT increased with difficulty ($F_{4, 68}$ = 23.47, p < 0.000001, η_p^2 = 0.58) (Fig. 7a). This effect was not found in type 2 RT ($F_{4, 68}$ = 1.59, p = 0.18, η_p^2 = 0.08) (Fig. 7b). The weak alternatives did not impact the RT on either of the tasks ($F_{3, 51}$= 0.26, p = 0.84, η_p^2 = 0.01 - type 1 task - Fig. 7a; $F_{3, 51}$ = 0.57, p = 0.63, η_p^2 = 0.03 - type 2 task - Fig. 7b). Regarding task performance, participants' accuracy decreased as the difficulty increased ($F_{3, 51}$ = 90.45, p < 0.000001, η_p^2 = 0.84) (Fig. 7c). Contrary, the dud alternatives did not affect participants' performance ($F_{3, 51}$ = 2.62, p = 0.06, η_p^2 = 0.13) (Fig. 7c).

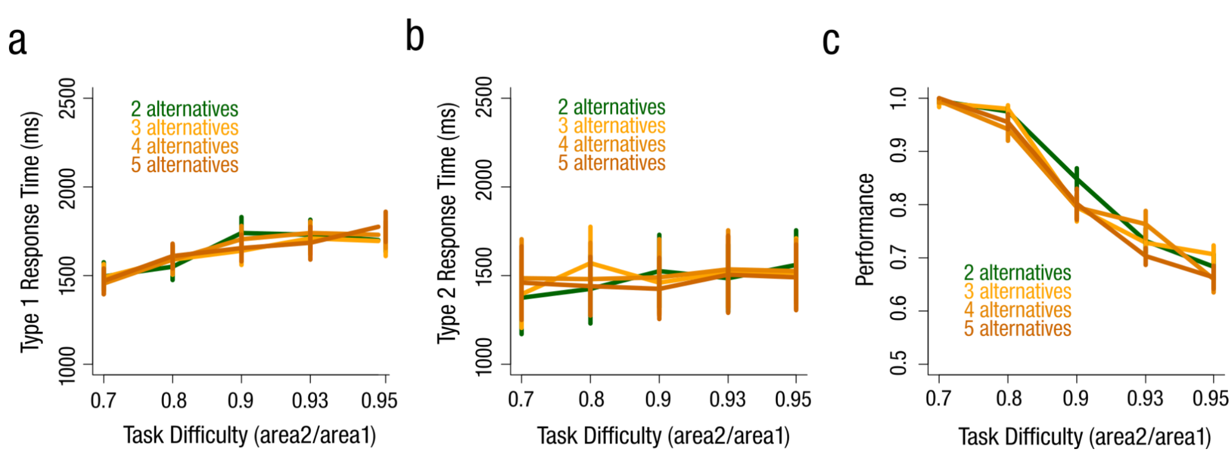

Fig. 7. As in Experiment 1, dud alternatives did not have an effect in RT, neither in (a) type 1 task or in (b) type 2 task. Performance did not vary with the addition of dud alternatives (c).

Confidence level decreased with task difficulty ($F_{4, 68}$ = 54.63, p < 0.000001, η_p^2 = 0.76) (Fig. 8a). Replicating the main finding in Experiment 1 confidence increased with the presence of irrelevant alternatives ($F_{3, 51}$ = 3.81, p = 0.01, η_p^2 = 0.18) (Fig. 8a). Furthermore, confirming the prediction of the Average-residual Model, confidence increased monotonically with the number of duds (Fig. 8a and 8b). We conducted a Tukey HSD post hoc analysis, showing that there was a significant difference between confidence with 2 and 5 alternatives (p = 0.008). To further explore this result, we conducted ANOVAs separately for each remaining conditions (3 and 4 alternatives vs 2 alternatives). We found a significant effect on confidence for the 3-alternative condition ($F_{1, 17}$ = 9.06, p = 0.007, η_p^2 = 0.34) not detected by the post hoc test. No interaction was found on this condition ($F_{4, 68}$ = 0.5, p = 0.72, η_p^2 = 0.03). For the condition with 4 alternatives we found a marginally significant effect ($F_{1, 17}$ = 3.58, p = 0.075, η_p^2 = 0.17) and a significant interaction between the amount of alternatives and the task difficulty ($F_{4, 68}$ = 2.61, p = 0.04, η_p^2 = 0.13). We did not replicate the interaction between difficulty and the irrelevant alternatives found in Experiment 1 ($F_{12, 204}$ = 1.13, p = 0.33, η_p^2 = 0.06) (Fig. 8a). Thirteen out of eighteen participants showed the effect of increased confidence in the dud alternative conditions (Fig. 8c). There was a significant but small effect of the dud alternative size on confidence ($F_{6, 102}$ = 2.49, p = 0.03, η_p^2 = 0.13).

