



HAL
open science

Confidence can be automatically integrated across two visual decisions

David Aguilar-Lleyda, Mahiko Konishi, Jérôme Sackur, Vincent de Gardelle

► **To cite this version:**

David Aguilar-Lleyda, Mahiko Konishi, Jérôme Sackur, Vincent de Gardelle. Confidence can be automatically integrated across two visual decisions. *Journal of Experimental Psychology: Human Perception and Performance*, 2021, 47 (2), pp.161-171. 10.1037/xhp0000884 . hal-03184996

HAL Id: hal-03184996

<https://hal.science/hal-03184996>

Submitted on 20 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 Confidence can be automatically 2 integrated across two visual decisions

3 David Aguilar-Lleyda ^{1*}, Mahiko Konishi ^{2,3}, Jérôme Sackur ^{2,4}, Vincent
4 de Gardelle ⁵

5 ¹ Centre d'Économie de la Sorbonne (CNRS & Université Paris 1 Panthéon-Sorbonne), 112
6 Boulevard de l'Hôpital, 75013 Paris, France

7 ² Laboratoire de Sciences Cognitives et Psycholinguistique (LSCP), Département d'Études
8 Cognitives de l'École Normale Supérieure, Centre National de la Recherche Scientifique, École
9 des Hautes Études en Sciences Sociales, Paris Sciences et Lettres Research University, Paris,
10 France

11 ³ ONERA, The French Aerospace Lab, Information Processing and Systems Department,
12 13661, Salon Cedex Air, France

13 ⁴ Laboratoire Interdisciplinaire de l'X, École Polytechnique, Palaiseau, France

14 ⁵ Paris School of Economics & CNRS

15
16 *corresponding author:

17 email: aguilarlleyda@gmail.com

18 address: Maison des Sciences Économiques, 112 Boulevard de l'Hôpital, 75013 Paris, France

19

20 Abstract:

21 Humans can estimate their confidence in making correct decisions, but these confidence
22 judgments are biased by their other estimations, an effect known as confidence leak. However,
23 it remains unclear whether this effect arises automatically. Here, we address this issue by
24 having participants make two visual decisions and give confidence ratings for one or for both
25 decisions within each trial. Using the well-known interaction between task difficulty and
26 response accuracy as a proxy for confidence, we found that confidence ratings for one decision
27 were greater when the other decision was also associated with greater confidence, even when
28 the latter was not explicitly rated. For one of the two tasks, this confidence leak also occurred
29 when participants knew in advance that no confidence rating would be required for the other
30 task. Our results support the idea that confidence is not only automatically computed but also
31 automatically integrated across decisions.

32

33 Keywords: confidence, metacognition, decision-making, vision, perception

1 Introduction

2 When we make a choice, we can often assess our confidence about it. This ability of evaluating
3 our own decisions can serve many purposes. For instance, it can be used to guide learning in
4 the absence of feedback (Guggenmos et al. 2016; Daniel and Pollmann 2012; Hainguerlot,
5 Vergnaud, and de Gardelle 2018), to regulate information accumulation (van den Berg et al.
6 2016; Desender, Boldt, and Yeung 2018), or to compare different tasks (de Gardelle and
7 Mamassian 2014; de Gardelle, Le Corre, and Mamassian 2016).

8
9 It has been long known that confidence can reflect the accuracy of the decision (Peirce and
10 Jastrow 1884; Dallenbach 1913). However, confidence ratings are also affected by variables
11 other than accuracy. Among them, confidence about the current decision may be attracted
12 towards confidence ratings expressed for other decisions (Rahnev et al. 2015; Kantner et al.
13 2019). This effect, known as confidence leak, would be consistent with observers assuming that
14 the quality of their perceptual evidence is relatively stable in time, as usual in natural situations.
15 Observers would then exploit this regularity when evaluating their performance.

16
17 However appealing this general view is, its support only comes from experimental situations
18 where both decisions required an explicit evaluation of confidence. In other words, it is not clear
19 whether participants' confidence judgments rely on other decisions for which no explicit
20 evaluation of confidence was made. If confidence leak depends on confidence being explicitly
21 stated, then it would only matter in a handful of laboratory situations, excluding those situations
22 where - as usual in life - confidence is not explicitly mapped onto a scale. Although there is
23 evidence for confidence being computed automatically (Lebreton et al., 2015), this automatic
24 evaluation could be an epiphenomenon without consequence. Our goal here is to evaluate
25 whether this automatic computation of confidence also sets the context of further metacognitive
26 evaluations. We hypothesize that our evaluation of a decision can be influenced by our
27 confidence in another decision, even when the latter was not explicitly expressed.

28
29 One methodological difficulty to test our hypothesis is to experimentally estimate an unreported
30 confidence judgment. Here, we relied on the previously demonstrated interaction between
31 response accuracy and task difficulty, by which confidence should both increase for correct
32 responses, and decrease for errors, as a task becomes easier (Kepecs et al. 2008; Sanders,
33 Hangya, and Kepecs 2016). Specifically, in a dual-task paradigm, we evaluated whether
34 confidence ratings for one task would be affected by the interaction between response accuracy
35 and task difficulty for the other task. If it was, we could conclude the existence of a confidence
36 leak from that other task to the task being rated.

37
38 We report three experiments in which we evaluate confidence leak in different settings. In
39 Experiment 1, on each trial participants saw a perceptual stimulus for which they had to make
40 two decisions, each followed by a confidence rating. Our goal was to replicate the original
41 confidence leak finding, to provide a baseline against which to compare the other experiments.
42 In Experiment 2, confidence was only required for one of the two tasks, randomly chosen on
43 each trial. We evaluated the confidence leak from an unreported confidence, using the accuracy

1 x difficulty interaction as our proxy for this unreported confidence. In Experiment 3, we also told
2 participants in advance which task would require a rating (this was kept constant within a block
3 of trials as well) to eliminate any reason for participants to compute confidence for the non-rated
4 task. Finding a confidence leak in this situation would provide clear evidence for the automatic
5 nature of confidence integration across decisions.

6 Method

7 Participants

8 One-hundred and one healthy adults took part in three experiments (Experiment 1 = 34,
9 Experiment 2 = 39, Experiment 3 = 28). They were recruited from the Laboratoire d'Économie
10 Experimentale de Paris volunteer database. They reported normal or corrected-to-normal vision
11 and gave written consent. They were naïve of the objective of the study. Each participant only
12 took part in one experiment. Final payoff depended on performance and confidence (see
13 supplementary material), with an average payoff of 16€. The study was approved by the Paris
14 School of Economics ethics committee. Experiments were run in two sessions of up to 20
15 participants each. Final sample sizes depended on how many people attended the sessions.

