Neural correlates of metacognition across the adult lifespan

Helen Overhoff[a,b,c*], Yiu Hong Ko[a,b,c], Daniel Feuerriegel[b], Gereon R. Fink[a,d], Jutta Stahl[c], Peter H. Weiss[a,d], Stefan Bode[b], & Eva Niessen[c]

[a]Cognitive Neuroscience, Institute of Neuroscience and Medicine (INM-3), Research Centre Jülich, Leo-Brandt-Str. 5, 52425 Jülich, Germany

[b]Melbourne School of Psychological Sciences, University of Melbourne, Parkville Campus, Parkville 3010, Victoria, Australia

[c]Department of Individual Differences and Psychological Assessment, University of Cologne, Pohligstr. 1, 50969 Cologne, Germany

[d]Department of Neurology, Faculty of Medicine, University Hospital Cologne, University of Cologne, Kerpener Str. 62, 50937 Cologne, Germany

*Corresponding author. Cognitive Neuroscience, Institute of Neuroscience and Medicine (INM-3), Research Centre Jülich, Leo-Brandt-Str. 5, 52425 Juelich, Germany.

Email: h.overhoff@fz-juelich.de

AGEING AND METACOGNITION

Highlights:

- Metacognition was assessed in a complex flanker task across the lifespan
- Metacognitive accuracy declined with age beyond reduced task performance
- Modulation of behavioural adaptations by confidence was stable across the lifespan
- Error-specific age-related decline was observed for $N_{e/c}$ but not $P_{e/c}$ amplitudes
- Neural correlates of confidence did not reflect changes in metacognitive accuracy

Neural correlates of metacognition across the adult lifespan

Abstract

Metacognitive accuracy describes the degree of overlap between the subjective perception of one's decision accuracy (i.e. confidence) and objectively observed performance. With older age, the need for accurate metacognitive evaluation increases; however, error detection rates typically decrease. We investigated the effect of ageing on metacognitive accuracy using event-related potentials (ERPs) reflecting error detection and confidence: the error/correct negativity ($N_{e/c}$) and the error/correct positivity ($P_{e/c}$). Sixty-five healthy adults (20 to 76 years) completed a complex Flanker task and provided confidence ratings. We found that metacognitive accuracy declined with age beyond the expected decline in task performance, while the adaptive adjustment of behaviour was well preserved. P_e amplitudes following errors varied by confidence rating, but they did not mirror the reduction in metacognitive accuracy. N_e amplitudes decreased with age for low confidence errors. The results suggest that age-related difficulties in metacognitive evaluation could be related to an impaired integration of decision accuracy and confidence information processing. Ultimately, training the metacognitive evaluation of fundamental decisions in older adults might constitute a promising endeavour.

Keywords: confidence, error detection, aging, error(-related) negativity (N_e, ERN), error positivity (P_e), behavioral adaptation

1. Introduction

We are continuously monitoring and controlling our behaviour in order to achieve goals and avoid errors. The internal evaluation of our behaviour and our decisions, also referred to as *metacognition*, is crucial in everyday life, because it guides our present and future behaviour (Desender et al., 2019b; Rabbitt, 1966). Metacognition comprises both the detection of committed errors and a feeling of confidence that accompanies a decision (Fleming and Frith, 2014; Shekhar and Rahnev, 2020). When we feel less confident about a decision, we might try to adjust it, seek more information, or recruit additional cognitive processes to optimise performance (Desender et al., 2019a, 2019b). As ageing is usually associated with declining cognitive functions and higher rates of decision errors in daily activities, decisions and corresponding motor actions need to be adjusted more often (Hertzog, 2015; Ruitenberg et al., 2014). This might be achieved, for example, by increasing efforts for an efficient metacognitive evaluation of one's behaviour.

In general, metacognitive judgements are highly predictive of actual task performance, yet there is strong evidence that metacognition constitutes a dissociable process from the execution of the initial task (Galvin et al., 2003; Song et al., 2011). The degree to which subjective perceptions and objectively observed performance overlap, that is, the *accuracy* of metacognitive judgements, varies across individuals and task demands (Fleming & Dolan, 2012; Hertzog & Hultsch, 2000; Rahnev et al., 2020). Metacognitive accuracy has been addressed in two separate but arguably related fields of research: studies on error detection, focussing on the recognition of errors, and studies on decision confidence, investigating processes related to beliefs regarding the likelihood of having made a correct choice. In most cases, low confidence implies a higher probability of having committed an error. It has been suggested that error detection and confidence judgements might even share similar underlying computations, whereby error detection arises from low

confidence that a correct decision has been made (Boldt and Yeung, 2015; Yeung and Cohen, 2006; Yeung and Summerfield, 2014).

1.1 Neural correlates of metacognition

Neural correlates of metacognition have been studied by measuring event-related potentials (ERPs) of the human scalp electroencephalogram (EEG). The error negativity (N_e) is a negative deflection peaking around 100 ms after an overt behavioural response at fronto-central electrodes and typically has larger amplitudes for errors than correct responses (N_c for correct responses; i.e. correct negativity; Falkenstein et al., 1991; Falkenstein et al., 2000; Vidal et al., 2003). The component is classically associated with conflict monitoring, assuming that it tracks conflict between the given response and continuously-accumulated post-decision evidence favouring the correct response (Falkenstein et al., 1991; Yeung et al., 2004). Moreover, it has been shown that the N_e amplitude scales with confidence, that is, it decreases from perceived errors to uncertain responses (guesses) to trials where the participant is confident about its correctness (Boldt and Yeung, 2015; Scheffers and Coles, 2000). The more posterior error positivity (P_e; P_c for correct responses, i.e. correct positivity) with a maximum amplitude around 250 ms after a response, is considerably larger for detected compared to undetected errors and has therefore been associated with explicit error awareness (Endrass et al., 2012a; Nieuwenhuis et al., 2001). Notably, the P_e has also been found to increase in amplitude with decreasing confidence in perceptual decisions (Boldt and Yeung, 2015; Rausch et al., 2019).

Concerning the mechanisms underlying these two components, Di Gregorio et al. (2018) designed a sophisticated task to provide evidence that the P_e, but not the N_e, was present when it was evident for participants that an error had been made, but they did not know the correct answer. These findings suggest that the P_e does not require a representation of the correct response to emerge, but instead accumulates post-decisional error evidence from widely distributed neural

sources (Di Gregorio et al., 2018; Murphy et al., 2015; Steinhauser and Yeung, 2010; Yeung and Summerfield, 2014). Thus, while both classical components of error processing, N_e and P_e, have been shown to vary with reported confidence, the Pe appears to be more specifically associated with conscious metacognitive processes (Boldt and Yeung, 2015; Nieuwenhuis et al., 2001; Scheffers and Coles, 2000).

1.2 Metacognition and ageing

Metacognitive abilities in older age have been shown to vary across cognitive domains (Fitzgerald et al., 2017; Hertzog & Hultsch, 2000). For instance, while older adults tend to underestimate the prevalence of their decision errors in everyday life, metacognitive judgements of certain memory aspects (e.g., memory encoding) seem to be well preserved (Castel et al., 2016; Harty et al., 2013; Mecacci and Righi, 2006). Previous studies on decision making and metacognition yielded relatively consistent findings of a significant decline in error detection rate with higher age across multiple tasks (Harty et al., 2013; Rabbitt, 1990), even when task performance was comparable (Harty et al., 2017; Niessen et al., 2017; Wessel et al., 2018). In a large sample of healthy adults, Palmer et al. (2014) investigated decision confidence using a measure of metacognitive accuracy that takes task performance into account (Maniscalco & Lau, 2012). The authors found that age was not correlated with metacognitive abilities in a memory task, but that it was negatively correlated with metacognitive abilities in a perceptual discrimination task.

Effects of ageing on the neural correlates of metacognition have primarily been investigated in the field of error detection. Here, both the difference between N_e and N_c (Endrass, Schreiber, et al., 2012; Falkenstein et al., 2001; Schreiber et al., 2011), and the $P_{e/c}$ amplitude (Clawson et al., 2017; Harty et al., 2017; Niessen et al., 2017) was smaller in older adults, while the decrease in P_e, in particular, was linked to a lower error detection rate. Notably, the processing of the stimulus can also affect subsequent response-related processes, and variations with age in two ERPs (namely

the N2 and the P300; Groom & Cragg, 2015; Polich, 2007) have been documented (Korsch et al., 2016; Larson et al., 2016; Lucci et al., 2013; Niessen et al., 2017). With the decline in behavioural performance reported above, this suggests an impaired error evidence accumulation process in older age, possibly due to limited cognitive resources (Harty et al., 2017; Niessen et al., 2017). Surprisingly, neither $N_{e/c}$ nor $P_{e/c}$ have been investigated using confidence ratings to assess age-related variations of metacognitive abilities. Some evidence from neuroimaging studies point to age-related structural differences in the neural basis of metacognition (Chua et al., 2009; Hoerold et al., 2013; Sim et al., 2020). However, a conclusive account that explains individual differences in metacognitive accuracy is still missing, for which the use of ERPs with high temporal resolution might be well-suited to provide valuable insights (Dully et al., 2018; Fleming and Dolan, 2012; Yeung and Summerfield, 2014).

1.3 The current study

This study aimed to investigate task performance and metacognition in older adults with a novel perceptual task to determine how generalizable the findings of decreased metacognitive accuracy in older age are (Palmer et al., 2014). For this, we used a colour-flanker task, in which participants had to identify the colour of a target stimulus that was flanked by two squares of the same or a different colour. We assessed decision accuracy, measured confidence using a four-point rating scale, and examined the impact of metacognitive accuracy on adaptations of subsequent behaviour (Desender et al., 2019a; Fleming et al., 2012; Ruitenberg et al., 2014). Furthermore, we investigated whether the amplitudes of $N_{e/c}$ and $P_{e/c}$, which are described as neural correlates of metacognition, track changes in decision confidence across the lifespan.

We hypothesised that metacognitive accuracy in our decision task would decrease with age (Niessen et al., 2017; Palmer et al., 2014). Independent of confidence, we expected an error-specific attenuation of ERP component amplitudes in older adults, which should result in a smaller

difference between the neural responses related to errors and correct decisions (Endrass et al., 2012b; Larson et al., 2016). Independent of age, reported confidence was expected to show a positive association with the $N_{e/c}$ and a negative association with the $P_{e/c}$ amplitude (Boldt and Yeung, 2015; Scheffers and Coles, 2000). Based on findings from error detection studies showing an age-related decrease in the P_e amplitude of detected, but not undetected errors (Harty et al., 2017; Niessen et al., 2017), as well as reports linking the P_e to confidence (Boldt and Yeung, 2015), we expected a specific decrease in P_e amplitude for low confidence errors with increasing age.