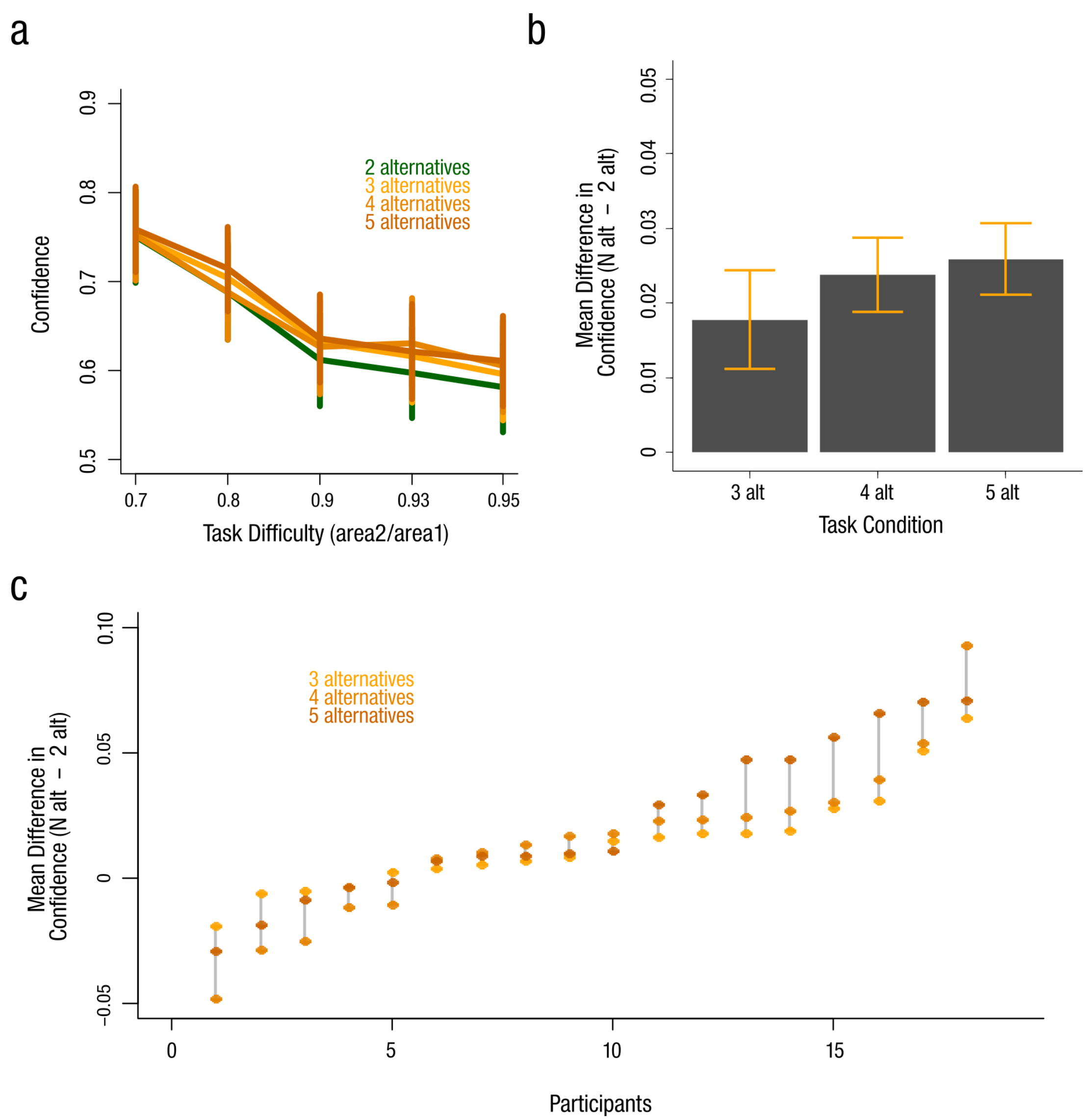

Fig. 8. As in Experiment 1, confidence (a) decreased with task difficulty but increased when weak alternatives were present. Panel (b) depicts the confidence difference between the 3, 4 and 5 alternative conditions and 2-alternative condition considering the highest levels of difficulty (0.9, 0.93 and 0.95). Panel (c) shows the average difference in confidence by condition and participant.

Experiment 3

To evaluate whether the dud-alternative effect replicates in a different multi-alternative task, we adapted the target categorization task from Li & Ma (2020) (Fig. 9).