16 Apparatus

17 Experiments were powered by MATLAB Psychtoolbox (Brainard 1997). Stimuli, appearing on a
18 grey background, were viewed approximately 60cm away from the screen (17", 1024x768 pixel
19 resolution, 60 Hz refresh rate).

20 Perceptual task

21 Trials began with a 500 ms fixation cross, followed by a 300 ms blank period, and then the
22 stimulus for 1000 ms. The stimulus, adapted from Rahnev et al. (2015), was an array of 80
23 randomly positioned letters, each letter being an O or an X, colored blue or orange (see Figure
24 1). Each letter, in Arial font, occupied approximately 0.5° of visual angle. Elements were
25 presented within a 10° wide imaginary square. After stimulus offset, participants reported
26 whether there were more blue or orange elements (the color task), and more Os or Xs (the letter
27 task). Across trials, the order of the two tasks was random, and so was the dominant element
28 for each task. The response screen for each task presented two horizontally aligned boxes
29 featuring an X and an O (letter task) or an orange and a blue square (color task), randomly
30 allocated to the right or left box. Participants pressed the 'E' or 'R' key of the keyboard to select
31 the left or right box (see Figure 1).
32

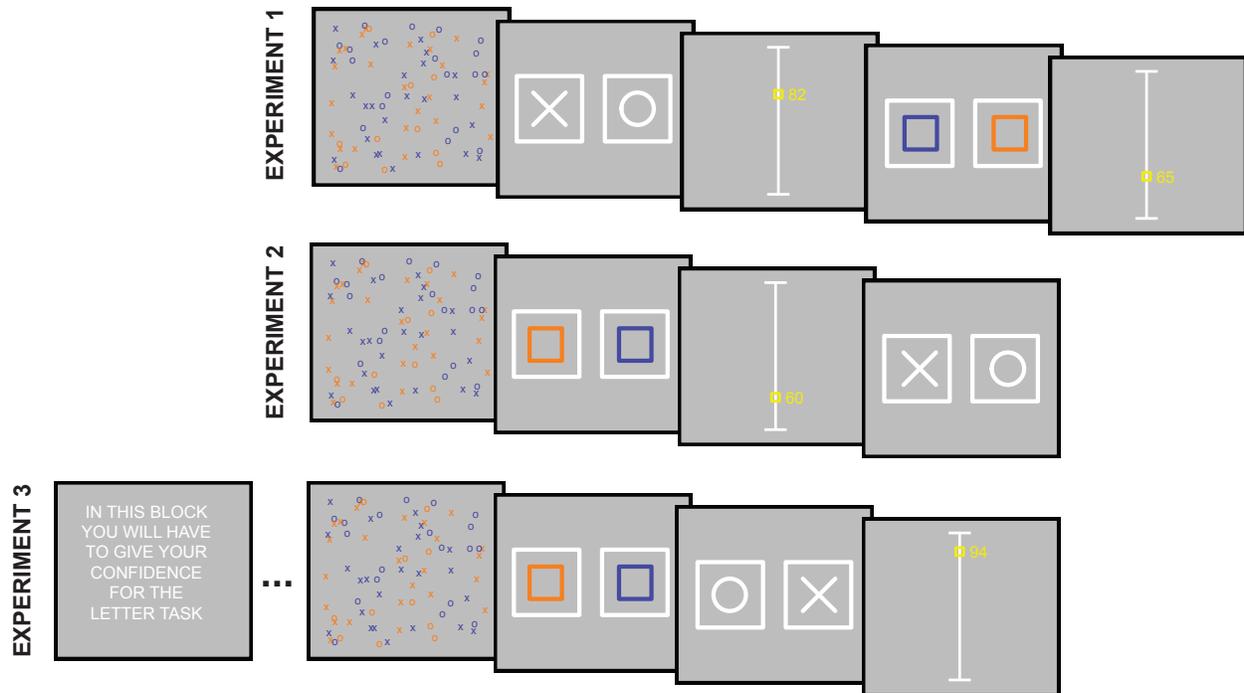


Figure 1. For all experiments, overview of a trial. Participants were presented with an array of Os and Xs, colored blue or orange, and indicated the predominant color and letter in the stimulus, by selecting the corresponding response boxes. The order of the two tasks was random within each trial. We required a confidence rating for both tasks within a trial in Experiment 1, for a task selected randomly on each trial in Experiment 2, and for a task announced to participants at the beginning of the block in Experiment 3.

Participants completed 4 parts of 4 blocks each, each block having 24 trials. A 15-second rest separated the blocks. After each part participants were given a self-timed rest, and informed of their accuracy (percentage of correct choices, pooling across both tasks) in the last part.

Confidence ratings for the different experiments

In Experiment 1, participants rated their confidence immediately after each choice. A white vertical confidence scale appeared together with a randomly positioned yellow cursor. The cursor was accompanied by an integer indicating the confidence rating, from 50 at the bottom to 100 at the top of the scale. Participants moved the cursor with their mouse, and clicked to select the desired rating. Participants were instructed that a rating of 50 would reflect total uncertainty (i.e. random choice), while 100 would reflect total certainty. After reporting confidence, the response screen for the second task appeared. After the second response, the confidence scale was presented for this second task.

In Experiment 2, confidence was asked for only one task within each trial, either the color or the letter task (counterbalanced across trials). Participants only knew which task required a

1 confidence rating when the confidence scale was presented. For the other task, no confidence
2 scale was presented.

3
4 In Experiment 3, participants gave only one confidence rating per trial, but the task to be rated
5 was known in advance: it was announced at the beginning of each block and kept constant
6 during the block. The rated task was also counterbalanced across blocks.

7 Staircase and difficulty manipulation

8 On each trial, the proportion of items in the dominant category was controlled so that the task
9 could be easy (90% expected performance) or hard (60% expected performance). Difficulty was
10 controlled independently for each task, leading to a 2 x 2 factorial design. Within a block, each
11 combination was presented equally often. The two difficulty levels were estimated for each
12 participant in an initial psychophysical staircase of 96 trials (see supplementary material).