2. Methods

2.1 Participants

We recruited 82 healthy adults with a broad age range from 20 to 81 years (49.8 $\pm$ 1.9 years [all results are indicated as mean $\pm$ standard error of the mean; *SEM*]; 35 female, 47 male). We aimed for an approximately uniform distribution of age and thus tested at least 10 participants per decade. Inclusion criteria were right-handedness according to the Edinburgh Handedness Inventory (EDI; Oldfield, 1971), fluency in German, (corrected-to-) normal visual acuity, no colour-blindness and no history of neurological or psychiatric diseases. Any signs of cognitive impairment (Mini-Mental-State Examination score lower than 24; MMSE; Folstein et al., 1975) or depression (Beck's Depression Inventory score higher than 17; BDI; Hautzinger, 1991) led to the exclusion of participants (one participant was excluded). Additionally, we excluded four participants who had more than one third of invalid trials (e.g., responses were too slow to fall into the pre-defined response window for analysis, or they showed recording artefacts). Another four participants were excluded because of an error rate (ER) higher than the chance level of 75%. Finally, eight participants were excluded because of combinations of very low accuracy, a high number of invalid trials, the selective use of single response keys, and errors in the colour discrimination task

(described below), which suggested a lack of understanding of the task or the use of heuristic response strategies instead of trial-by-trial decisions. After exclusions, the final sample consisted of 65 healthy adults (45.5 ± 2.0 years; 20 to 76 years; 26 female, 39 male).

The study was approved by the ethics committee of the German Psychological Society (DGPs) and conformed to the declaration of Helsinki. All participants gave written informed consent before participating in the experiment.

2.2 Experimental paradigm

The main experimental task consisted of a modified version of the Eriksen flanker task using coloured squares as stimuli and four response options (Eriksen & Eriksen, 1974; Maier & Steinhauser, 2017; Figure 1A). Participants were asked to respond as fast and accurately as possible to a centrally presented target by pressing a button with one of their index or middle fingers, mapped onto four designated target colours. In each trial, the target consisted of one of these target colours, and the flankers, located right and left to the target, consisted either of the same colour as the target (congruent condition), of another target colour (incongruent condition), or of one of three additional neutral colours, which were not mapped to any response (neutral condition [Maier et al., 2008]; see Figure 1B). Both the incongruent and the neutral condition were used to induce conflict as they provided information distinct from the target. We chose this version of the classical flanker paradigm in order to increase task difficulty and thereby maximise the number of errors without tapping into other cognitive processes that might be affected by ageing (e.g., spatial, lexical, or semantic cognition). The colour-finger mapping was fixed over the course of the experiment for each participant and counterbalanced across participants.

Each trial started with a fixation cross for 500 ms. Then, flankers were presented for 50 ms before the target was added to the display for another 100 ms. Showing the (task-irrelevant) flankers before the target was expected to increase the induced conflict (Mattler, 2003). We used a

response deadline of 1,200 ms because this timing provided a good balance between a desirable number of errors and feasibility for all participants. If no response was registered before this deadline, the trial was terminated and the feedback ‘zu langsam’ (German for ‘too slow’) was presented on the screen. If a response was given, a confidence rating scale appeared after a black screen of 800 ms. The delay was introduced to avoid that EEG activity related to the first response overlapped with the confidence assessment. Participants were asked to indicate their confidence in their decision on a four-point rating scale from ‘surely wrong’ to ‘surely correct’ using the same keys as for the initial response. The maximum time for the confidence judgment was 2,000 ms. Trials were separated by a jittered intertrial interval of 400 to 600 ms. The sequence of an experimental trial is depicted in Figure 1C.

2.3 Procedures

Prior to testing, participants were asked to provide demographic details and complete the handedness questionnaire. Afterwards, they completed a brief colour discrimination task (without EEG) to ensure that all participants were able to correctly discriminate the seven different colours used in the experimental paradigm (see Figure 1B). The discrimination task was followed by the EEG preparation and the main task. The neuropsychological tests were administered after the experiment. In addition, we assessed sustained attention span and processing speed using the d2-test (Brickenkamp, 2002), which have been shown to be positively associated with error processing abilities (Larson et al., 2011).

All stimuli in both tasks were presented on a black screen (LCD monitor, 60 Hz) in an electrically shielded and noise-insulated chamber with dimmed illumination, using Presentation software (Neurobehavioural Systems, version 14.5) for the colour discrimination task and uVariotest software (Version 1.978) for the main task. A chin rest ensured a viewing distance of

70 cm to the screen and minimised movements. To record participants' responses, we used custom-made force-sensitive keys with a sampling rate of 1024 Hz (see Stahl et al., 2020).

The experiment started with a practice block of 18 trials in which participants received feedback about the accuracy of their response, which could be repeated if the participant considered it necessary. After that, two additional blocks with 72 trials without feedback and confidence assessments served as training blocks, allowing the participants to memorise the colour-finger mapping and to get accustomed to the response keys. Afterwards, another practice block introduced the confidence rating to ensure that participants understood and correctly applied the rating scale. The following main experiment consisted of five blocks with 72 trials each. Participants were allowed to take self-timed breaks after each block. The entire session lasted approximately three hours.

2.4 Electroencephalography recording and preprocessing

The EEG was recorded using 61 active electrodes (Acticap, Brain Products) aligned according to the international 10-20 system (Jasper, 1958). The electrodes were online referenced against the posterior Iz electrode close to the inion. Horizontal eye movements were measured using two electrodes at the outer canthi of the eyes (horizontal electrooculogram [EOG]), and another electrode underneath the left eye measured vertical movements (vertical EOG). The EEG signal was recorded continuously at a sampling rate of 500 Hz using a digital BrainAmp DC amplifier (Brain Products). Data were filtered between 0.1 Hz and 70 Hz, and a notch filter of 50 Hz was applied to remove line noise.

EEG data were preprocessed following a standardised pipeline using the MATLAB-based toolboxes EEGLAB and ERPLAB (Delorme and Makeig, 2004; Lopez-Calderon and Luck, 2014). The signal was segmented from -150 to 2,000 ms relative to target stimulus presentation (note that the flankers were presented at -50 ms). Epochs were visually inspected for artefacts and noisy

electrodes. Epochs with artefacts were removed and identified noisy channels were interpolated using spherical spline interpolation. To identify and remove eyeblinks, we ran an Independent Component Analysis (ICA) using the infomax algorithm implemented in EEGLAB and afterwards baseline-corrected the epochs using the period of -150 ms to -50 ms to avoid influences of early perceptual processes related to the flanker presentation. Next, data were locked to the response, epoched from -150 ms to 800 ms relative to response onset and baseline-corrected using the 100 ms before the response. The additional analysis of conflict-related stimulus-locked ERPs can be found in the supplementary material S4. Remaining artefacts exceeding $\pm$ 150 μV were removed (Niessen et al., 2017), and a current source density (CSD) analysis was conducted using the CSD toolbox (Kayser and Tenke, 2006) allowing for better spatial isolation of ERP components and for obtaining a reference-independent measure (Perrin et al., 1989).

2.5 Behavioural data analysis

Trials with invalid responses (i.e. responses that were too slow) or recording artefacts, as well as responses faster than 200 ms were excluded from further analysis. The error rate (ER) was calculated as the proportion of errors relative to valid responses. Response time (RT) was defined as the time between stimulus onset and the initial crossing of the force threshold (40 cN) by any of the response keys.

To inspect how the confidence scale was used across participants, raw distributions of confidence ratings within all incorrect and correct responses were extracted. We computed Friedman ANOVAs for the proportion of each of each rating level for errors and correct responses with the factor confidence (4 levels). This analysis revealed that only a limited number of trials was available for the two middle confidence rating levels (‘maybe wrong’, ‘maybe correct’), and we therefore collapsed those to create one category for all further analyses representing ‘unsure’ responses, i.e. confidence ratings expressing uncertainty.

For the analysis of metacognitive accuracy, we computed the Phi (Φ) correlation coefficient, which is a simple trial-wise correlation between task accuracy and reported confidence. It describes the extent to which the distributions of confidence ratings for correct and incorrect trials differ, while still depending on primary task performance and individual biases in confidence judgements (Fleming and Lau, 2014; Kornell et al., 2007; Nelson, 1984). Phi was calculated by correlating accuracy, coded as 0 (error) and 1 (correct response), and confidence (that the given response was correct), coded as 1 ('surely wrong'), 2 ('unsure'), and 3 ('surely correct'), for each participant. This provided us with one measure of metacognitive ability per participant that comprises both the accuracy and the confidence rating of each trial (e.g., Phi = 1 means that correct trials were successfully identified as such without uncertainty; while a Phi = 0 means that all errors were rated as 'surely correct', or all correct trials were rated as 'surely incorrect').

To assess the impact of accuracy and confidence on trial *n* on adaptations of behavioural responses, we computed a measure of response caution by multiplying the accuracy and RT on trial *n*+1 (Desender et al., 2019a). Response caution captures the trade-off between speed and accuracy in a decision, with higher values indicating a more cautious response strategy that is characterised by slower, and at the same time, more accurate responses. For this analysis, only pairs of two consecutive valid trials were included. Response caution was computed separately relative to a) initial trial accuracy (error, correct), and b) initial trial confidence ('surely wrong', 'unsure', 'surely correct').

At the group level, age-related effects on the d2-test score, the error rate, and Phi were computed using linear regressions. To rule out that metacognitive accuracy was confounded by age-related impairments in task performance or attention and processing speed, we performed additional multiple linear regressions to predict phi by age, adding the factors of error rate or d2-test score, respectively.

For the analysis of performance and confidence at the trial level, data were analysed using linear and generalised linear mixed effects models. We always used the between-subjects factor age as a predictor. The within-subject factor of interest was either accuracy (error, correct) or (pooled) confidence (3 levels). We fitted random intercepts for participants and, if possible, random slopes by participant for the within-subject factor of interest. For the outcome variables of RT, confidence and response caution, we fitted linear mixed models, for which *F* statistics are reported and degrees of freedom were estimated by Satterthwaite's approximation, and for accuracy we fitted generalised linear mixed models, for which X^2 statistics are reported. Model structures and coefficients are reported in the supplementary material S1.

Significant effects of confidence were followed up by pairwise comparisons across rating levels using paired-samples *t*-tests for linear mixed models and *Z*-tests for generalised linear mixed models. Significant interactions were followed up by (generalised) linear mixed regressions, separately for each level of a given within-subject factor to assess potential effects of age. We decided on these follow-up tests because our main interest was in the differential relations between accuracy, confidence, and behaviour across the lifespan rather than between the levels. Post-hoc test results were compared against Holm corrected significance thresholds to account for multiple comparisons.

Analyses were run in MATLAB R2019a (The Mathworks, Inc.) and R (version 4.0.5; R Core Team, 2021) using the lme4 package (version 1.1; Bates et al., 2015).

2.6 Electroencephalographic data analysis

One participant had to be removed from electroencephalographic analyses, because noisy EEG data led to the exclusion of more than half of the trials. Data were response-locked and analysed at the single trial level. To obtain meaningful time-windows for amplitude extraction, we first computed the grand-average for all participants, separately for errors and correct responses.