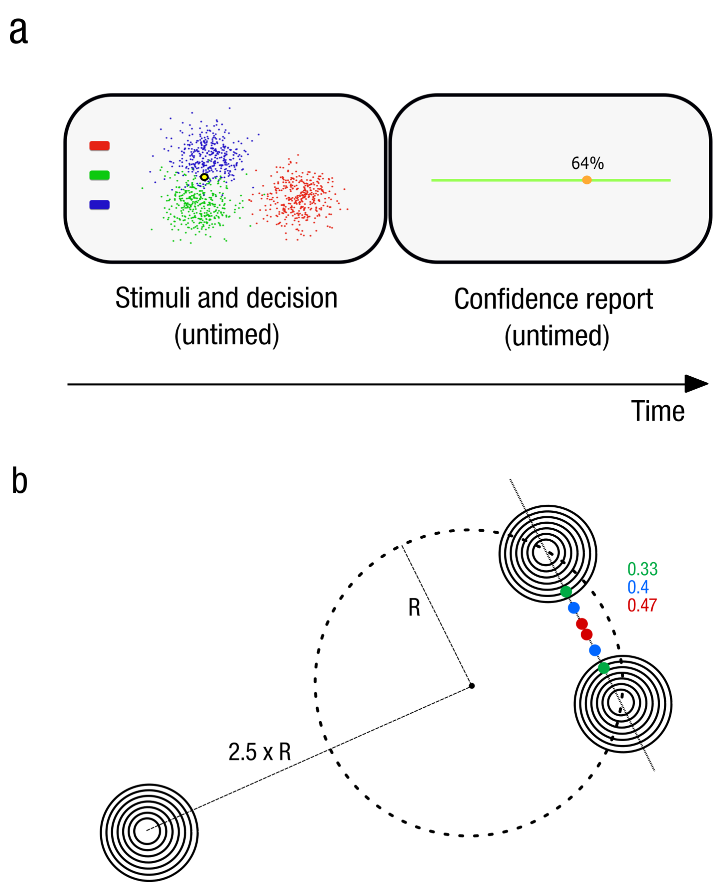

Fig. 9. Target categorization task. Participants (a) decided which cloud of dots the yellow target dot belonged to. Half of the trials had three alternatives and half two. When present, the third cloud was far from the target in order to make it ineligible, following the logic of the previous two studies (red cloud in the figure). After the decision, participants reported their confidence level on a continuous bar. Panel (b) depicts the underlying structure of the task (see Methods).

Methods

As the two previous experiments, the third experiment was conducted online on JATOS (Lange et al., 2015) and programmed in JavaScript using the library jQuery.

Participants: 63 participants took part in this experiment. The study was approved by the ethical committee of the Institute of Psychological Research (IIPsi - CONICET - Córdoba, Argentina). Participants should read and accept an informed consent prior to the realization of the experiment. All participants reported no psychiatric, illegal drug consumption or neurological history. The tasks' instructions were written on screen prior to the experiment. All participants completed the experiment on a computer.

Experimental design: the experiment consisted of a decision making task where participants had to decide to which cloud of dots a target dot belonged to. It had 120 trials, 60 with two alternatives and 60 with three alternatives (i.e., two or three clouds). Dots were normally distributed in space. The mean of the distributions were set in space such that the two main clouds were closer together. More specifically, the mean of these distributions were located in a circle of radius R centered at a fixation point and separated by an angle of 120 degrees. In trials where the third cloud was present, its mean was located at a distance 2.5xR from the fixation point and at an angle of 120 degrees from each of the mean locations of the other two clouds. Each dot cloud had 375 dots. Dot size and standard deviation of the distribution was responsively set, according to the presentation screen. The target dot was yellow with a black border. The target was located in a segment that begins and ends at the mean values of the 2 competing clouds distributions. The position of the target was parameterized by a “task difficulty” variable *d* that takes values from 0 to 0.5, where 0 represents one of the extremes and 0.5 represents the middle, equidistant from the mean values of the distributions. We used *d*={0.33, 0.4, 0.47} where 0.33 is the easiest and 0.47 the most difficult condition. See Fig. 9b for a graphical reference of the task. Stimuli remained on the screen until the participant made a decision, clicking on color coded screen buttons. Participants reported their confidence level on the decision on a continuous scale, as in Experiments 1 and 2.

Data analysis: we excluded 7 participants from the analysis (3 due to low performance -our accuracy cutoff point was an accuracy of 0.6-, and 4 due to lack of variability in the confidence report (confidence level at 100% in > 90% of the trials). The final sample consisted of 52 participants (including all participants does not modify the results). We excluded trials with RT lower than 200ms both for type 1 and type 2 tasks. We excluded trials with a response time greater than 20 s instead of 10 s as in our previous tasks since RTs are larger in this task. We conducted repeated measures ANOVAs and Tukey HSD post hoc tests (both with the Statistica software) to estimate the impact of the difficulty of the task and the dud alternatives in RT, performance and confidence. Our predefined p-value for statistical significance was 0.05.