13 Statistics

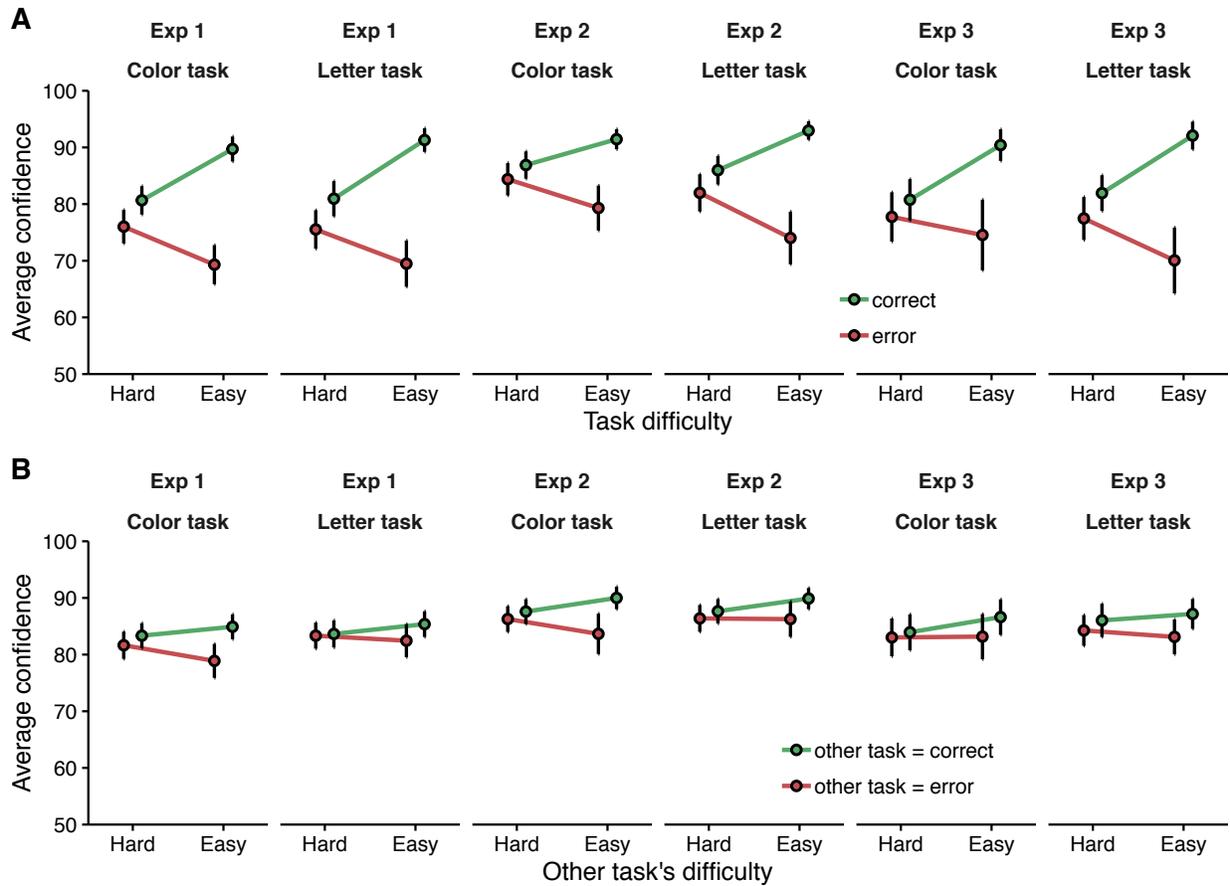
14 We analysed our data with linear mixed models (LMM) using the lme4 (Bates et al. 2015) and
15 lmerTest (Kuznetsova, Brockhoff, and Christensen 2017) packages in R, and report the results
16 of ANOVA tests conducted on the fitted models. Using these models allows us to take into
17 account every trial, instead of averaging them and giving the same weight to each level of the
18 factor. In our experiments, this is particularly important when comparing correct trials and errors,
19 which occur in highly different proportions. In our LMMs participants were always treated as a
20 random intercept.

21 Results

22 Previous work claimed the presence of a confidence leak between two tasks by showing a
23 positive correlation between the confidence ratings given for both. We started by trying to
24 replicate these results in Experiment 1, where participants reported their confidence for both
25 tasks within each trial. The correlation between the two confidence ratings across trials was
26 positive for 29 out of 34 participants, with r values of participants being overall significantly
27 larger than 0 ($t(33) = 7.000$, $p < 0.001$, 95% CI [0.132, 0.241], $d = 1.201$). In other words, in
28 Experiment 1 we replicated the original finding of a confidence leak across tasks.

29
30 We also verified the expected relation between accuracy, difficulty and confidence. For each
31 task (color and letter) separately, we fitted a LMM with rated confidence as dependent variable,
32 and response accuracy (correct vs. error) and task difficulty (hard vs easy) in the task as fixed
33 effects. The expected pattern was obtained: for both tasks we found not only significant main
34 effects of response accuracy (color task: $F(13026) = 986.315$, $p < 0.001$; letter task: $F(13022) =$
35 1288.025 , $p < 0.001$) and task difficulty (color task: $F(13022) = 19.814$, $p < 0.001$; letter task:
36 $F(13021) = 34.761$, $p < 0.001$), but also a significant interaction between accuracy and difficulty
37 (color task: $F(13024) = 399.138$, $p < 0.001$; letter task: $F(13022) = 464.092$, $p < 0.001$). When

1 performing this analysis for the remaining experiments, the interaction was always significant
 2 (all $p < 0.001$, see the supplementary material for the detailed information, as well as Figure
 3 2A).
 4



5
 6 *Figure 2. A. For each experiment and task, average confidence ratings across*
 7 *participants, as a function of task difficulty and response accuracy (color-coded). Error*
 8 *bars denote 95% confidence intervals. B. For each experiment and task, average*
 9 *confidence ratings across participants, as a function of task difficulty and response*
 10 *accuracy on the other task (color-coded). Error bars denote 95% confidence intervals.*
 11

12 Crucially, this interaction is the signature of confidence that we will rely on to evaluate
 13 confidence leak. In a nutshell, we reasoned that, if confidence ratings in one task were affected
 14 by confidence in the other task, then these ratings should be affected by the accuracy x difficulty
 15 interaction in the other task. Importantly, using this proxy, we can evaluate a confidence leak
 16 that may occur even when participants do not explicitly rate their confidence for task B.
 17

18 To do so, we built a LMM where the confidence rating for a task was predicted by the response
 19 accuracy and task difficulty of the other task within the same trial, in addition to the response
 20 accuracy and task difficulty of the task at hand. By doing so, we ensure that our measure of
 21 confidence leak is not contaminated by a potential correlation in confidence between the two
 22 tasks, driven by a correlation in their performance. Table 1 shows, for each task, how

1 confidence was affected by the accuracy and difficulty and their interaction for the current task,
 2 as well as for the other task.