The latency of the grand-average peak amplitude served as the time point around which individual mean amplitudes were extracted from the signal (± 50 ms). This was done to obtain meaningful time windows for statistical analyses, because data of single trials is too noisy to identify a meaningful peak (Clayson et al., 2013). On each trial, the $N_{e/c}$ local peak amplitudes were extracted from the response-locked data from the interval 0 to 150 ms following the response at FCz, and the $P_{e/c}$ local peak amplitudes were extracted from the interval 150 to 350 ms at Cz. This was based on visual inspection of the local maxima of the grand-average scalp topographies as well as previous literature (Falkenstein et al., 2000; Siswandari et al., 2019).

For statistical analyses of ERP amplitudes, we fitted the same linear mixed effects regression models as for the behavioural data. They included fixed effects of age and the within-subject factor accuracy (error, correct) for all trials combined (see supplementary material S3 for the analysis with the within-subject factor confidence for all trials) or confidence (3 levels) for separate analysis of errors and correct responses, a random intercept for each participant, and a random slope of the within subject factor by participant, if possible. The models were fitted to the CSD-transformed single trial mean ERP amplitudes of the $N_{e/c}$ and $P_{e/c}$. Model structures and coefficients are reported in the supplementary material S2.

3. Results

For brevity, only significant effects in the mixed effects regression analyses and relevant follow-up tests are reported in this section. For results of all tests as well as Bayesian analyses of relevant null effects, please refer to the supplementary material S1, S2 and S6.

3.1 Behavioural results

3.1.1 Attention

The average score for sustained attention and processing speed as assessed by the d2-test was 178.5 ± 5.6 ($M \pm SEM$) and showed the typical decline for older adults, as reflected in a significant prediction of the test scores by age [$F(1,63) = 27.819$, $p < .001$; $\beta = -1.536$, $SE = 0.291$].

3.1.2 Distribution of confidence ratings

In a first step, we were interested in how the confidence ratings were distributed across the four confidence levels across the lifespan (Figure 2). For this, we ran two Friedman ANOVAs for dependent measures for the proportion for each rating category, separately for errors and correct responses.

The ANOVA for errors showed that the proportion differed between confidence levels [$X^2(3) = 78.029$, $p < .001$; Figure 2A]. On average, most errors were rated as 'surely wrong' (42.8 %) and least errors as 'maybe wrong' (7.3%). Follow-up linear regressions on age-related differences for each rating category showed that the proportion of 'maybe correct' ratings was increased with higher age [$F(1,63) = 15.973$, $p < .001$; $\beta = 0.005$, $SE = 0.001$], whereas the ratio of 'surely wrong' ratings was decreased [$F(1,63) = 26.276$, $p < .001$; $\beta = -0.008$, $SE = 0.002$].

For correct responses, the ANOVA also revealed a main effect of confidence [$X^2(3) = 167.472$, $p < .001$]. Correct responses were most often rated as 'surely correct' (84.2 %) and least often as 'surely wrong' (0.7 %). Again, linear regression analyses on age-related differences showed that the proportion of 'maybe correct' ratings was increased with higher age [$F(1,63) = 24.653$, $p < .001$; $\beta = 0.006$, $SE = 0.001$], and the proportion of 'surely correct' ratings was decreased with age [$F(1,63) = 24.815$, $p < .001$; $\beta = -0.006$, $SE = 0.001$; Figure 2B].

As mentioned above, to ensure a sufficient number of trials for each level of confidence for each participant, we combined 'maybe wrong' and 'maybe correct' ratings into one category representing 'unsure' responses. Thus, for all following behavioural analyses including the factor confidence, the reported analyses use three confidence levels.

3.1.3 Error rate (ER)

The average error rate was 15.6 ± 1.6%, and the mixed effects regression model testing for effects of confidence and age on error rate showed that the error rate significantly increased with higher age [$X^2(1) = 4.704$, $p = .030$]. The analysis further showed an effect of confidence on error rate [$X^2(2) = 2200.020$, $p < .001$]. The error rate decreased across confidence levels from 94.0 ± 0.7% on trials rated as 'surely wrong' to 67.3 ± 0.8% on trials rated as 'unsure' and 6.6 ± 0.2% on trials rated as 'surely correct'. Pairwise comparisons indicated that all comparisons were statistically significant (all $p < .001$). Thus, on average, participants' confidence reflected their performance well (which further supports the notion that the current study's confidence scale was a meaningful assessment tool). Furthermore, the regression analysis revealed a significant interaction between confidence and age [$X^2(2) = 168.125$, $p < .001$]. In subsequent mixed effects regression analyses for each level of confidence, error rates only significantly increased with higher age for the 'surely correct' confidence level [$X^2(1) = 37.664$, $p < .001$].

3.1.4 Response time (RT)

A mixed effects regression model predicting RT and testing for the effects of accuracy and age showed a significant effect of accuracy [$F(1,61.6) = 5.572$, $p = .021$] with on average slower errors (752.3 ± 3.8 ms) than correct responses (716.5 ± 1.3 ms). Moreover, the model revealed the expected slowing with age [$F(1,62.9) = 17.358$, $p < .001$], which did not significantly differ between errors and correct responses.

The mixed effects regression with the within-subject factor confidence similarly revealed an age-related slowing [$F(1,63.7) = 13.305$, $p < .001$; Figure 4A]. Moreover, the analysis revealed an effect of confidence [$F(2,61.8) = 27.291$, $p < .001$] and a significant interaction between confidence and age [$F(2,56.6) = 5.187$, $p = .009$]. Pairwise comparisons indicated that all pairs were statistically significantly different (all $p < .010$), with trials associated with the 'unsure'

confidence level (815.6 ± 6.7 ms) being considerably slower than trials rated as 'surely correct' (702.1 ± 1.3 ms) or 'surely wrong' (736.6 ± 6.7 ms). Furthermore, trials were significantly slower with older age for the extreme ratings ['surely wrong': $F(1,68.4) = 13.592$, $p < .001$; 'surely correct': $F(1.62.4) = 18.358$, $p < .001$], but not for 'unsure' ratings.

In short, RT was associated with confidence, such that high certainty (i.e. 'surely correct/wrong') was associated with the fastest responses, and this confidence-related modulation of RT decreased with higher age.

3.1.5 Confidence

A linear mixed effects regression model predicting confidence (coded from 1 to 3) across all trials revealed a significant effect of accuracy [i.e. error vs. correct trials; $F(1,63.4) = 162.928$, $p < .001$] and a significant interaction between accuracy and age [$F(1,62.4) = 37.361$, $p < .001$], but no significant effect of age. The average confidence rating was lower for errors (1.991 ± 0.014) compared to correct responses (2.867 ± 0.003). Follow-up regression analyses predicting confidence as a function of age for errors and correct responses separately revealed that confidence increased with age for errors [$F(1,60.1) = 17.977$, $p < .001$], while for correct responses it decreased [$F(1,62.1) = 23.816$, $p < .001$; Figure 3B].

3.1.6 Metacognitive accuracy (Phi)

Phi had a mean of 0.579 ± 0.027 across the entire sample and was significantly predicted by age at the group level [$F(1,63) = 32.206$, $p < .001$; $\beta = -0.008$, $SE = 0.001$], indicating a decrease of metacognitive accuracy with age (Figure 3A). Moreover, a multiple linear regression including the additional factor of error rate did not show a significant interaction with age ($p = .535$), suggesting that the association between metacognitive accuracy and age was not affected by decreased task performance in older adults. Similarly, a multiple linear regression including the additional factor of d2-test scores (which provide a task-independent measure of attention)

suggested that the decrease in Phi with age was also independent of an age-related reduction in attentional capacity (interaction: $p = .091$).

3.1.7 Behavioural adaptation

To investigate the effect of accuracy and confidence in a given trial on the behaviour in the following trial, we computed response caution as the product of accuracy (coded as 0 and 1) and RT in the subsequent trial. The mixed effects regression with the within-subject factor accuracy (referring to the previous trial) revealed a significant effect of accuracy [$F(1,55.9) = 12.366$, $p < .001$] and an interaction between accuracy and age [$F(1,43.9) = 6.709$, $p = .013$], but no significant effect of age. Follow-up regression analyses for the subsets of errors or correct responses showed a nominal decrease in response caution with age for errors, but neither this nor the effect of age for correct responses was significant. Thus, these findings indicate that participants were on average more cautious after errors than after correct responses, and this effect did not significantly vary across age.

Next, we examined whether the response caution in the subsequent trial could also be predicted by the confidence rating in the preceding trial. As shown above, confidence and accuracy are strongly related; however, a significant modulation by confidence could also indicate that this internal confidence signal drives behavioural adaptations. The mixed effects regression on response caution with the within-subject factor confidence (referring to the previous trial) indeed revealed an effect of confidence [$F(2,54.9) = 7.306$, $p = .002$], but again, no effect of age and also no significant interaction (Figure 4B). Pairwise comparisons between the confidence levels showed that the response caution after trials rated as ‘surely correct’ was significantly lower compared to trials rated as ‘unsure’ or as ‘surely wrong’.

To summarise the effects of ageing on behaviour, we found the expected age-related general increase in error rates and response times, accompanied by a decrease in metacognitive ability,

which was mainly reflected in reduced use of confidence ratings at the extreme ends of the scale but more indications of being unsure. Response caution, on the other hand, was not affected by ageing. Caution increased after errors compared to correct responses, and was notably specifically modulated by previous trial confidence. With higher confidence, the response caution in the subsequent trial decreased.

3.2 Electrophysiological results

3.2.1 $N_{e/c}$ amplitudes

The mean amplitude of the $N_{e/c}$ was significantly larger for errors compared to correct responses, as reflected in an effect of accuracy in the mixed effects regression predicting the $N_{e/c}$ as a function of accuracy and age [$F(1,38.6) = 9.054$, $p = .005$; Figure 5A]. There was no main effect of age, but a significant interaction [$F(1,31.8) = 5.472$, $p = .026$] as the amplitude of the N_e [$F(1,55.4) = 5.030$, $p = .029$] but not the N_c was smaller with higher age.

For the analysis of confidence, we fitted separate linear mixed effects models to the N_e amplitude for errors and to the N_c amplitude for correct responses, with confidence as the within-subject factor and age as the between-subject factor. The regression analysis for errors showed effects of age [$F(1,57.4) = 4.068$, $p = .048$], confidence [$F(2,2706.4) = 4.007$, $p = .018$], and a significant interaction between age and confidence [$F(2,2731.5) = 3.662$, $p = .026$; Figure 6A]. Pairwise comparisons between the confidence levels indicated a significant difference between errors rated as ‘surely wrong’ and ‘surely correct’, and follow-up mixed effects regressions showed that specifically the N_e amplitudes of low confidence errors (i.e. rated as ‘surely wrong’) was decreased with older age [$F(1,58.1) = 9.735$, $p = .003$].

The regression analysis for correct responses with the within-subject factor confidence yielded no significant effects (Figure 6B).

