## Study Information

### 1. Title

Role of feedback on metacognitive training

### 2. Authorship, by alphabetical order

Vincent de Gardelle, Nathan Faivre, Elisa Filevich, Gabriel Reyes, Martin Rouy, Jérôme Sackur, Jean-Christophe Vergnaud

### 3. Research Questions

This study follows up on previous work by Carpenter and colleagues (2019), who showed that metacognitive performance can be improved through adaptive training. In their study, participants were asked to perform a perceptual discrimination task, and subsequently report the confidence in their response using a 4-point likert scale. Metacognitive training was quantified by comparing meta-performance measured in the first session (S1, pre-training) and in the final session (S10, post-training), after 8 sessions of training (S2-9). During these training sessions, to encourage participants to report confidence as accurately as possible (and therefore enhance meta-performance), a reward was provided at the end of each block of 27 trials, depending on how closely confidence ratings tracked first-order accuracy. To compute this reward, confidence was mapped from the 4-point scale onto a subjective probability of answering correctly, and transformed into points so that participants received the highest number of points when they were highly confident in their correct responses, and unconfident in their errors. Another group of participants took another version of the experiment which served as control. In the control group, participants received feedback on their decision performance during training, independently of confidence judgments. Results show better meta-performance in the treatment compared to the control group, as well as a transfer of metacognitive improvements between a perceptual and a memory task. The present study aims at clarifying the origins of metacognitive improvements by manipulating the type of feedback received by subjects.

In the original study subjects were instructed in the pre-training phase to report confidence on a four level scale, defined as 1 = “very low confidence”, 2 = “low confidence”, 3 = “high confidence” and 4 = “very high confidence”. Importantly no explicit mapping from confidence levels to subjective probabilities was given to participants. We believe that in this context, the correct interpretation of the lowest confidence rating is that of a 50% chance of being correct, and therefore that participants are provided with a half-scale of confidence. Yet in the training phase, feedback was computed assuming a full scale as confidence was mapped onto a probability of a response being correct from 0 to 1. As a result, confidence ratings 1 and 2 were to be used in case subjects thought they made an error (level 1 would be used when they were certain that they made an error), which rarely occurs in such experimental settings. The original study shows that 1) subjects used mostly confidence 3 and 4 after one training session, 2) the increase in meta-performance is mediated by this increase in mean confidence, and 3) meta-performance increases rapidly at the beginning of the training and then remains constant.

To assess whether the increase in meta-performance observed in the original study stems from an incongruence between instructions and feedback, we plan on training participants with feedback that is compatible with instructions. Namely, participants will be instructed to report confidence on a Likert scale with 1 = “very low confidence”, 2 = “low confidence”, 3 = “high confidence” and 4 = “very high confidence”. This is equivalent to the instructions participants saw at the beginning of the original study, but not later on during training and the rest of the experiment. As opposed to the original study, confidence will be mapped onto a probability of being correct between 0.5 and 1, as follows: P(correct) = (conf+2)/6. Subsequently the QSR score is obtained by computing 1 - (accuracy - P(correct))^2, for each trial. As in the original study, participants will be informed that in addition to base payments per session, they would receive bonus payments in each session if their score would exceed the score of a randomly chosen participant.

One other factor that might affect performance in the training phase is the introduction of incentives. Indeed, subjects were rewarded during training sessions, but no reward was offered pre and post-training. Therefore, one possibility to explain increased meta-performance is a difference in terms of task engagement between the pre and post-training sessions: subjects may have reported confidence more accurately as a result of incentives rather than increased meta-performance per se. To rule out this possibility, we will test metacognitive performance between S2 and S10, that is under constant rewarding scheme.

Concretely, we will implement the following changes in the experimental scripts used to collect data by Carpenter et al.:

1. After the instructions for each task (perceptual or memory) are provided, participants respond to two questions on a slider. Carpenter and colleagues asked:
   i. "*In the [perception/memory] task, what confidence rating should you give if you are **100% sure you are correct**?*"
   ii. "*In the [perception/memory] task, what confidence rating should you give if you are **100% sure you are incorrect**?*"
   Participants did not advance with the experiment until they provided the expected answer (4 and 1, respectively). We will change the questions to read:
   i. “*In the perception task, what confidence rating should you give if you are **very sure you are correct**?*”
   ii. “*In the perception task, what confidence rating should you give if you are **not at all sure you are correct**?”*
2. In their study, Carpenter and colleagues calculated the QSR score using the formulas:

$1 -(1 - (-1/3 + (\mathtt{confidence}/3))^2$ for correct trials and

$1 -(0 - (-1/3 + (\mathtt{confidence}/3))^2$ for incorrect trials.

We will calculate the QSR score using the following formulas instead:

$1 -(1 - (+1/3 + (\mathtt{confidence}/6))^2$ for correct trials and

$1 -(0 - (+1/3 + (\mathtt{confidence}/6))^2$ for incorrect trials.