3

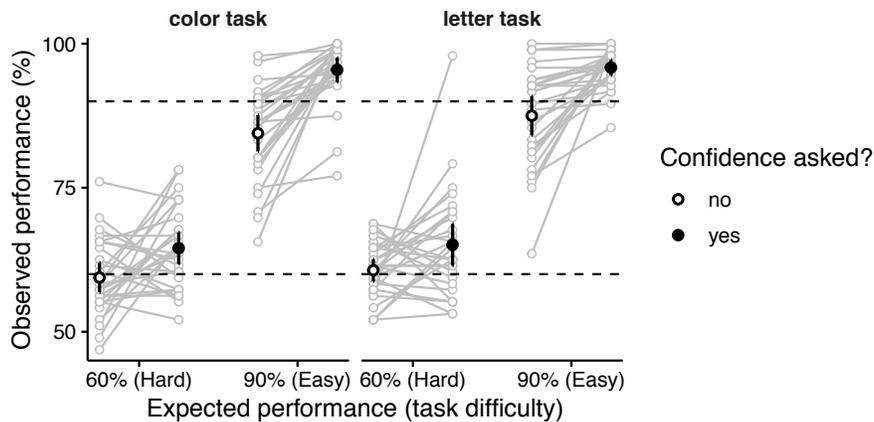
4 Critically, the accuracy x difficulty interaction based on the other task (rightmost columns in
 5 Table 1), was significant for both tasks in Experiment 1, for both tasks in Experiment 2, and for
 6 the color task in Experiment 3. In sum, we found evidence that ratings of confidence for one
 7 task can be influenced by the signature of confidence from another task (Experiment 1), even
 8 when confidence is not explicitly reported for that other task (Experiment 2), and even when
 9 participants know in advance that no confidence rating will be required for this other task
 10 (Experiment 3). Note however that the leak from color to letter was not significant in Experiment
 11 3, and that this difference between the color-to-letter leak and letter-to-color leak in Experiment
 12 3 was statistically significant, when including both tasks in the LMM (triple interaction task type x
 13 other accuracy x other difficulty, $F(10715) = 7.596, p = 0.006$).

14

		effect of task-relevant variables			leak from the other task		
		d.f.	F	p	d.f.	F	p
experiment 1		<u>color -> color</u>			<u>letter -> color</u>		
	Accuracy	13022.156	972.209	< 0.001	13019.511	47.947	< 0.001
	Difficulty	13019.363	21.491	< 0.001	13018.354	0.962	0.327
	Interaction	13020.628	392.766	< 0.001	13019.990	6.763	0.009
		<u>letter -> letter</u>			<u>color -> letter</u>		
	Accuracy	13018.869	1277.811	< 0.001	13021.884	6.644	0.010
Difficulty	13017.920	36.327	< 0.001	13018.748	3.263	0.071	
Interaction	13019.263	457.565	< 0.001	13019.788	6.693	0.010	
experiment 2		<u>color -> color</u>			<u>letter -> color</u>		
	Accuracy	7452.862	259.765	< 0.001	7447.707	78.842	< 0.001
	Difficulty	7446.741	2.789	0.094	7446.095	0.085	0.770
	Interaction	7447.594	106.703	< 0.001	7446.859	37.666	< 0.001
		<u>letter -> letter</u>			<u>color -> letter</u>		
	Accuracy	7449.883	463.810	< 0.001	7452.373	20.936	< 0.001
Difficulty	7447.628	8.720	0.003	7446.821	12.788	< 0.001	
Interaction	7448.781	169.093	< 0.001	7447.551	3.889	0.049	
experiment 3		<u>color -> color</u>			<u>letter -> color</u>		
	Accuracy	5346.829	208.313	< 0.001	5346.323	27.558	< 0.001
	Difficulty	5344.857	32.201	< 0.001	5343.594	3.172	0.075
	Interaction	5345.079	103.870	< 0.001	5344.230	19.743	< 0.001
		<u>letter -> letter</u>			<u>color -> letter</u>		
	Accuracy	5345.749	299.843	< 0.001	5347.009	19.493	< 0.001
Difficulty	5343.561	18.505	< 0.001	5343.148	2.255	0.133	
Interaction	5343.562	138.020	< 0.001	5343.598	1.879	0.171	

15
 16 *Table 1. For each experiment and task, results of the LMMs where confidence for a task*
 17 *is predicted by response accuracy and task difficulty for that task (task-relevant*
 18 *variables) and by response accuracy and task difficulty for the other task (leak from the*
 19 *other task).*

1
2 In Experiment 3, participants may have devoted more cognitive resources to the task for which
3 confidence had to be rated, simply because they were required more information for this task
4 than for the other. To evaluate this possibility, we conducted an ANOVA on performance, with
5 task type, task difficulty and whether confidence had been asked for that task or not, as factors.
6 Performance was affected by task difficulty ($F(1,27) = 1510.331, p < 0.001, \eta_p^2 = 0.982$), but also
7 by whether confidence was asked ($F(1,27) = 23.073, p < 0.001, \eta_p^2 = 0.461$) and by the
8 interaction between these two factors ($F(1,27) = 17.288, p < 0.001, \eta_p^2 = 0.390$). No other main
9 effects or interactions were significant. In short, performance was higher for the rated task, in
10 particular for easier than for hard tasks, as illustrated in Figure 3, suggesting that participants
11 may have dedicated more attention to the task that was associated with a confidence rating. In
12 the case of the letter task, even when it received less attention this task still generated a leak.
13



14
15 *Figure 3. For each task of Experiment 3, observed performance as a function of the*
16 *expected performance, split by whether confidence was asked for that task. Bigger dots*
17 *represent average performance across participants expressed as a percentage, with*
18 *error bars denoting 95% confidence intervals. Smaller dots represent individual*
19 *participants. Top and bottom dashed lines help indicate where 90% and 60% observed*
20 *performance would lay, respectively.*

21 Discussion

22 Previous work suggests that confidence in one decision may leak into confidence in another
23 decision, thus producing a confidence integration across decisions. However, evidence for this
24 leak has only come from experimental situations where both decisions required an explicit
25 judgment of confidence. In the present study, we show that even an unreported confidence can
26 leak into the confidence on another decision. To do so, we rely on a proxy for confidence:
27 specifically, we look at whether the confidence rating for a task depends on the interaction
28 between response accuracy and task difficulty for the other task. We found that confidence in
29 one dimension was influenced by confidence in the other dimension when the latter was
30 reported (Experiment 1), but also when this other confidence was not reported (Experiments 2

1 and 3), and even when participants knew in advance that they would not have to report it
2 (Experiment 3, although only from the letter to the color task). Beyond providing further
3 evidence that confidence is computed automatically (Lebreton et al. 2015), we show that this
4 computation of confidence is automatically integrated across decisions.