3.2.2 $P_{e/c}$ amplitudes

The mixed effects regression on the $P_{e/c}$ amplitude with the within-subject factor accuracy revealed a significant effect of accuracy with larger amplitudes for errors compared to correct responses [$F(1,55.3) = 10.378$, $p = .002$; Figure 5B]. There was no effect of age, but a significant interaction between accuracy and age [$F(1,49.2) = 6.443$, $p = .014$]. However, in follow-up regression analyses, no significant associations were found for errors or correct responses.

Next, responses were again split by their accuracy, and separate linear mixed effects models were fitted to the P_e and P_c amplitudes, respectively, with the within-subject factor confidence. Neither the analysis for errors nor the analysis for correct responses yielded any significant effects on the P_e and P_c amplitudes (Figure 6C & D).

However, due to previous evidence suggesting a strong relation between $P_{e/c}$ amplitude and error detection or confidence ratings (Boldt and Yeung, 2015; Nieuwenhuis et al., 2001), we were specifically interested in the modulation of the P_e by confidence. To replicate previous findings, we fitted an additional, exploratory mixed effects model to the P_e amplitudes, including only the factor of confidence. The analysis revealed a significant, albeit small difference in P_e amplitude between errors rated as 'surely wrong' and errors rated as 'surely correct' [$F(2,2723.6) = 6.627$, $p = .001$], as confirmed in follow-up multiple comparisons between confidence levels ($t = 3.617$, $p < .001$). This exploratory analysis implies that the P_e was modulated by confidence when assessed independent of age.

4. Discussion

We conducted a complex four-choice flanker task with adult participants covering an age range from 20 to 76 years, allowing us to investigate confidence and metacognitive accuracy as well as neural indices thereof across the lifespan. We found that error rates and response times (RT) increased with age. Metacognitive accuracy, quantified as Phi, gradually decreased across the

lifespan and was characterised by differential use of confidence ratings. In contrast, we did not find differences between younger and older adults in the ability to adapt behaviour in accordance with reported confidence. As expected, the $N_{e/c}$ and $P_{e/c}$ amplitudes declined with higher confidence in having made a correct response, which was specifically observed for trials with response errors. While the N_e amplitude was smaller with older age whenever participants were sure they made an error, the variation in the P_e amplitude with reported confidence was surprisingly not affected by ageing. In the following, we will first discuss potential processes underlying age-related differences in metacognitive accuracy and their relation to task performance and confidence, before comparing the pattern we observed at the behavioural level to the patterns we observed in the ERPs. Finally, we argue that older adults' preserved ability to adapt their behaviour to their perceived confidence could be related to the $P_{e/c}$ amplitude.

4.1 Differential use of confidence scale as a marker of age-related metacognitive decline

In the present study, metacognitive accuracy (Phi) was reduced with increasing age. This is consistent with the findings of Palmer et al. (2014) who used a metacognitive efficiency measure, which further considered the individual performance in their perceptual discrimination task. As this measure was not directly applicable in our four-choice flanker task, we confirmed (by calculating multiple linear regressions taking into account the error rate and the d2-test score) that the observed decline in metacognitive accuracy was not merely a reflection of general age-related performance or attention deficits (d2-test; see also Larson & Clayson, 2011). Our results, therefore, show that Palmer et al.'s (2014) findings also hold for a more complex, speeded decision task, which was not based on stimulus ambiguity.

The question remains as to how the age-related differences in confidence emerge. Given the nature of Phi, a smaller value could either indicate more undetected errors or correct responses rated as being incorrect, or a generally higher uncertainty (i.e. rating all correct responses as 'maybe

correct' will result in a lower Phi value than rating the same number of correct responses as 'surely correct'). Indeed, we observed that older adults used the extreme ends of the confidence scale considerably less often than younger adults.

For errors, this pattern resulted in a higher mean confidence with age. This disproportional rise in reported confidence has similarly been shown in error detection studies, indicated by a lower error detection rate in older adults (Harty et al., 2017, 2013; Niessen et al., 2017). For correct decisions, we observed a *lower* mean confidence due to the tendency of the older adults to use the middle of the confidence scale, whereas previous studies rather reported an *over*-confidence in older age (Dodson et al., 2007; Hansson et al., 2008; Ross et al., 2012).

Interestingly, participants in our study responded slowest in case of uncertainty, i.e. 'unsure' ratings. In contrast, studies on decision confidence typically report increasing RT with decreasing confidence (Kiani et al., 2014; Rahnev et al., 2020; Weidemann and Kahana, 2016). Most of these studies specifically measured confidence in having made a correct decision (i.e. the lowest confidence indicates guessing, while in our study it indicates high certainty in being incorrect), and typical paradigms in these studies are two-choice signal detection tasks in which the degree of sensory evidence, for instance, perceptual discriminability is manipulated (Kiani et al., 2014; Moran et al., 2015; Rollwage et al., 2020). In our task, we ensured (using a designated colour discrimination test) that all stimuli were perceptually discriminable without time pressure, and our data showed no signs of age-related differences in stimulus processing (even though it remains possible that slight impairments in colour perception, or other untested factors such as attention, working memory, etc., might have contributed to the age-related slowing we observed; see supplementary material S4). Instead, potential sources for errors could be, for instance, stimulus conflict caused by the flankers and the similarity of the stimulus colours, or difficulties in remembering the stimulus response mapping. Using a comparable paradigm, Stahl et al. (2020)

found slow errors to be associated with lower confidence than fast, impulsive errors and inferred that those error types should predominantly be caused by weak stimulus-response representations (i.e. due to weak memory traces).

As such conclusions could not be drawn from classical error processing studies requiring only a binary error detection rating, our findings provide an important link between those and decision confidence studies. In a typical error processing paradigm that posed higher demands on the older adults (as indicated, for instance, by higher error rates), our results could be interpreted as their impaired metacognitive evaluation (assessed via confidence ratings) being partly related to more frequent memory-related errors, which appear to be more challenging to assess consciously (Maier and Steinhauser, 2017; Stahl et al., 2020).

4.2 Neural correlate of confidence is stable across age

The $P_{e/c}$ is an established marker of metacognition, reflecting variations in subjective error awareness and decision confidence (Boldt and Yeung, 2015; Nieuwenhuis et al., 2001). In the present study, the $P_{e/c}$ showed the well-known accuracy effect of larger amplitudes for errors than correct responses. Moreover, we could replicate prior findings of the P_e increasing with decreasing confidence, - for the first time - for a very broad age range (Boldt and Yeung, 2015; Rausch et al., 2019). This also replicates findings from error detection studies showing increased P_e amplitudes for detected compared to undetected errors (Endrass et al., 2012a; Nieuwenhuis et al., 2001).

The main interest of our study was to investigate the modulation of the $P_{e/c}$ by metacognition in the context of healthy ageing. Remarkably, the $P_{e/c}$ amplitude did not show an overall reduction with age, nor a differential modulation by confidence across the lifespan, suggesting that the accumulation of error evidence was well preserved in older age. This is contrary to the error detection literature (Harty et al., 2017; Niessen et al., 2017). Since these studies did not assess confidence on multiple levels, participants did not have the chance to express uncertainty.

Assuming more 'unsure' cases with older age, their observed age-related decrease in P_e amplitude for detected errors might thus be confounded, as higher uncertainty was generally associated with reduced P_e amplitudes (Boldt and Yeung, 2015). Following this logic, it is also possible to explain the lack of a significant age-related modulation of the $P_{e/c}$ amplitude in the present study: If older adults' internal threshold for rating an error as 'surely wrong' was generally raised, the errors that *were* rated as 'surely wrong' should be trials with particularly high P_e amplitudes, as they were absolutely sure of having committed an error. As a result, a putative age-related decrease in the P_e amplitude of low confidence errors could be masked in our data, because the same reported rating levels might reflect a different sense of confidence for younger and older adults. Thus, the current pattern of results suggests that the P_e amplitude does *not* serve as a direct index of metacognitive accuracy across participants, but rather reflects the degree of confidence, irrespective of objective performance (Di Gregorio et al., 2018; Larson and Clayson, 2011; Pouget et al., 2016; Stahl et al., 2020).

4.3 Impaired neural processing of conflict modulates metacognitive decline

The marked behavioural decline in older adults' metacognitive accuracy was not mirrored in age-related variations of the $P_{e/c}$ amplitude, but rather in a differential modulation of the N_e across the lifespan. The modelling results revealed that the N_e amplitude was also affected by the interaction between confidence and age. With older age, the N_e declined for all errors in which high conflict was perceived. In other words, only the N_e amplitude of errors which were rated as 'surely wrong' varied in amplitude across the lifespan. As the $N_{e/c}$ is sensitive to conflict between the given and the actual correct response, older adults seemed to having had difficulties internally representing the correct response in highly conflicting situation (Yeung et al., 2004). Notably, this effect was error-specific, that is, we cannot draw conclusions about internal processes for correct

responses, as the N_c amplitude did not show a relation to confidence that could have varied with age.

We suggest that the reduced N_e amplitude of low confidence errors with higher age could be related to the observed decrease in metacognitive accuracy in our flanker task. If older adults did not perceive high conflict due to difficulties in forming an accurate internal representation of the correct response, this information was necessarily missing for the metacognitive evaluation. Thus, the impaired neural integration of conflict detection and confidence could have led to the observed behavioural difficulties matching confidence ratings and objective accuracy.

4.4 Adults of all ages base future behaviour on subjective confidence

Ultimately, proper metacognitive evaluation should improve behaviour. Interestingly, response caution was not only enhanced after errors, but we also found evidence that it was modulated by the reported confidence in the preceding trial. Given that the participants did not receive any external feedback about the accuracy of their response (as it is often the case in real-life decisions), it seems plausible that they used their best available estimate, i.e. the subjective sense of confidence, to regulate subsequent behaviour (Desender et al., 2019a). Specifically, low confidence (reflecting a belief that an error had been committed) or uncertainty about a decision were associated with higher response caution in the subsequent trial. Possibly, participants sought more evidence before committing to their next decision, leading to slower but more accurate responses (Desender et al., 2019a, 2019b).

Translating our findings to error detection studies, the increase in response caution with lower previous trial confidence converges with findings of error detection studies reporting increased slowing (i.e. a sign of behavioural adaptation) after detected compared to undetected errors (Nieuwenhuis et al., 2001; Stahl et al., 2020; Wessel et al., 2018; for a review on post-error adjustments see Danielmeier & Ullsperger, 2011).

Notably, response caution was similarly affected by accuracy and confidence across the lifespan. Thus, while metacognitive accuracy was reduced in older age, a neural correlate of error confidence magnitude, the P_e amplitude, and the behavioural adaptations relative to the reported confidence were consistent across the lifespan. This suggests that it is the perceived confidence that shapes future behaviour, irrespective of metacognitive accuracy: Despite their failure in matching confidence to task performance, older adults seem to be equally able to use internal states of confidence to change future behaviour adaptively.