3. [Not sure we want to mention this, because we never saw Carpenter's actual study, maybe this wasn't an issue somehow]. We fixed a small error in the code shared by Carpenter and colleagues that led to some of the images presented in the memory shapes condition to be presented more than once in each memory miniblock, and other images to never be displayed. This led to several invalid trials in the first-order discrimination memory task, where there was effectively no correct answer, as none of the two images presented in the discrimination task had ever been presented. (We note that if this error occurred, it was both for the pre- and post-training and does not, *per se*, invalidate the conclusions about training effects.)
4. Critically, we now provide detailed instructions about how to map confidence to correct trials after the titration tasks in Session 1 but *before* any task where participants rate confidence. These instructions include a predefined set of demonstration trials and a series of practice trials with trial-wise feedback about whether confidence ratings were correctly assigned to correct or incorrect trials.
5. In their original study, Carpenter et al. awarded points during the visual tasks in the training sessions alone, and not for the pre-and post-training sessions. We will award points to the visual (but not memory) tasks in pre- and post-training sessions, as well.
6. In the pre- and post-training sessions, participants in the study by Carpenter et al. could either start with the memory tasks or the perception tasks. As a consequence of the changes described in point 5, participants will always start with the perception task. This is to allow for continuity in the explanation of how points are calculated and assigning points to participants.
7. Carpenter et al. ran the initial titration staircase (according to the code, this is not reported in the paper) until a fixed number of reversals was reached, or a maximum of 60 trials. We will run the titration staircase for a fixed number of 60 trials.
8. We will test only one group of participants receiving feedback according to confidence (i.e., intervention group). This group will be compared to the control group from the original study receiving feedback on first-order performance.

### 4. Hypotheses

Primary hypotheses regarding the role of feedback:

- H1: meta-performance as assessed with log M-ratio differs between S1 and S10, even when issues related to feedback are corrected.
- H0: meta-performance as assessed with log M-ratio does not differ between S1 and S10

If H1 is true, we can further test whether the metacognitive training effect is general across metacognitive tasks, following the same approach as in the original study.

Secondary hypotheses regarding the role of incentives:

- H1: meta-performance as assessed with log M-ratio differs between S2 and S10, even when issues related to incentives are corrected.
- H0: meta-performance as assessed with log M-ratio does not differ between S2 and S10

## Sampling Plan

### 1. Existing data

Registration prior to creation of data: As of the date of submission of this research plan for preregistration, the data in the treatment group have not yet been collected, created, or realized.

### 2. Explanation of existing data

Data in the treatment group will be compared to the control group recruited by Carpenter et al, available online: https://github.com/metacoglab/CarpenterMetaTraining

This group was rewarded depending on their first-order performance, independent of confidence.

### 3. Data collection procedures.

Data will be collected on MTurk, using the same scripts and materials as Carpenter and colleagues.

### 4. Sample size

We will adopt an open-ended sequential Bayes factor design, whereby we will test our primary effect of interest after each participant and decide to stop data collection whenever there is moderate evidence for either H0 (no difference of log M-ratio between S1 and S10)  or H1 (significant difference of log M-ratio between S1 and S10).

### 5. Sample size rationale

Following Schönbrodt & Wagenmakers (2018), we ran simulations for Bayesian simulation factor analysis. Assuming an effect size of d = 0.3 and standard prior distribution (t with df = 1 and scale factor = sqrt(2)/2), these simulations show that a one-sided paired bayesian t-test yields conclusive evidence under H1 (i.e BF > 5) in 74% of simulations, and under H0 (i.e BF < 0.2) in 89% of simulations with N = 100.

### 6. Stopping rule

Data will be acquired until a Bayes factor of 0.2 or 5 is obtained.

### Variables

### 7. Manipulated variables

### 8. Measured variables

### Behavioral variables

First-order accuracy (binary: correct/incorrect on the visual discrimination task)
First-order reaction time (continuous: time to respond to the first-order task)
Confidence (ordinal: Likert scale 1-4)
Second-order reaction time (continuous: time to report confidence)

### 9. Indices

See analysis plan section below

## Design Plan

### 10. Study type

Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.

### 11. Blinding

No blinding is involved in this study.

### 12. Study design

Identical to the study by Carpenter et al, except for the rule determining feedback (see above).

### 13. Randomization

As in the original study, bonuses will be distributed pseudorandomly to ensure equivalent financial motivation irrespective of performance.

## Analysis Plan

### 14. Statistical models

We will employ the same analysis plan as Carpenter and colleagues.

### 15. Transformations

NA

### 16. Follow-up analyses

NA

### 17. Inference criteria

Moderate evidence in favor of H0 or H1 (Bayes factor > 3 or < ⅓).

### 18. Data exclusion

The same exclusion criteria as defined in the original study by Carpenter and colleagues will be used.

A subject is excluded in case of:

- floor or ceiling performance in the pre-training baseline session
- first-order performance in the range of 55 – 95% in at least one condition/session
- average difficulty level calculated across all sessions dropping below 2.5 standard deviations below the group mean difficulty level
- reporting the same confidence level on 95% of trials for 3 or more sessions.

Trials with reaction times > 2000ms or < 200ms will be excluded.

### 19. Missing data

NA