Results

Task difficulty had a significant effect on type 1 RT ($F_{2, 102} = 34.40$, $p < 0.000001$, $\eta_p^2 = 0.40$) (Fig. 11a). RT were also larger for the 3-alternative condition ($F_{1, 51} = 14.38$, $p = 0.0004$, $\eta_p^2 = 0.22$) (Fig. 10a). No interaction was found between the number of alternatives and task difficulty ($F_{2, 102} = 0.26$, $p = 0.77$, $\eta_p^2 = 0.005$) (Fig. 10a). Task difficulty affected type 2 RT ($F_{2, 102} = 3.63$, $p = 0.03$, $\eta_p^2 = 0.07$). The number of alternatives did not impact type 2 RT ($F_{1, 51} = 3.20$, $p = 0.08$, $\eta_p^2 = 0.059$). No interaction was found between task difficulty and the number of alternatives ($F_{2, 102} = 1.01$, $p = 0.37$, $\eta_p^2 = 0.02$).

Performance, as expected, decreased as the task became more difficult ($F_{2,\ 102}$ = 418.85, p < 0.000001, η_p^2 = 0.89) (Fig. 10c). The presence of a third alternative did not have an effect on performance ($F_{1,\ 51}$ = 0.78, p = 0.38, η_p^2 = 0.015) (Fig. 10c).

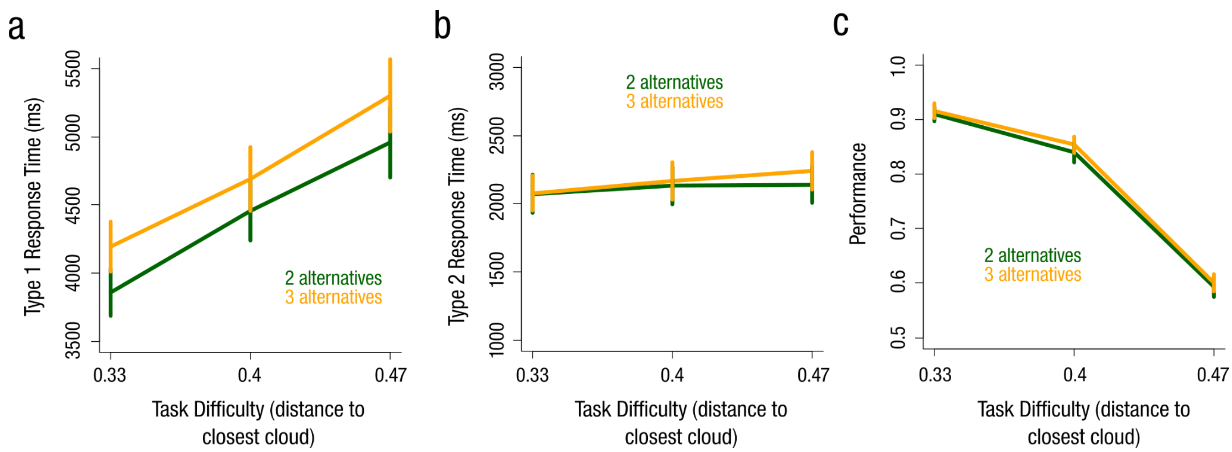

Fig. 10. Both task difficulty and the presence of the dud alternative increased response times in (a) the type 1 task. Contrary to type 1 task, (b) response times for the confidence report were not affected neither by task difficulty nor the amount of alternatives. Performance decreased as the task difficulty increased, whereas the dud alternative did not affect it (c).

Confidence, as expected and in line with performance, decreased when the task got more difficult ($F_{2,\ 102}$ = 97.59, p < 0.000001, η_p^2 = 0.66) (Fig. 11a). In contrast with the results from Experiments 1 and 2, we did not find evidence for higher confidence in the 3-alternative condition ($F_{1,\ 51}$ = 0.485, p = 0.49, η_p^2 = 0.009) (Fig. 11b) and there was not an interaction between the number of alternatives and task difficulty ($F_{2,\ 102}$ = 1.84, p < 0.16, η_p^2 = 0.035).