4.5 Limitations and implications

One limitation of the present study is the number of participants retained for the analyses. When designing the experiment, we tried to find an optimal balance between task difficulty, feasibility for all ages, and gaining many trials while ensuring that especially older adults were not exhausted at the end of the experiment. However, the combination of a substantial number of response alternatives, time pressure, and discriminability of stimuli was demanding, leading to an undesirably large number of participants to be excluded from the analyses (17 of the initial 82 participants).

A second shortcoming is the confined number of trials available for analysis after defining conditions of interest. Due to an unforeseen highly skewed use of the confidence scale, it was impossible to apply a factorial design while retaining four distinct confidence levels. In particular for correct trials, the variance in confidence ratings was low, which is a common problem in metacognition research (for a review, see Wessel, 2012). However, the application of linear mixed effects modelling provided us with a powerful tool that can account for varying trial numbers across participants and importantly, the multi-level structure of our data.

Nevertheless, our findings provide important insights into ageing effects on metacognition, integrating approaches from error detection and decision confidence research. In contrast to the

metacognitive evaluation itself, the effect of confidence on subsequently adapting response caution was well preserved in older adults. Thus, training the metacognitive evaluation of fundamental decisions in older adults might constitute a promising endeavour (and has been shown to work for mathematical problem solving [Pennequin et al., 2010]).

5. Conclusion

The study of error detection and confidence in the context of healthy ageing have advanced largely in parallel. Our study demonstrates that confidence shapes our behavioural and neural processing of decisions and should be considered to investigate age-related effects on error processing and metacognitive abilities. Interestingly, the N_e, but not the P_e amplitude was differentially modulated by confidence across the lifespan, suggesting that the decreasing accuracy of metacognitive judgements with older age might be related to impaired integration of neural correlates of conflict detection and decision confidence.

Disclosure statement

The authors declare no conflict of interest.

Acknowledgements

We thank all colleagues from the Institute of Neuroscience and Medicine (INM-3), Cognitive Neuroscience, and the Decision Neuroscience Lab at the Melbourne School of Psychological Sciences for valuable discussions and their support.

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation – Project-ID 431549029 – SFB 1451; GRF and PHW), and the Australian Research Council (Discovery Project Grant; DP160103353; SB).

References

Bates, D., Mächler, M., Boker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using {lme4}. J. Stat. Softw. 67, 1–48. https://doi.org/10.18637/jss.v067.i01

Boldt, A., Yeung, N., 2015. Shared neural markers of decision confidence and error detection. J. Neurosci. 35, 3478–3484. https://doi.org/10.1523/JNEUROSCI.0797-14.2015

Brickenkamp, R., 2002. The d2 Test of attention. (1st ed.), 9th ed. Verlag für Psychologie Hogrefe, Goettingen.

Castel, A.D., Middlebrooks, C.D., McGillivray, S., 2016. Monitoring Memory in Old Age: Impaired, Spared, and Aware., in: Dunlosky, J., Tauber, S.K. (Eds.), The Oxford Handbook of Metamemory. Oxford University Press, pp. 519–534. https://doi.org/10.1093/oxfordhb/9780199336746.013.3

Chua, E., Schacter, D.L., Sperling, R., 2009. Neural basis for recognition confidence. Psychol. Aging 24, 139–153. https://doi.org/10.1037/a0014029.Neural

Clawson, A., Clayson, P.E., Keith, C.M., Catron, C., Larson, M.J., 2017. Conflict and performance monitoring throughout the lifespan: An event-related potential (ERP) and temporospatial component analysis. Biol. Psychol. 124, 87–99. https://doi.org/10.1016/j.biopsycho.2017.01.012

Clayson, P.E., Baldwin, S.A., Larson, M.J., 2013. How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. Psychophysiology 50, 174–186. https://doi.org/10.1111/psyp.12001

Danielmeier, C., Ullsperger, M., 2011. Post-error adjustments. Front. Psychol. 2, 1–10.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. 134, 9–21.

Desender, K., Boldt, A., Verguts, T., Donner, T.H., 2019a. Confidence predicts speed-accuracy tradeoff for subsequent decisions. Elife. https://doi.org/10.1101/466730

Desender, K., Murphy, P.R., Boldt, A., Verguts, T., Yeung, N., 2019b. A Postdecisional Neural Marker of Confidence Predicts Information-Seeking in Decision-Making. J. Neurosci. 39, 3309–3319. https://doi.org/10.1101/433276

Di Gregorio, F., Maier, M.E., Steinhauser, M., 2018. Errors can elicit an error positivity in the absence of an error negativity: Evidence for independent systems of human error monitoring. Neuroimage 172, 427–436. https://doi.org/10.1016/j.neuroimage.2018.01.081

Dodson, C.S., Bawa, S., Krueger, L.E., 2007. Aging, metamemory, and high-confidence errors: A misrecollection account. Psychol. Aging 22, 122–133. https://doi.org/10.1037/0882-7974.22.1.122

Dully, J., McGovern, D.P., O'Connell, R.G., 2018. The impact of natural aging on computational and neural indices of perceptual decision making: A review. Behav. Brain Res. 355, 48–55. https://doi.org/10.1016/j.bbr.2018.02.001

Endrass, T., Klawohn, J., Preuss, J., Kathmann, N., 2012a. Temporospatial dissociation of Pe subcomponents for perceived and unperceived errors. Front. Hum. Neurosci. 6, 1–10. https://doi.org/10.3389/fnhum.2012.00178

Endrass, T., Schreiber, M., Kathmann, N., 2012b. Speeding up older adults: Age-effects on error processing in speed and accuracy conditions. Biol. Psychol. 89, 426–432. https://doi.org/10.1016/j.biopsycho.2011.12.005

Eriksen, B.A., Eriksen, C.W., 1974. Effects of noise letters upon the identification of a target letter in a nonsearch task. Percept. Psychophys. 16, 143–149.

Falkenstein, M., Hohnsbein, J., Hoormann, J., Blanke, L., 1991. Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. Electroencephalogr. Clin. Neurophysiol. 78, 447–455. https://doi.org/10.1016/0013-4694(91)90062-9

Falkenstein, M., Hoormann, J., Christ, S., Hohnsbein, J., 2000. ERP components on reaction errors and their functional significance: A tutorial. Biol. Psychol. 51, 87–107. https://doi.org/10.1016/S0301-0511(99)00031-9

Falkenstein, M., Hoormann, J., Hohnsbein, J., 2001. Changes of error-related ERPs with age. Exp. Brain Res. 138, 258–262. https://doi.org/10.1007/s002210100712

Fitzgerald, L.M., Arvaneh, M., Dockree, P.M., 2017. Domain-specific and domain-general processes underlying metacognitive judgments. Conscious. Cogn. 49, 264–277. https://doi.org/10.1016/j.concog.2017.01.011

Fleming, S.M., Dolan, R.J., 2012. The neural basis of metacognitive ability. royalsocietypublishing.org 367, 1338–1349. https://doi.org/10.1098/rstb.2011.0417

Fleming, S.M., Dolan, R.J., Frith, C.D., 2012. Metacognition: Computation, biology and function. Philos. Trans. R. Soc. B Biol. Sci. 367, 1280–1286. https://doi.org/10.1098/rstb.2012.0021

Fleming, S.M., Frith, C.D., 2014. The Cognitive Neuroscience of Metacognition, The Cognitive Neuroscience of Metacognition. Springer. https://doi.org/10.1007/978-3-642-45190-4

Fleming, S.M., Lau, H.C., 2014. How to measure metacognition. Front. Hum. Neurosci. 8, 1–9. https://doi.org/10.3389/fnhum.2014.00443

Folstein, M.F., Folstein, S.E., McHugh, P.R., 1975. “Mini-mental state”. A practical method for grading the cognitive state of patients for the clinician. J. Psychiatr. Res. 12, 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

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. Psychon. Bull. Rev. 10, 843–876. https://doi.org/10.3758/BF03196546

Groom, M.J., Cragg, L., 2015. Differential modulation of the N2 and P3 event-related potentials by response conflict and inhibition. Brain Cogn. 97, 1–9. https://doi.org/10.1016/j.bandc.2015.04.004

Hansson, P., Rönnlund, M., Juslin, P., Nilsson, L.G., 2008. Adult Age Differences in the Realism of Confidence Judgments: Overconfidence, Format Dependence, and Cognitive Predictors. Psychol. Aging 23, 531–544. https://doi.org/10.1037/a0012782

Harty, S., Murphy, P.R., Robertson, I.H., O’Connell, R.G., 2017. Parsing the neural signatures of reduced error detection in older age. Neuroimage 161, 43–55. https://doi.org/10.1016/j.neuroimage.2017.08.032

Harty, S., O’Connell, R.G., Hester, R., Robertson, I.H., 2013. Older adults have diminished awareness of errors in the laboratory and daily life. Psychol. Aging 28, 1032–1041. https://doi.org/10.1037/a0033567

Hautzinger, M., 1991. The Beck Depression Inventory in clinical practice. Nervenarzt 62, 689–96.

Hertzog, C., 2015. Aging and Metacognitive Control, The Oxford Handbook of Metamemory. https://doi.org/10.1093/oxfordhb/9780199336746.013.31

Hertzog, C., Hultsch, D., 2000. Metacognition in adulthood and old age., in: Craik, F.I.M., Salthouse, T.A. (Eds.), The Handbook of Aging and Cognition. Lawrence Erlbaum Associates Publishers, pp. 417–466.

Hoerold, D., Pender, N., Robertson IH, 2013. mende. Neuropsychologia 51, 385–391.

Jasper, H., 1958. The ten twenty electrode system of the international federation. Electroencephalogr. Clin. Neurophysiol. 10, 371–375.