5
6 The finding that confidence leak is observed across two different tasks suggests that the
7 representation involved is relatively abstract and task-independent, as argued before (Rahnev
8 et al., 2015). The present study provides more evidence along the same lines, and furthermore
9 shows that this computation of confidence affects our judgments even when it is not made
10 explicitly. Moreover, we show that this influence from the non-rated task happens despite the
11 fact that this task receives less resources, as indicated by a reduced perceptual performance.

12
13 We relied on the interaction between accuracy and difficulty as a proxy for the confidence in the
14 non-rated task. As this interaction is indeed considered a signature of confidence (Kepecs et al.
15 2008; Sanders, Hangya, and Kepecs 2016), our approach ensured that the influence could be
16 attributed to confidence in the non-rated task, and not simply to the difficulty or to the accuracy
17 in the non-rated task. In addition, since our analyses incorporated accuracy and difficulty for the
18 rated task, our results cannot be explained by a correlation of performance between the two
19 tasks (e.g. due to arousal) that would introduce a correlation of confidence between the tasks
20 masquerading as a confidence leak. Not requiring a confidence rating also ensured that the leak
21 was not due to an anchoring effect by which participants would use similar regions within the
22 displayed confidence scale for the two ratings.

23
24 In previous studies (Rahnev et al. 2015; Kantner et al. 2019; Mueller and Weidemann 2008),
25 confidence has also been shown to be auto-correlated across consecutive trials, either within
26 the same task, or between two different tasks. However, as detailed in the supplementary
27 materials, our data indicated no widespread effect of such leak across trials. It may be that, in a
28 dual task paradigm, any effect coming from the previous trial may vanish because of the greater
29 delay between consecutive trials, or because of the influence of the other task within the same
30 trial. As a matter of fact, in the aforementioned studies, confidence leak across consecutive
31 trials has only been found when only one task was presented per trial.

32
33 One obvious question that remains is why in Experiment 3 confidence for the letter task was
34 clearly integrated into confidence for the color task, but not vice versa. A possible answer lies in
35 how each task may be solved. We can speculate that a decision for the color task is reached at
36 a glance, without having to identify whether each individual element is blue or orange, whereas
37 the letter task requires a more active engagement in terms of visual search. This greater
38 deliberation or effort could have made confidence for the letter task more salient, and thus more
39 likely to leak into color confidence, but also more impervious from influence by the color task.
40 While the present data does not allow us to test these hypotheses, further studies could clarify
41 the cognitive processes engaged during the two tasks, and could evaluate how participants form
42 confidence on these tasks (Reyes and Sackur 2014). One exciting question for future research
43 is whether the saliency of confidence computation, and thus the strength of confidence leak,
44 would depend on the automaticity of the tasks themselves. This suggested link between

1 automaticity and availability for the monitoring system echoes theoretical proposals made in the
2 context of metacognition and consciousness studies (Cleeremans, 2006).

3 References

- 4 Bates, D., M. Maechler, B. Bolker, and S. Walker. 2015. "Fitting Linear Mixed-Effects Models
5 Using lme4." *Journal of Statistical Software*.
- 6 Berg, Ronald van den, Ariel Zylberberg, Roozbeh Kiani, Michael N. Shadlen, and Daniel M.
7 Wolpert. 2016. "Confidence Is the Bridge between Multi-Stage Decisions." *Current Biology*:
8 *CB* 26 (23): 3157–68.
- 9 Brainard, D. H. 1997. "The Psychophysics Toolbox." *Spatial Vision* 10 (4): 433–36.
- 10 Cleeremans, A. (2006). Conscious and unconscious cognition: A graded, dynamic perspective.
11 In Q. Jing, M.R. Rosenzweig, G. d' Ydewalle, H. Zhang, H.-C. Chen & K. Zhang (Eds.),
12 Progress in Psychological Science Around the World Volume I: Neural, Cognitive, and
13 Developmental Issues (Proceedings of the 28th International Congress of Psychology),
14 Hove, UK: Psychology Press, pp. 401- 418.
- 15 Dallenbach, Karl M. 1913. "The Relation of Memory Error to Time Interval." *Psychological*
16 *Review* 20 (4): 323.
- 17 Daniel, Reka, and Stefan Pollmann. 2012. "Striatal Activations Signal Prediction Errors on
18 Confidence in the Absence of External Feedback." *NeuroImage* 59 (4): 3457–67.
- 19 Desender, Kobe, Annika Boldt, and Nick Yeung. 2018. "Subjective Confidence Predicts
20 Information Seeking in Decision Making." *Psychological Science* 29 (5): 761–78.
- 21 Gardelle, Vincent de, François Le Corre, and Pascal Mamassian. 2016. "Confidence as a
22 Common Currency between Vision and Audition." *PloS One* 11 (1): e0147901.
- 23 Gardelle, Vincent de, and Pascal Mamassian. 2014. "Does Confidence Use a Common
24 Currency across Two Visual Tasks?" *Psychological Science* 25 (6): 1286–88.
- 25 Guggenmos, Matthias, Gregor Wilbertz, Martin N. Hebart, and Philipp Sterzer. 2016.
26 "Mesolimbic Confidence Signals Guide Perceptual Learning in the Absence of External
27 Feedback." *eLife* 5 (March). <https://doi.org/10.7554/eLife.13388>.
- 28 Hainguerlot, Marine, Jean-Christophe Vergnaud, and Vincent de Gardelle. 2018. "Metacognitive
29 Ability Predicts Learning Cue-Stimulus Associations in the Absence of External Feedback."
30 *Scientific Reports* 8 (1): 5602.
- 31 Kantner, Justin, Lisa A. Solinger, David Grybinas, and Ian G. Dobbins. 2019. "Confidence
32 Carryover during Interleaved Memory and Perception Judgments." *Memory & Cognition* 47
33 (2): 195–211.
- 34 Kepecs, Adam, Naoshige Uchida, Hatim A. Zariwala, and Zachary F. Mainen. 2008. "Neural
35 Correlates, Computation and Behavioural Impact of Decision Confidence." *Nature* 455
36 (7210): 227–31.
- 37 Kuznetsova, Alexandra, Per B. Brockhoff, and Rune Haubo Bojesen Christensen. 2017.
38 "lmerTest Package: Tests in Linear Mixed Effects Models." *Journal of Statistical Software*
39 82 (13). [http://orbit.dtu.dk/portal/en/publications/id\(2e6e11c6-6a44-43df-b809-
40 d96b897d9bf2\).html](http://orbit.dtu.dk/portal/en/publications/id(2e6e11c6-6a44-43df-b809-d96b897d9bf2).html).
- 41 Lebreton, Maël, Raphaëlle Abitbol, Jean Daunizeau, and Mathias Pessiglione. 2015. "Automatic
42 Integration of Confidence in the Brain Valuation Signal." *Nature Neuroscience* 18 (8): 1159–
43 67.
- 44 Mueller, Shane T., and Christoph T. Weidemann. 2008. "Decision Noise: An Explanation for
45 Observed Violations of Signal Detection Theory." *Psychonomic Bulletin & Review* 15 (3):
46 465–94.
- 47 Peirce, Charles Sanders, and Joseph Jastrow. 1884. "On Small Differences in Sensation."