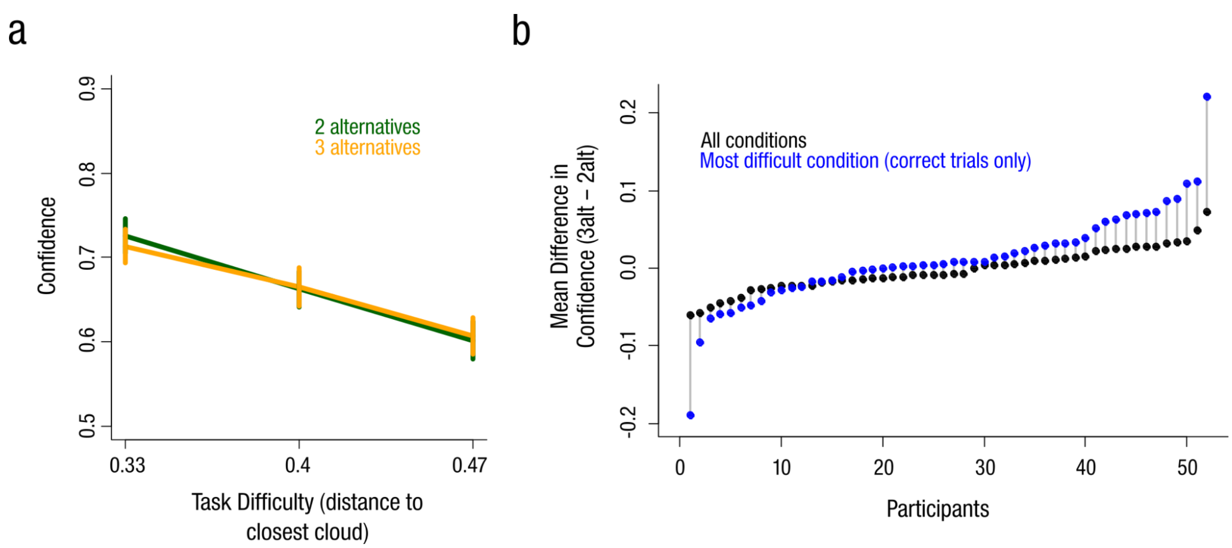

Fig. 11. Effect of number of alternatives on confidence. Confidence (a) diminished with task difficulty

and is not affected by the presence of an irrelevant alternative. Panel (b) shows the confidence difference between 3 and 2 alternative conditions, per participant.

We explored the impact of the presence of the third alternative on confidence level separately for correct and incorrect responses, as in Experiment 1 (Fig. 12). For correct trials only, confidence was not affected by the number of alternatives ($F_{1,\ 51}$ = 0.03, p = 0.86, η_p^2 = 0.0006), but we found a marginally significant interaction between task difficulty and the number of alternatives ($F_{2,\ 102}$ = 2.61, p = 0.078, η_p^2 = 0.05), pointing to a possibly present effect of the third alternative in confidence at high task difficulty (Fig. 12b). To explore this further, we repeated the analysis but only considering trials with RTs lower than 10 s (as in Experiments 1 and 2). We found a significant interaction ($F_{2,\ 102}$ = 3.49, p = 0.003, η_p^2 = 0.06). For incorrect trials, we did not found an effect of the amount of alternatives ($F_{1,\ 25}$ = 0.045, p = 0.83, η_p^2 = 0.002) or an interaction between task difficulty and the number of alternatives ($F_{2,\ 50}$ = 1.40, p = 0.26, η_p^2 = 0.05) (Fig. 12c).