Kayser, J., Tenke, C.E., 2006. Principal components analysis of Laplacian waveforms as a generic method for identifying ERP generator patterns: II. Adequacy of low-density estimates. Clin. Neurophysiol. 117, 369–380. https://doi.org/10.1016/j.clinph.2005.08.033

Kiani, R., Corthell, L., Shadlen, M.N., 2014. Choice certainty is informed by both evidence and decision time. Neuron 84, 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

Kornell, N., Son, L.K., Terrace, H.S., 2007. Transfer of metacognitive skills and hint seeking in monkeys: Research article. Psychol. Sci. 18, 64–71. https://doi.org/10.1111/j.1467-9280.2007.01850.x

Korsch, M., Frühholz, S., Herrmann, M., 2016. Conflict-specific aging effects mainly manifest in early information processing stages-an ERP study with different conflict types. Front. Aging Neurosci. 8, 1–12. https://doi.org/10.3389/fnagi.2016.00053

Larson, M.J., Clayson, P.E., 2011. The relationship between cognitive performance and electrophysiological indices of performance monitoring. Cogn. Affect. Behav. Neurosci. 11, 159–171. https://doi.org/10.3758/s13415-010-0018-6

Larson, M.J., Clayson, P.E., Keith, C.M., Hunt, I.J., Hedges, D.W., Nielsen, B.L., Call, V.R.A.V.R.A., 2016. Cognitive control adjustments in healthy older and younger adults: Conflict adaptation, the error-related negativity (ERN), and evidence of generalized decline with age. Biol. Psychol. 115, 50–63. https://doi.org/10.1016/j.biopsycho.2016.01.008

Larson, M.J., South, M., Clayson, P.E., 2011. Sex differences in error-related performance monitoring. Neuroreport 22, 44–48. https://doi.org/10.1097/WNR.0b013e3283427403

Lopez-Calderon, J., Luck, S.J., 2014. ERPLAB: An open-source toolbox for the analysis of event-related potentials. Front. Hum. Neurosci. 8, 1–14. https://doi.org/10.3389/fnhum.2014.00213

Lucci, G., Berchicci, M., Spinelli, D., Taddei, F., Di Russo, F., 2013. The Effects of Aging on Conflict Detection. PLoS One 8. https://doi.org/10.1371/journal.pone.0056566

Maier, M.E., Steinhauser, M., 2017. Working memory load impairs the evaluation of behavioral errors in the medial frontal cortex. Psychophysiology 54, 1472–1482. https://doi.org/10.1111/psyp.12899

Maier, M.E., Steinhauser, M., Hübner, R., 2008. Is the Error-related Negativity Amplitude Related to Error Detectability? Evidence from Effects of Different Error Types. J. Cogn. Neurosci. 20, 2263–2273.

Maniscalco, B., Lau, H., 2012. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious. Cogn. 21, 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Mattler, U., 2003. Delayed flanker effects on lateralized readiness potentials. Exp. Brain Res. 151, 272–288. https://doi.org/10.1007/s00221-003-1486-5

Mecacci, L., Righi, S., 2006. Cognitive failures, metacognitive beliefs and aging. Pers. Individ. Dif. 40, 1453–1459. https://doi.org/10.1016/j.paid.2005.11.022

Moran, R., Teodorescu, A.R., Usher, M., 2015. Post choice information integration as a causal determinant of

confidence: Novel data and a computational account. Cogn. Psychol. 78, 99–147. https://doi.org/10.1016/j.cogpsych.2015.01.002

Murphy, P.R., Robertson, I.H., Harty, S.S., O'Connell, R.G., O'Connell, R.G., 2015. Neural evidence accumulation persists after choice to inform metacognitive judgments. Elife 4, 1–23. https://doi.org/10.7554/eLife.11946

Nelson, T., 1984. A comparison of current measures of the accuracy of feeling-of-knowing predictions. Psychol. Bull. 95, 109–133.

Niessen, E., Fink, G.R., Hoffmann, H.E.M.M., Weiss, P.H., Stahl, J., 2017. Error detection across the adult lifespan: Electrophysiological evidence for age-related deficits. Neuroimage 152, 517–529. https://doi.org/10.1016/j.neuroimage.2017.03.015

Nieuwenhuis, S., Richard Ridderinkhof, K., Blom, J., Band, G.P.H., Kok, A., 2001. Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. Psychophysiology 38, 752–760. https://doi.org/10.1017/S0048577201001111

Oldfield, R.C., 1971. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia 9, 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Palmer, E.C., David, A.S., Fleming, S.M., 2014. Effects of age on metacognitive efficiency. Conscious. Cogn. 28, 151–160. https://doi.org/10.1016/j.concog.2014.06.007

Pennequin, V., Sorel, O., Mainguy, M., 2010. Metacognition, executive functions and aging: The effect of training in the use of metacognitive skills to solve mathematical word problems. J. Adult Dev. 17, 168–176. https://doi.org/10.1007/s10804-010-9098-3

Perrin, F., Pernier, J., Bertrand, O., Echallier, J.F., 1989. Spherical splines for scalp potential and current density mapping. Electroencephalogr. Clin. Neurophysiol. 72, 184–187. https://doi.org/10.1016/0013-4694(89)90180-6

Polich, J., 2007. Updating P300: an integrative theory of P3a and P3b. Clin. Neurophysiol. 118, 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Pouget, A., Drugowitsch, J., Kepecs, A., 2016. Confidence and certainty: distinct probabilistic quantities for different goals. nature.com. https://doi.org/10.1038/nn.4240

Rabbitt, P., 1990. Age, IQ and awareness, and recall of errors. Ergonomics 33, 1291–1305. https://doi.org/10.1080/00140139008925333

Rabbitt, P.M.A., 1966. Errors and error correction in choice-response tasks. Exp. Psychol. 71, 264–272. https://doi.org/10.1037/h0022853

Rahnev, D., Desender, K., Lee, A.L.F., Adler, W.T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L.Y., Balcı, F., Bang, J.W., Bègue, I., Birney, D.P., Brady, T.F., Calder-Travis, J., Chetverikov, A., Clark, T.K., Davranche, K., Denison, R.N., Dildine, T.C., Double, K.S., Duyan, Y.A., Faivre, N., Fallow, K., Filevich, E., Gajdos, T., Gallagher, R.M., de Gardelle, V., Gherman, S., Haddara, N., Hainguerlot, M., Hsu, T.Y., Hu, X., Iturrate, I., Jaquiery, M., Kantner, J., Koculak, M., Konishi, M., Koß, C., Kvam, P.D., Kwok, S.C., Lebreton, M., Lempert, K.M., Ming Lo, C., Luo, L., Maniscalco, B., Martin, A., Massoni, S., Matthews, J., Mazancieux, A., Merfeld, D.M., O'Hora, D., Palser, E.R., Paulewicz, B., Pereira, M., Peters, C., Philiastides, M.G., Pfuhl, G., Prieto, F., Rausch, M., Recht, S., Reyes, G., Rouault, M., Sackur, J., Sadeghi, S., Samaha, J., Seow, T.X.F., Shekhar, M., Sherman, M.T., Siedlecka, M., Skóra, Z., Song, C., Soto, D., Sun, S., van Boxtel, J.J.A., Wang, S., Weidemann, C.T., Weindel, G., Wierzchoń, M., Xu, X., Ye, Q., Yeon, J., Zou, F., Zylberberg, A., 2020. The Confidence Database. Nat. Hum. Behav. 4, 317–325. https://doi.org/10.1038/s41562-019-0813-1

Rausch, M., Zehetleitner, M., Steinhauser, M., Maier, M.E., 2019. Cognitive modelling reveals distinct electrophysiological markers of decision confidence and error monitoring. bioRxiv 860379. https://doi.org/10.1101/860379

Rollwage, M., Loosen, A., Hauser, T.U., Moran, R., Dolan, R.J., Fleming, S.M., 2020. Confidence drives a neural confirmation bias. Nat. Commun. 11. https://doi.org/10.1038/s41467-020-16278-6

Ross, L.A., Dodson, J., Edwards, J.D., Ackerman, M.L., Ball, K., 2012. Self-rated Driving and Driving Safety in Older Adults. Accid. Anal. Prev. 48, 523–527. https://doi.org/10.1038/jid.2014.371

Ruitenberg, M.F.L., Abrahamse, E.L., De Kleine, E., Verwey, W.B., 2014. Post-error slowing in sequential action: An aging study. Front. Psychol. 5, 1–8. https://doi.org/10.3389/fpsyg.2014.00119

Scheffers, M.K., Coles, M.G.H., 2000. Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors. J. Exp. Psychol. Hum. Percept. Perform. 26, 141–151. https://doi.org/10.1037/0096-1523.26.1.141

Schreiber, M., Pietschmann, M., Kathmann, N., Endrass, T., 2011. ERP correlates of performance monitoring in elderly. Brain Cogn. 76, 131–139. https://doi.org/10.1016/j.bandc.2011.02.003

Shekhar, M., Rahnev, D., 2020. The Nature of Metacognitive Inefficiency in Perceptual Decision Making. Psychol. Rev. 1–83. https://doi.org/10.1037/rev0000249

Sim, J., Brown, F., O'Connell, R., Hester, R., 2020. Impaired error awareness in healthy older adults: an age group comparison study. Neurobiol. Aging. https://doi.org/10.1016/j.neurobiolaging.2020.08.001

Siswandari, Y., Bode, S., Stahl, J., 2019. Performance monitoring beyond choice tasks: The time course of force execution monitoring investigated by event-related potentials and multivariate pattern analysis. Neuroimage 197, 544–556. https://doi.org/10.1016/j.neuroimage.2019.05.006

Song, C., Kanai, R., Fleming, S.M., Weil, R.S., Schwarzkopf, D.S., Rees, G., 2011. Relating inter-individual differences in metacognitive performance on different perceptual tasks. Conscious. Cogn. 20, 1787–1792. https://doi.org/10.1016/j.concog.2010.12.011

Stahl, J., Mattes, A., Hundrieser, M., Kummer, K., Mück, M., Niessen, E., Porth, E., Siswandari, Y., Wolters, P., Dummel, S., 2020. Neural correlates of error detection during complex response selection: Introduction of a novel eight-alternative response task. Biol. Psychol. 156. https://doi.org/10.1016/j.biopsycho.2020.107969

Steinhauser, M., Yeung, N., 2010. Decision processes in human performance monitoring. J. Neurosci. 30, 15643–15653. https://doi.org/10.1523/JNEUROSCI.1899-10.2010

Team, R.C., 2021. No Title.

Vidal, F., Burle, B., Bonnet, M., Grapperon, J., Hasbroucq, T., 2003. Error negativity on correct trials: a reexamination of available data. Biol. Psychol. 64, 265–282. https://doi.org/10.1016/S0301-0511(03)00097-8

Weidemann, C.T., Kahana, M.J., 2016. Assessing recognition memory using confidence ratings and response times. R. Soc. Open Sci. 3, 150670. https://doi.org/10.1098/rsos.150670

Wessel, J.R., 2012. Error awareness and the error-related negativity: evaluating the first decade of evidence. Front. Hum. Neurosci. 6, 1–16. https://doi.org/10.3389/fnhum.2012.00088

Wessel, J.R., Dolan, K.A., Hollingworth, A., 2018. A blunted phasic autonomic response to errors indexes age-related deficits in error awareness. Neurobiol. Aging 71, 13–20. https://doi.org/10.1016/j.neurobiolaging.2018.06.019

Yeung, N., Botvinick, M.M., Cohen, J.D., 2004. The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol. Rev. 111, 931–959. https://doi.org/10.1037/0033-295X.111.4.931

Yeung, N., Cohen, J.D., 2006. The Impact of Cognitive Deficits on Conflict MonitoringPredictable Dissociations Between the Error-Related Negativity and N2. Psychol. Sci. 17, 164–171.