1 <https://philarchive.org/archive/PEIOSD>.
2 Rahnev, Dobromir, Ai Koizumi, Li Yan McCurdy, Mark D'Esposito, and Hakwan Lau. 2015.
3 "Confidence Leak in Perceptual Decision Making." *Psychological Science* 26 (11): 1664–
4 80.
5 Reyes, Gabriel, and Jérôme Sackur. 2014. "Introspection during Visual Search." *Consciousness*
6 *and Cognition* 29 (October): 212–29.
7 Sanders, Joshua I., Balázs Hangya, and Adam Kepecs. 2016. "Signatures of a Statistical
8 Computation in the Human Sense of Confidence." *Neuron* 90 (3): 499–506.

9 Declaration of competing interests

10 The authors declare no competing interests.

11

1 Supplementary information

2 Training and staircase

3 Before the main part of the experiment, analyzed in the results section, participants went
4 through two initial parts.

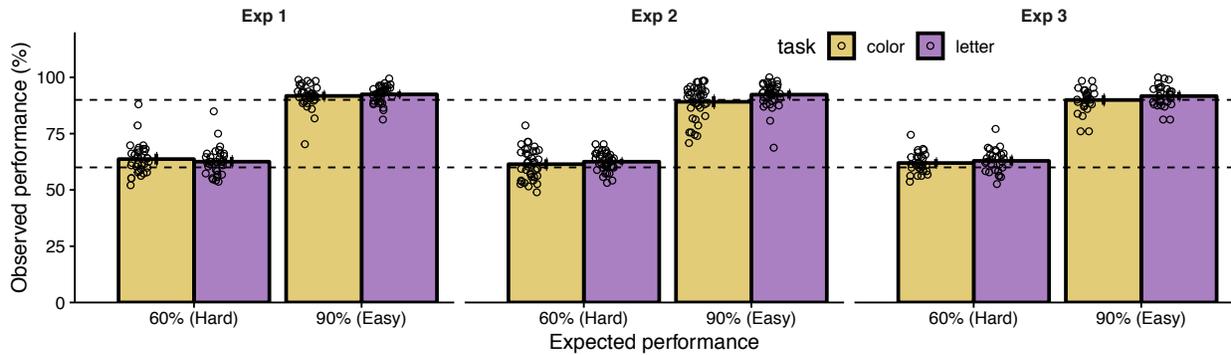
5
6 At the beginning of the experiment, participants were familiarized with the stimulus and the way
7 to give responses. To start with the easiest setting, they practised with trials for which only one
8 choice and subsequent confidence had to be given. Participants completed a short block of 5
9 trials concerning only the color task, and another block of 5 trials concerning only the letter task,
10 with the order randomized across participants.

11
12 After this short training, participants completed 96 trials very much resembling the main part of
13 the experiment. On each trial, they gave the responses for both tasks and, depending on the
14 experiment, rated their confidence for both or one task. However, there was a crucial difference
15 with the subsequent main part of the experiment: for each task, we implemented a staircase
16 procedure. On each trial, each task's difficulty was updated based on the last trial's response for
17 that task. Difficulty was altered by changing the proportion of elements (blue over orange, or O
18 over X). For the first trial, 64 of the 80 elements belonged to the dominant category. On
19 subsequent trials, a correct response would decrease the dominant category (make the task
20 more difficult) by one element, while an error would increase (make the task easier) it by 4
21 elements. In order to avoid participants tampering with the staircase procedure to make final
22 difficulties easier, the staircase trials were disguised as an initial part of 4 blocks. Except for the
23 adjustment of the stimulus, this part mimicked the design of the other parts of the experiment,
24 including the specific details of the confidence rating for that experiment.

25
26 Once the staircase trials had finished, we used the data from this part to obtain the proportion of
27 elements that would be used in the main part when presenting easy and hard trials. For the
28 color task, we estimated the psychometric function representing the probability of responding
29 blue as a function of the number of blue elements, fitted with a cumulative Gaussian. To
30 calculate the number of elements leading to a 90% expected performance (easy condition), we
31 took 40 (50% of the elements present on a trial) and added the the semi-difference between the
32 number of blue elements for which the psychometric curve predicted a 90% and 10% of blue
33 choices, rounded to its nearest integer. This procedure assumes that participants were not
34 biased towards responding more blue or orange. A similar procedure was used to obtain the
35 proportion of elements for the hard condition, and for both conditions of the letter task.

36
37 To demonstrate that the staircase procedure worked, Figure S1 shows, for all task types and
38 experiments, expected performance against observed performance in the main part of the
39 experiment. As can be seen, they were well-matched.

40



1
 2 *Figure S1. For each experiment, observed performance as a function of expected performance,*
 3 *both in percentage, split by task type. Bars represent averages across participants, with error*
 4 *bars denoting 95% confidence intervals. Dots represent individual participants. Top and bottom*
 5 *dashed lines help indicate where 90% and 60% observed performance would lay, respectively.*

6 Payoff

7 At the end of the experiment, participants were paid proportionally to both their performance in
 8 the perceptual tasks and to how accurately their confidence ratings reflected their performance.
 9 In Experiment 1, for each of the 16 blocks, a trial was randomly selected. Then, also at random,
 10 we chose the color or the letter task. We compared the confidence rating that had been given
 11 for that task, with a random number picked from a uniform distribution ranging from 50 to 100. If
 12 the random number was smaller than the confidence rating, payoff depended on performance:
 13 1€ was given if the choice for that task within that trial had been correct, and 0€ if it had been an
 14 error. If the random number was bigger than the confidence rating, payoff depended on a
 15 lottery: we compared the previous random number with another number was sampled from a
 16 uniform distribution between 0 and 100, and 1€ was given if the former was bigger than the
 17 latter, and 0€ otherwise.

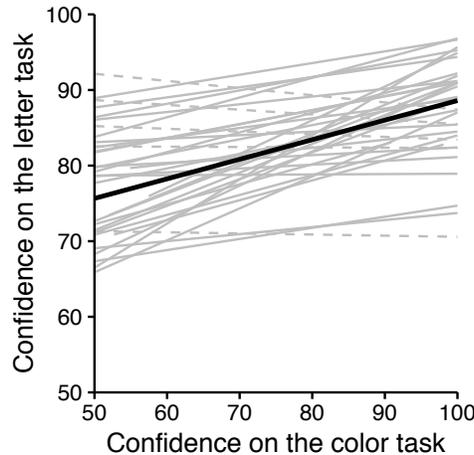
18
 19 In experiments 2 and 3 the payoff system was slightly different. Since only one confidence
 20 rating was given per trial, this was always what was compared with the initial random number. If
 21 payoff depended on performance, we then randomly chose whether it would depend on the
 22 correctness of the color or the letter task.