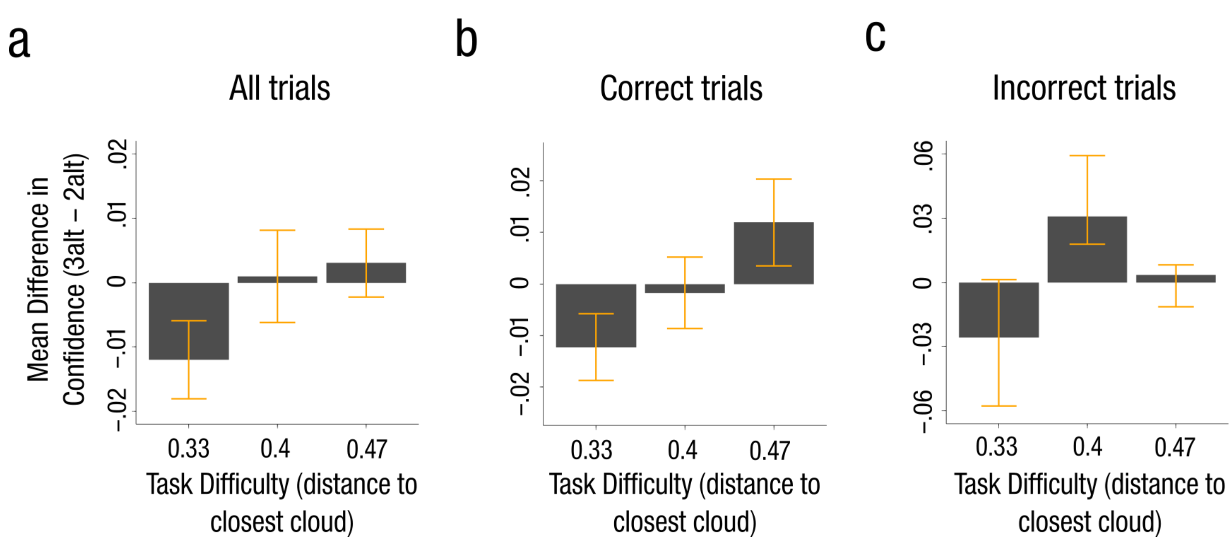

Fig. 12. The impact of the third alternative on confidence dissociated by (b) correct and (c) incorrect responses. Considering (a) all trials, the third alternative did not increase confidence at any level of difficulty. For (b) correct trials, an interaction between task difficulty and the amount of alternatives was found. For (c) incorrect trials, confidence only increased at the middle level of difficulty.

Discussion

In the present study, we investigated whether irrelevant alternatives increase confidence in a perceptual decision. The underlying motivation was that this kind of effect was reported in likelihood judgements (Windschitl & Chambers, 2004), line-up identifications

(Charman et al., 2011) and associative memory (Hanczakowski et al., 2014) but never tested in the present context. Furthermore, most accepted computational models of confidence in perceptual decisions (H.-H. Li & Ma, 2020; Meyniel et al., 2015; Sanders et al., 2016) state that the presence of irrelevant alternatives should not modify it or, in any case, should decrease it. Contrary, our main finding is that these alternatives can boost confidence in perceptual decisions, a result that seems at odds with normative models of confidence.

These weak alternatives, however, do not impact RTs (Experiments 1 and 2) or performance. These null effects are important because they discard that changes in performance or overconfidence usually present in difficult tasks (Baranski & Petrusic, 1994; Gigerenzer et al., 1991) are raising confidence. The lack of effect in performance also implies that divisive normalization is not at play in this task. In agreement with Gluth et al. (2020) —and extending this negative result to a perceptual decision making task—, we did not found supportive evidence for a distractor effect in our task (Louie et al., 2013).

Our results from Experiment 1 and 2 cannot be explained by current views of the field proposing that confidence arises directly from the probability of being correct (Meyniel et al., 2015; Sanders et al., 2016) or direct comparisons between the more probable stimulus and the second one (H.-H. Li & Ma, 2020). This depicts why considering the effect of irrelevant alternatives (Charman et al., 2011; Hanczakowski et al., 2014; Windschitl & Chambers, 2004) in models of confidence is needed. To account for the effect in non-perceptual tasks, Charman et al. (2011), Windschitl & Chambers (2004) and Hanczakowski et al. (2014) proposed various theoretical explanations. Interestingly, the preferred one was the Contrast model, that we computationally implemented: confidence level is obtained by a series of pairwise comparisons between the chosen option and the rest. Nevertheless, our computational modeling results suggest that, although the Contrast model captures the effect of increased confidence levels when a dud-alternative is present, only the Average-residual model replicates the pattern of our data. Specifically, the Average-residual model states that people construct their confidence by first judging the support offered by the evidence for chosen option and the support offered by the average of the rest of the alternatives and then taking the difference between them. In this model having dud alternatives decreases the value of the competing options, inflating confidence. Furthermore, the model predicts that confidence should increase with the number of dud alternatives, which the results of Experiment 2 confirmed.

Prior research suggests that maintaining mental representations for all alternatives is costly (Maniscalco et al., 2016). In that scenario, confidence only relies on the evidence in favor of the chosen option, also known as “positive evidence bias” (Koizumi et al., 2015; Maniscalco et al., 2016; Peters et al., 2017; Zylberberg et al., 2012). However, unchosen options can affect confidence (H.-H. Li & Ma, 2020). Our results and model are in line with this posture, since even very weak alternatives are taken into account to make confidence judgments. Moreover, our findings also deviate from the notion that, in multi-alternative perceptual decisions, individuals making decisions only have access to the level of activity of the alternative with highest internal activity (Yeon & Rahnev, 2020). Indeed, the Average-residual model necessarily implies that the level of activity of all alternatives is accessible for the observer in order to compute confidence.

The picture, however, seems more complicated. Our results from Experiment 3 show that, in that particular task, the dud alternative barely affects confidence. If it does, the effect is restricted to trials with correct responses. This constitutes a main challenge for the Average-residual model. We can only speculate why this experiment bears no effect. On one hand, the slower reaction times in the 3-alternative condition may indicate that observers were more uncertain and needed more time to make their decision; this could counteract the dud-alternative effect. On the other hand, one key element could be presentation time (restricted to 1 second in Experiments 1 and 2, and untimed in Experiment 3). In perceptual tasks with brief stimuli presentations, confidence computations rely more on an internal representation (Rahnev, 2020). Thus, mechanisms acting in Experiments 1 and 2 may differ from the ones acting in Experiment 3. In fact, the dud alternative effect is more prominent in timed tasks (Windschitl & Chambers, 2004).