Yeung, N., Summerfield, C., 2014. Shared mechanisms for confidence judgements and error detection in human decision making, in: The Cognitive Neuroscience of Metacognition. Springer-Verlag Berlin Heidelberg, pp. 147–167. https://doi.org/10.1007/978-3-642-45190-4_7

Figure captions

Figure 1. (A) The left panel shows an example of a trial in the flanker task, where one central target and two flankers were presented, and the participant had to press the finger that was assigned to the respective target colour (illustrated by the grey arrow). The confidence rating (right panel) consisted of four squares, and the ends of the scale were labelled with the German words for 'surely wrong' on the left and 'surely correct' on the right side. The fingers were mapped onto the four squares according to their spatial location. (B) Colours used in the flanker task. Flanker stimuli could consist of target or neutral colours, whereas target stimuli could only consist of one of the four target colours. (C) Sequence of one trial (here, incongruent). Each trial started with a fixation cross, followed by the presentation of the flankers, to which the target was added shortly after. Then, the screen turned black until a response was registered (maximum 1,200 ms), followed by another blank screen. If a response had been given, the rating scale appeared until a rating was given (maximum 2,000 ms). If no response had been given within the designated time window, the German words for 'too slow' were shown instead. The trial ended with another blank screen for a random intertrial interval.

Figure 2. Distributions of confidence ratings for errors (A) and correct responses (B). Errors were most often rated as 'surely wrong', and correct responses as 'surely correct'. Dots represent the individual proportions of the particular confidence response amongst all errors or correct responses, respectively. A median split by age (*Mdn* = 46) was conducted for illustration purposes. Older adults are shown in green, younger adults in orange. With increasing age, participants used the 'surely correct/wrong' ratings less often, and the middle of the confidence scale more often.

Figure 3. Metacognition across the lifespan. (A) Metacognitive accuracy (Phi) decreased with age. Dots represent means of individual participants. (B) Confidence ratings for errors and correct trials were significantly predicted by age (in years). With increasing age, confidence was reduced for correct responses and increased for errors.

Figure 4. Modulation of response time (RT; A) and response caution (B) by confidence and age (in years). (A) Trials rated as 'unsure' showed slower RTs than trials associated with any of the 'surely' rating categories, and this difference was smaller with increasing age. (B) Adaptation of response caution depending on previous trial confidence rating. Response caution was computed as the product of the accuracy and RT of subsequent trials. Across the lifespan, participants responded less cautiously after higher confidence ratings.

Figure 5. Response-locked event-related potentials for errors and correct responses and topographical maps of errors after current source density transformation. (A) $N_{e/c}$ is computed at electrode FCz and (B) $P_{e/c}$ at electrode Cz. Errors are shown in red, correct trials in black. Scalp topographies depict the mean activity for all error trials averaged across the time windows for the N_e (0-150 ms) and the P_e (150-350 ms). Grey squares indicate time windows for the identification of peak amplitudes, which served to compute the adaptive mean amplitudes.

Figure 6. Regression of response-locked event-related potentials on age (in years) by confidence, separately for errors and correct responses after current source density transformation. Errors are shown in the left panel, correct trials in the right panel. The N_e (A) and N_c (B) are shown at electrode FCz, and P_e (C) and P_c (D) are shown at electrode Cz. For errors, the amplitudes increased with lower confidence, while for correct responses, they were not modulated by confidence. Age predicted a decrease in N_e amplitude of 'surely wrong' errors.

Ms. No.: NBA-21-145
Title: **Neural correlates of metacognition across the adult lifespan**
Corresponding Author: **Miss Helen Overhoff**
Authors: **Yiu Hong Ko; Daniel Feuerriegel; Gereon R. Fink; Jutta Stahl; Peter H. Weiss; Stefan Bode; Eva Niessen**

We are grateful to the Reviewers for their constructive feedback. We hope that we have satisfactorily addressed all points raised by the Reviewers. Changes made to the manuscript are described in detail here and are highlighted in yellow in the manuscript document. Reviewer comments are presented in bold text. Our replies are in standard text, and changes to the manuscript are in italics and quotations. Page numbers corresponding to quoted pieces of text are specified.

Reviewer #1: Comments to the Authors

This study investigated effects of aging on metacognition. In a flanker task, older participants responded slower and less accurately and showed reduced metacognitive capacity. Ne and Pe were reduced with higher confidence that a response was correct, and the Ne, but not the Pe was reduced with older age. Age modulated the relation between the Ne and confidence, but not the relation between the Pe and confidence.

The study is interesting, the paper well-written and a publication would provide a valuable contribution to the literature. I just have some points the Authors should address.

1. Figures 2 and 4: Line graphs of means with standard errors are a bit difficult to see when plotted over the single data points in the same colors especially where the data points become crowded. Could the Authors maybe use different colors for the means and standard errors?

We thank the Reviewer for pointing out a lack of clarity in the figures. Based on the Reviewer's comment, and after implementing major changes in our analysis (i.e. all behavioural and electrophysiological analyses are now conducted at the single-trial level using linear mixed effects models; please see comment 1 by Reviewer #2), we adjusted all figures to better represent the results of the new analysis. Figure 2 now displays the single data points as well as the shapes of the distributions for each condition.

Adjusted Figure 2:

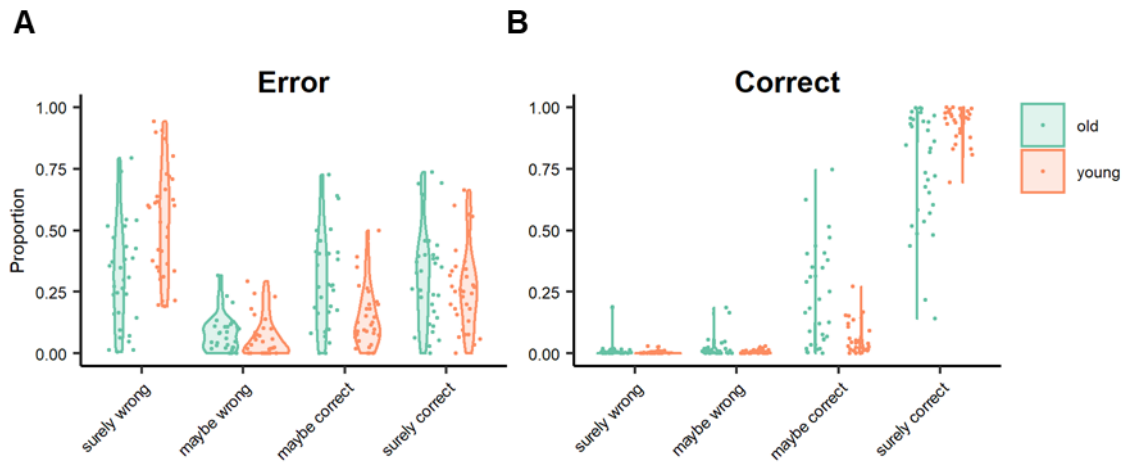


Figure 2. Distributions of confidence ratings for errors (A) and correct responses (B). Errors were most often rated as 'surely wrong', and correct responses as 'surely correct'. Dots represent the individual proportions of the particular confidence response amongst all errors or correct responses, respectively. A median split by age (*Mdn* = 46) was conducted for illustration purposes. Older adults are shown in green, younger adults in orange. With increasing age, participants used the 'surely correct/wrong' ratings less often, and the middle of the confidence scale more often.

Adjusted Figure 4:

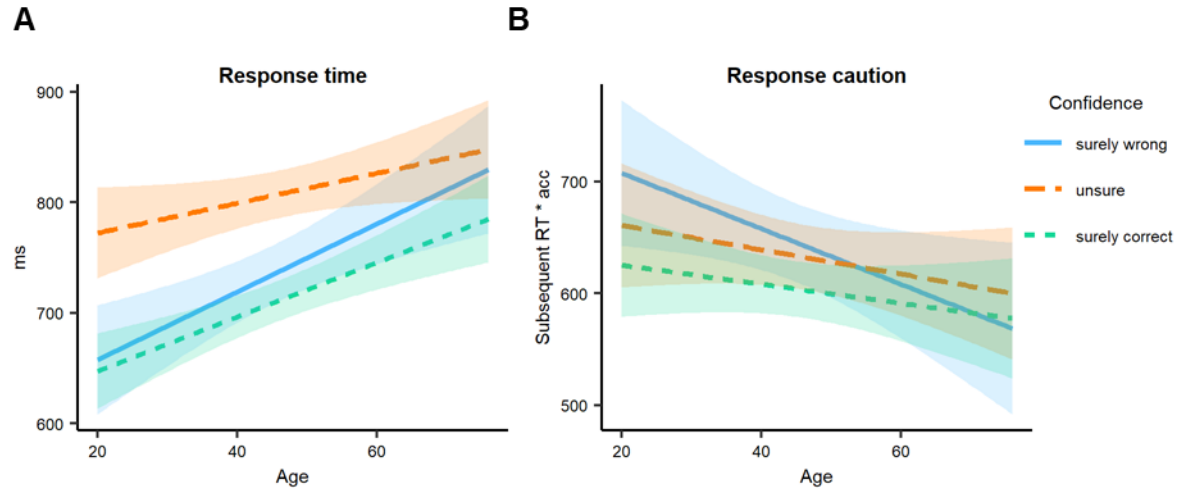


Figure 4. Modulation of response time (RT; A) and response caution (B) by confidence and age (in years). (A) Trials rated as 'unsure' showed slower RTs than trials associated with any of the 'surely' rating categories, and this difference was smaller with increasing age. (B) Adaptation of response caution depending on previous trial confidence rating. Response caution was computed as the product of the accuracy and RT of subsequent trials. Across the lifespan, participants responded less cautiously after higher confidence ratings.

2. I wonder whether some of the results related to age and metacognition (e.g., the negative correlation of metacognitive capacity and age) could be explained by a poorer ability of older participants to discriminate between the colors in the flanker task (which could be an additional reason for why older participants had longer RT and higher error rates in the flanker task). The Authors already made some efforts to make a case against this argument, such as the partial correlations between phi as a measure of metacognitive capacity and age while controlling for error rate (3.1.6). Did the correlation between phi and age also survive controlling for RT?

Due to the changes in our analysis, we are now computing multiple linear regressions instead of partial correlations. Note that these analyses also allowed us to control for RT. The regression analysis including the predictor of RT yielded no significant interaction between age and RT (p = .455), suggesting that the age-related decline in metacognitive accuracy did not moderate the general slowing of responses in older age.

Regarding the discrimination of colours, we cannot rule out the possibility that it took longer for older adults to discriminate them. However, note that we used a colour discrimination test prior to the experiments to ensure that all participants were able to distinguish between the colours. It remains possible that the longer RTs in older age reflect a combination of several (more or less) impaired processes, e.g., selective attention, colour identification, memory retrieval of response mapping, action initiation, etc. Unfortunately, our paradigm does not allow to reveal these processes. We now mention this possibility in the discussion section of the manuscript.

Changes in manuscript (page 21 f.):

"In our task, we ensured (using a designated colour discrimination test) that all stimuli were perceptually discriminable without time pressure, and our data showed no signs of age-related differences in stimulus processing (even though it remains possible that slight impairments in colour perception, or other untested factors such as attention, working memory, etc., might have contributed to the age-related slowing we observed; see supplementary material S4)."