23
 24 This payoff system was adapted from (Massoni, Gajdos, and Vergnaud 2014). Its objectives are
 25 two. First, making participants be focused at all time, since any trial could be picked for payoff.
 26 Second, forcing participants to give a confidence rating that accurately reflected their belief in
 27 their choices to be correct. If they expressed a high confidence, they would be more likely to be
 28 paid based on their choice, whereas a low confidence would more easily lead to a lottery. In the
 29 long run, accurately reporting confidence would be translated into a higher payoff. This system
 30 and its rationale were explained to participants before the experiment.

1 **Correlation of confidence between tasks for Experiment 1**

2 In the main text we already reported how, in Experiment 1, the rating between the color and the
3 letter task within a trial were positively correlated. Figure S2 offers graphic support for that
4 finding.

5



6

7

8 *Figure S2. For Experiment 1, confidence on the color task against confidence on the letter task.*
9 *Each grey line corresponds to a linear regression for one participant, made using the confidence*
10 *ratings for all trials. For visualization purposes, the few participants with a negative r value are*
11 *depicted with a dashed line. The black line corresponds to a linear regression using the whole*
12 *data across participants.*

13 **Confidence rating as a function of accuracy and difficulty of that**
14 **task, split by experiment**

15 Here we report the results of the ANOVA on the LMM that we conducted to investigate how a
16 confidence rating changed based on the parameters of the task for which that rating had been
17 given. Unlike in the main text, where we collapsed our dataset across experiment, here we give
18 the results individually for each experiment, showing that the pattern does not change. Indeed,
19 note that the interaction between accuracy and difficulty is always very significant.

20

1 *Table S1. For each experiment, results of the ANOVA on an LMM where confidence for a task*
 2 *is predicted as a function of that task's type, response accuracy for that task on that trial, and*
 3 *difficulty for that task on that trial.*

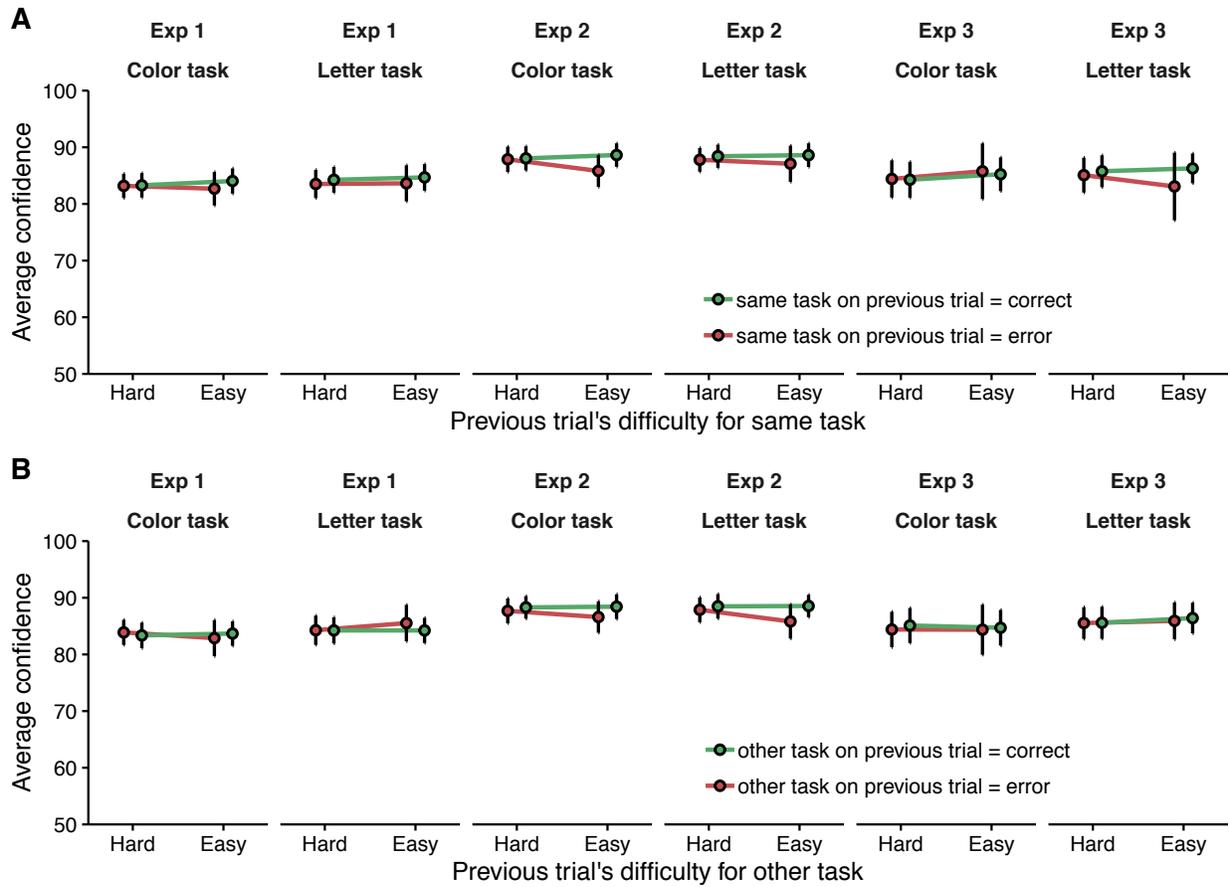
Experiment number	Task type	Parameter (always from that task)	Degrees freedom	F	p
1	Color	Accuracy	13026	986.315	< 0.001
		Difficulty	13022	19.814	< 0.001
		Accuracy * Difficulty	13024	399.138	< 0.001
	Letter	Accuracy	13022	1288.025	< 0.001
		Difficulty	13021	34.761	< 0.001
		Accuracy * Difficulty	13022	464.092	< 0.001
2	Color	Accuracy	7456	276.900	< 0.001
		Difficulty	7450	2.037	0.154
		Accuracy * Difficulty	7451	108.895	< 0.001
	Letter	Accuracy	7453	478.969	< 0.001
		Difficulty	7451	7.788	0.005
		Accuracy * Difficulty	7452	174.081	< 0.001
3	Color	Accuracy	5350	216.795	< 0.001
		Difficulty	5348	30.124	< 0.001
		Accuracy * Difficulty	5348	108.154	< 0.001
	Letter	Accuracy	5349	306.084	< 0.001
		Difficulty	5347	18.378	< 0.001
		Accuracy * Difficulty	5347	139.311	< 0.001