In this line, an alternative explanation for our results might be that the dud alternative effect depends on a form of variance misperception (Zylberberg et al., 2014), an overconfidence bias found in perceptual tasks when signal is low and perceptual noise is added (M. K. Li et al., 2018; Rahnev et al., 2011; Solovey et al., 2015). The presence of dud alternatives might add perceptual noise, causing the participants to have higher levels of confidence if they do not adjust their type 2 criteria accordingly. Moreover, recently reported evidence in value-based decisions indicate that irrelevant alternatives remain at an attentional periphery (Gluth et al., 2020), where there are effects of variance misperception and overconfidence (Winter & Peters, 2021). Larger stimuli presentations would also rule out possible variance misperception effects, due to a higher signal to noise ratio. Nevertheless, why the dud effect was not present in Experiment 3 is still unclear, since it was already reported in non-perceptual tasks (Charman et al., 2011; Hanczakowski et al., 2014; Windschitl & Chambers, 2004). Future research should address why confidence is better explained by the Diff model in the Experiment 3 context.

In conclusion, our results imply that —at least in some contexts— confidence in multi-alternative decisions deviates from the traditional Bayesian confidence hypothesis and from the Difference model, recently developed for decisions with multiple alternatives. Moreover, our findings suggest that unchosen options affect confidence. Future computational models should consider this effect to more accurately explain human confidence levels in perceptual decision-making.

Author contributions

N. A. Comay, M. Sigman, G. Solovey and P. Barttfeld developed the study concept and design. G. Della Bella and P. Lamberti helped with online implementation of the task. Data collection was performed by N.A. Comay. N.A. Comay, G. Solovey and P. Barttfeld performed the data analysis and model simulations. N. A. Comay, G. Solovey, M. Sigman and P. Barttfeld drafted the manuscript. All authors approved the final version of the manuscript for submission.

Acknowledgements

This research was supported by a grant from Agencia Nacional de Promoción Científica y Tecnológica (Argentina) PICT (Grant #2018-03614). Nicolás A. Comay and Gabriel Della Bella have a scholarship from the National Scientific and Technical Research Council (CONICET - Argentina).

Data and code availability

All data and materials will be stored in https://github.com/nicolascomay/confidence_dud. This includes the source code of the 3 experiments, the .jzip archives for uploading the experiments to JATOS and R code for model fitting and replication of main figures .

References

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition*, *146*, 377-386. https://doi.org/10.1016/j.cognition.2015.10.006

Ariely, D., & Wallsten, T. S. (1995). Seeking Subjective Dominance in Multidimensional Space: An Explanation of the Asymmetric Dominance Effect. *Organizational*

Behavior and Human Decision Processes, *63*(3), 223-232. https://doi.org/10.1006/obhd.1995.1075

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*(4), 412-428. https://doi.org/10.3758/BF03205299

Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identifications. *Law and Human Behavior*, *35*(6), 479-500. https://doi.org/10.1007/s10979-010-9261-1

Drugowitsch, J., Moreno-Bote, R., & Pouget, A. (2014). Relation between Belief and Performance in Perceptual Decision Making. *PLOS ONE*, *9*(5), e96511. https://doi.org/10.1371/journal.pone.0096511

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. https://doi.org/10.1017/CBO9781316272503

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00443

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843-876. https://doi.org/10.3758/BF03196546

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506-528. https://doi.org/10.1037/0033-295X.98.4.506

Gluth, S., Kern, N., Kortmann, M., & Vitali, C. L. (2020). Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nature Human Behaviour*, *4*(6), 634-645. https://doi.org/10.1038/s41562-020-0822-0

Grimaldi, P., Lau, H., & Basso, M. A. (2015). There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making. *Neuroscience & Biobehavioral Reviews*, *55*, 88-97. https://doi.org/10.1016/j.neubiorev.2015.04.006

Hanczakowski, M., Zawadzka, K., & Higham, P. A. (2014). The dud-alternative effect in memory for associations: Putting confidence into local context. *Psychonomic Bulletin & Review*, *21*(2), 543-548. https://doi.org/10.3758/s13423-013-0497-x

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*(2), 181-197. https://doi.org/10.1017/S0952523800009640

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322-1337. https://doi.org/10.1098/rstb.2012.0037

Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception, & Psychophysics*, *77*(4), 1295-1306. https://doi.org/10.3758/s13414-015-0843-3

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PLOS ONE*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Li, H.-H., & Ma, W. J. (2020). Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nature Communications*, *11*(1), 2004. https://doi.org/10.1038/s41467-020-15581-6

Li, M. K., Lau, H., & Odegaard, B. (2018). An investigation of detection biases in the unattended periphery during simulated driving. *Attention, Perception, & Psychophysics*, *80*(6), 1325-1332. https://doi.org/10.3758/s13414-018-1554-3

Louie, K., Grattan, L. E., & Glimcher, P. W. (2011). Reward Value-Based Gain Control: Divisive Normalization in Parietal Cortex. *Journal of Neuroscience*, *31*(29), 10627-10639. https://doi.org/10.1523/JNEUROSCI.1237-11.2011

Louie, K., Khaw, M. W., & Glimcher, P. W. (2013). Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National Academy of Sciences*, *110*(15), 6139-6144. https://doi.org/10.1073/pnas.1217854110

Luce, R. D. (1979). *Individual choice behavior: A theoretical analysis*. Greenwood Press.

Ma, W. J. (2019). Bayesian Decision Models: A Primer. *Neuron*, *104*(1), 164-175. https://doi.org/10.1016/j.neuron.2019.09.037

Mamassian, P. (2016). Visual Confidence. *Annual Review of Vision Science*, *2*(1), 459-481. https://doi.org/10.1146/annurev-vision-111815-114630

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422-430. https://doi.org/10.1016/j.concog.2011.09.021

Maniscalco, B., Peters, M. A. K., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity.

Attention, Perception, & Psychophysics, *78*(3), 923-937. https://doi.org/10.3758/s13414-016-1059-x

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian Probability: From Neural Origins to Behavior. *Neuron*, *88*(1), 78-92. https://doi.org/10.1016/j.neuron.2015.09.039

Niwa, M., & Ditterich, J. (2008). Perceptual Decisions between Multiple Directions of Visual Motion. *Journal of Neuroscience*, *28*(17), 4435-4445. https://doi.org/10.1523/JNEUROSCI.5564-07.2008

Northoff, G., & Mushiake, H. (2020). Why context matters? Divisive normalization and canonical microcircuits in psychiatric disorders. *Neuroscience Research*, *156*, 130-140. https://doi.org/10.1016/j.neures.2019.10.002

Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., Halgren, E., & Lau, H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, *1*(7), 0139. https://doi.org/10.1038/s41562-017-0139

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864-901. https://doi.org/10.1037/a0019737

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366-374. https://doi.org/10.1038/nn.4240

Rahnev, D., Balsdon, T., Charles, L., de Gardelle, V., Denison, R. N., Desender, K., Faivre, N., Filevich, E., Fleming, S. M., Jehee, J., Lau, H., Lee, A. L. F., Locke, S.

M., Mamassian, P., Odegaard, B., Peters, M. A. K., Reyes, G., Rouault, M., Sackur, J., … Zylberberg, A. (2021). Consensus goals for the field of visual metacognition [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/z8v5x

Rahnev, D. (2020). Confidence in the Real World. *Trends in Cognitive Sciences*, *24*(8), 590-591. https://doi.org/10.1016/j.tics.2020.05.005

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*, e223. https://doi.org/10.1017/S0140525X18000936

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., de Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, *14*(12), 1513-1515. https://doi.org/10.1038/nn.2948

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review*, *120*(3), 697-719. https://doi.org/10.1037/a0033152

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron*, *90*(3), 499-506. https://doi.org/10.1016/j.neuron.2016.03.025

Solovey, G., Graney, G. G., & Lau, H. (2015). A decisional account of subjective inflation of visual perception at the periphery. *Attention, Perception, & Psychophysics*, *77*(1), 258-271. https://doi.org/10.3758/s13414-014-0769-1

Windschitl, P. D., & Chambers, J. R. (2004). The Dud-Alternative Effect in Likelihood Judgment. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *30*(1), 198-215. https://doi.org/10.1037/0278-7393.30.1.198

Winter, C. J., & Peters, M. A. K. (2021). *Variance misperception under skewed empirical noise statistics explains overconfidence in the visual periphery* [Preprint]. Neuroscience. https://doi.org/10.1101/2021.01.08.425966

Yeon, J., & Rahnev, D. (2020). The suboptimality of perceptual decision making with multiple alternatives. *Nature Communications*, *11*(1), 3857. https://doi.org/10.1038/s41467-020-17661-z

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*. https://doi.org/10.3389/fnint.2012.00079

Zylberberg, A., Roelfsema, P. R., & Sigman, M. (2014). Variance misperception explains illusions of confidence in simple perceptual decisions. *Consciousness and Cognition*, *27*, 246-253. https://doi.org/10.1016/j.concog.2014.05.012