4

5 Analyses on the across-trial confidence leak

6 Our reported analyses so far have tried to determine the presence of a confidence leak from
 7 one task to another within the same trial. However, previous work has documented a leak from
 8 the confidence reported for tasks on the previous trial. Thus, we also tested for the existence of
 9 such leaks in our data. We explored whether the rating for a task was affected by confidence on
 10 the same task for the previous trial, but also by confidence on the other task for the previous
 11 trial. We approached this by expanding the LMM used to check for a confidence leak in the
 12 results section of the main text. The added fixed effects were accuracy and difficulty for the
 13 same task on the previous trial, and those two parameters for the other task on the previous
 14 trial. We then identified any significant interaction between the two parameters for one of the
 15 tasks. The interaction between accuracy and difficulty of the last trial's same task was only

1 significant for the color task of Experiment 2 ($F(7131.904) = 7.547, p = 0.006$), while it was
 2 marginally significant for the color task of Experiment 1 ($F(12470.525) = 3.594, p = 0.058$). The
 3 interaction for the parameters of the last trial's other task was significant for the color task of
 4 Experiment 2 ($F(7131.884) = 4.077, p = 0.044$), and marginally significant for the letter task of
 5 the same experiment ($F(7126.753) = 3.045, p = 0.081$). These results show that although there
 6 was evidence for across-trial confidence leak for some experiments and tasks, this was not as
 7 widespread as the within-trial leak. Figure S3 illustrates how confidence on a task depended on
 8 the parameters for the previous trial's same (Figure S3A) and other (Figure S3B) tasks.
 9



10
 11 *Figure S3. A. For each experiment and task, average confidence ratings across participants, as*
 12 *a function of the difficulty of the same task type on the past trial, and split according to the*
 13 *accuracy of the same task type on the past trial (color-coded). Error bars denote 95%*
 14 *confidence intervals. B. For each experiment and task, average confidence ratings across*
 15 *participants, as a function of the difficulty of the other task type on the past trial, and split*
 16 *according to the accuracy of the other task type on the past trial (color-coded). Error bars*
 17 *denote 95% confidence intervals.*

1 Rated confidence as a function of the order in which the rated 2 task was responded to

3 In our 3 experiments we always presented two tasks per trial. An interesting question not asked
4 in the main text is whether confidence ratings changed as a function of the order for which the
5 corresponding task had to be done: first or second within the trial. A confidence rating for the
6 first-responded task was only preceded by that task. A rating after the second-responded task
7 was preceded also by the other task, and in Experiment 1 also by its confidence rating. It could
8 be that confidence suffered a memory decay, such that, for a second-responded task,
9 confidence on that decision had somewhat waned. We investigated this possibility by taking the
10 LMM used in the main text to test for the confidence leak, and adding as a factor the order in
11 which that rating's task had been responded.

12

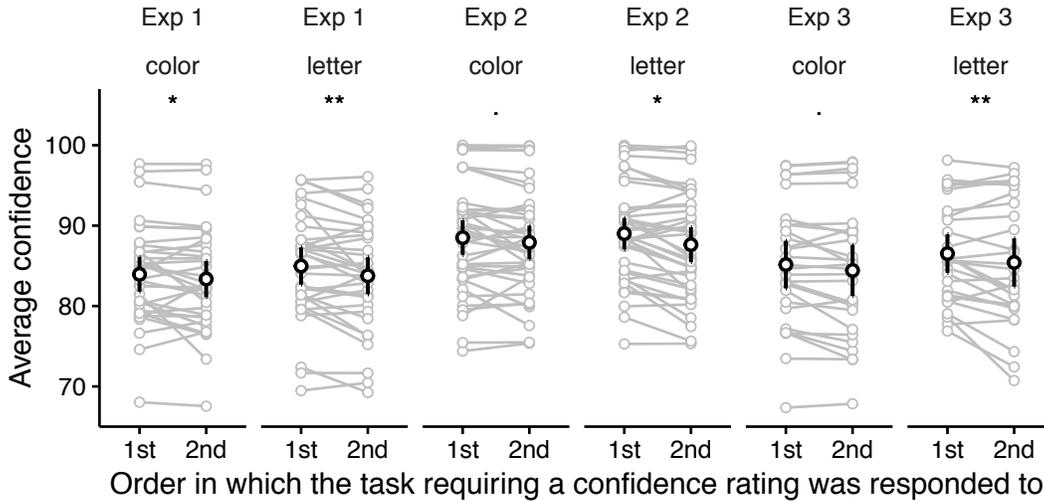
13 The main effect of order was significant or marginally significant for all experiments and tasks
14 (Exp. 1, color task: $F(13009.625) = 3.921$, $p = 0.048$; Exp. 1, letter task: $F(13009.509) = 8.112$,
15 $p = 0.004$; Exp. 2, color task: $F(7436.940) = 2.836$, $p = 0.092$; Exp. 2, letter task: $F(7437.229) =$
16 4.029 , $p = 0.045$; Exp. 3, color task: $F(5335.440) = 3.501$, $p = 0.061$; Exp. 3, letter task:
17 $F(5335.899) = 6.958$, $p = 0.008$). The direction of this effect revealed that, indeed, confidence
18 tended to be higher for first-responded tasks. Figure S4 gives a hint of this tendency.

19

20 In order to see whether order changed the confidence leak, we checked for a significant triple
21 interaction among accuracy for the other task, difficulty for the other task, and order. Only in
22 Experiment 2 for the color task did this interaction approach significance ($F(7347.829) = 3.016$,
23 $p = 0.083$). In summary, there was a tendency for confidence ratings to be higher for first-
24 responded tasks than for second-reported tasks, but overall that order difference did not affect
25 the confidence leak.

26

1



2

3

4

5

6

7

8

9

Figure S4. For all experiments and task types, average confidence as a function of whether the confidence rating was given for the first- or for the second-responded task. Individual participants are depicted in grey. Black, bigger dots represent across-participant averages, with 95% confidence intervals. Asterisks above the plotted data denote the significance level for the main effect of order given by the ANOVA on the LMM described in this section. A dot instead of an asterisk denotes marginal significance.

9

10 References

11 Massoni, Sébastien, Thibault Gajdos, and Jean-Christophe Vergnaud. 2014. "Confidence
 12 Measurement in the Light of Signal Detection Theory." *Frontiers in Psychology* 5
 13 (December): 1455.

14