3. Please add x-axes labels to Figures 4, 5 and 6. I noticed this missing especially for the correlation scatter-plots (amplitudes and age) in Figure 6.

Following the Reviewer's advice, we added the missing x-axes labels in all adjusted figures.

4. In the text (3.2.1 & 3.2.2), the reader is referred to Figures 5A & B for the Ne/c analyses and 5C & D for the Pe/c analyses. As Figure 5 only has panels A (Ne/c) and B (Pe/c), something seems to have gone wrong, please double-check.

We have corrected this issue and refer to the correct figures now.

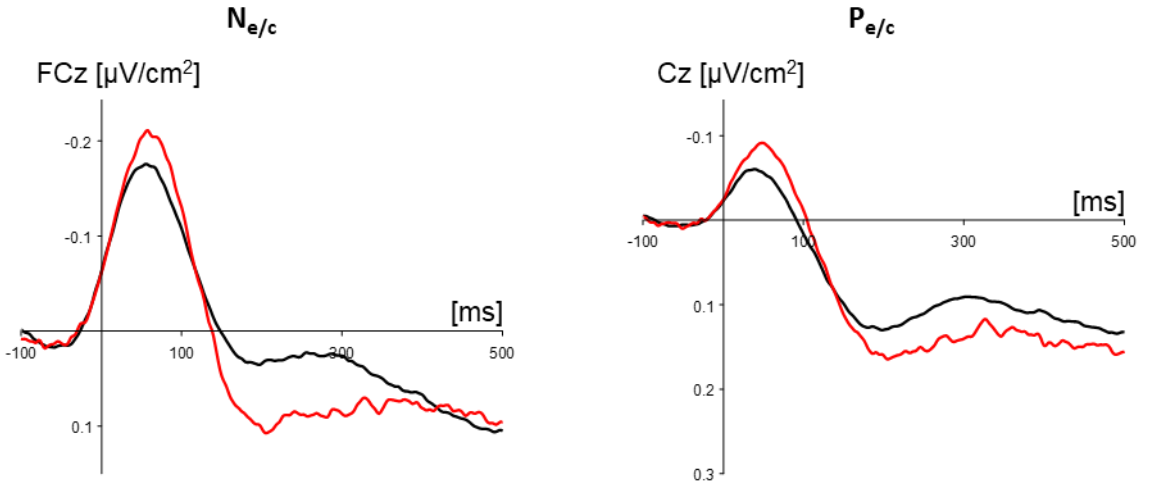
5. I would like to see the ERP analyses on untransformed raw data. CSD-transformations often reduce the Ne, but increase the Nc. If, for instance, the Nc was actually increased with older age, this effect could look different (or even disappear) with a CSD-transformation. The same holds for the Ne, in the opposite direction. The Pe analyses should additionally be presented for untransformed data as well, for the same reasons.

We thank the Reviewer for raising this important issue. Following the Reviewer's advice, we provide the ERP waveforms for errors and correct responses using untransformed raw data below. As can be seen in Figure R1, the CSD transformation indeed decreased the N_e and increased the N_c amplitude. Thus, the transformation makes our analysis even more conservative.

To additionally show this statistically, we fitted the same linear mixed effects models to the raw ERP data that are described in the manuscript for the CSD-transformed data. For the $N_{e/c}$, we found the same effects for the raw data, namely an effect of accuracy [$F(1,48.4) = 23.326$, $p < .001$] and an interaction between accuracy and age [$F(1,41.4) = 10.754$, $p = .002$]. The amplitude for errors was significantly larger than the amplitude for correct responses. For the $P_{e/c}$, accuracy significantly predicted the ERP amplitude [$F(1,60.6) = 5.338$, $p = .024$], as error amplitudes were larger compared to correct amplitudes. The CSD-transformed data also revealed an effect of accuracy and no effect of age. The only discrepancy between the approaches was that we found an interaction between age and accuracy on the $P_{e/c}$ amplitude using the transformed, but not the raw data. As this effect was not large in our reported analysis, the increased noise in the raw data might have concealed the effect in this analysis. In sum, the pattern of results for the raw ERP data is very similar to the CSD-transformed data.

In general, the CSD-transformation leads to a clearer, reference-free separation of ERP components, because it serves as a spatial high-pass filter by removing contributions of temporally overlapping components of different neural generators (Kayser and Tenke, 2015; Luck, 2014). The re-analysis of our EEG data at the single-trial level additionally encouraged us to retain our original approach. However, we now provide the full results of the analyses reported here in the supplementary material S5.

CSD



Raw ERPs

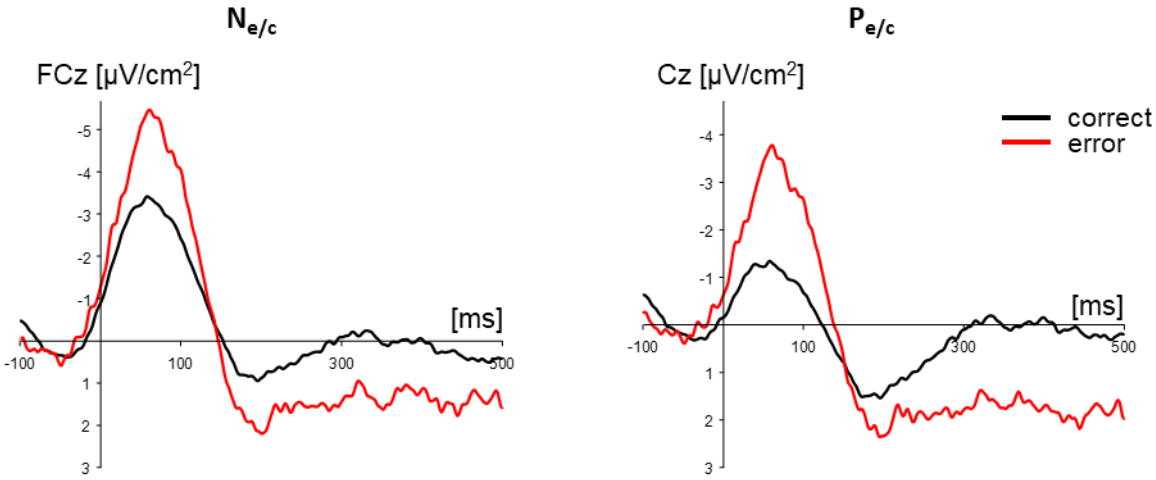


Figure R1. Response-locked event-related potentials for errors and correct responses after current source density transformation (top row) and using the raw ERP data (bottom row). $N_{e/c}$ (left column) is computed at electrode FCz and $P_{e/c}$ (right column) at electrode Cz. Errors are shown in red, correct trials in black.

Reviewer #2:

In this work, Overhoff and colleagues report the results of a large-scale EEG experiment with participants across a wide age range, looking into metacognition. The behavioral data show a decrease in metacognitive accuracy with age, despite adaptive adjustments of behavior uncorrelated with age. The neural data show that the relation between confidence and Ne/c changes with age, whereas this is not the case for the Pe/c. The manuscript is very well written and analyses of such a large EEG dataset is noteworthy and will definitely be of interest. Below I outline a couple of ways in which I think the manuscript can still be improved in important ways.

Major points

1. One drawback of the current analysis approach is that data from a lot of participants cannot be used due to a lack of trials. As a consequence, some of the most interesting analysis are done on only 44 participants, which gives very low power to detect correlations with age. Moreover, a cut off of 6 trials only (p. 13) seems very low to me to compute average ERPs. A far better approach would be to use linear mixed modeling which allows to analyze data at the trial level and retain all participants in all analyses (the imbalanced data are handled by the assumption that individual estimates come from a normal distribution around the group mean). Given these concerns, in this case this method of analyses seems far superior. I realize that this induces a lot of extra work, but in this case I am convinced it would make the argument much stronger.

We are grateful for this comment and agree that using linear mixed effects models has great advantages under these circumstances. Conducting these analyses allowed us to model our data in a better way, drawing a much clearer picture of subtle changes in the processing of confidence across the lifespan.

We applied the suggested linear mixed effects models to the single trial behavioural data and ERP amplitude data, which enabled us to include almost all participants in the analyses (note that one participant still had to be excluded due to too noisy EEG data). Moreover, we were now able to differentiate three levels of confidence in the separate analysis of errors and correct trials, thus avoiding that we have to combine confidence levels as in the previous version. Accordingly, major modifications have been made to the methods and results sections. As they are rather extensive, we decided to briefly summarise them here, and we refer to the main manuscript for all tracked changes.

For the analysis of EEG data, the linear mixed effects regression models replaced the ANCOVAs. We added the between-subject factor age and the within-subject factor accuracy, or, for the separate analysis of errors and correct responses, the within-subject factor confidence. Random slopes for the within-subject factor of interest were added, if possible. In order to run

analyses at the single trial level, we extracted adaptive mean amplitudes (± 50 ms around the peak) instead of peak ERP amplitudes from the EEG data, because they are more robust and less affected by noise (Clayson et al., 2013). For post-hoc analysis of significant effects of accuracy or confidence, we applied pairwise comparisons and for significant interactions, we applied linear mixed models for all levels of accuracy or confidence to test for associations with age.

The same analyses were applied to the behavioural data, except for the group-level analysis of Phi. Here, we used simple linear regressions and multiple linear regressions to test for interaction effects of ER (error rate) and d2-test score.

Importantly, this major methodological modification led to a highly similar set of results. However, there were some changes, which were most likely due to the enhanced sensitivity and reduction of noise when including all participants and trials:

For response caution, we now find a significant modulation by confidence, whereby participants responded faster and made more errors after high confidence ratings compared to medium or low confidence ratings. This effect did not reach the significance threshold before.

For the regression on the $N_{e/c}$ amplitude, the model using all trials with the factor accuracy did not reveal an effect of age anymore. For the analysis of errors, we now show an effect of confidence, and an interaction with age, as 'surely wrong' errors had larger N_e amplitudes than 'surely correct' errors, which decreased with age. For correct trials, on the other hand, we did not find any significant effects anymore.

For the $P_{e/c}$ component, we also did not find effects of age or confidence on the correct trials anymore. The pattern of results for the errors was unchanged.

In sum, the majority of our conclusions remain valid after adjusting all of our analyses. This robustness is reassuring, because we agree with the Reviewer that these new analyses are far superior, making the interpretation of our findings much stronger. We again thank the Reviewer for this suggestion.

Specific changes to the discussion can be found below. Please see pages 11-19 in the revised manuscript for all detailed changes in Methods and Results.

Changes in manuscript (page 23 f.):

"…but rather in a differential modulation of the N_e across the lifespan. The modelling results revealed that the N_e amplitude was also affected by the interaction between confidence and age. With older age, the N_e declined for all errors in which high conflict was perceived. In other words, only the N_e amplitude of errors that were rated as 'surely wrong' varied in amplitude across the lifespan. As the $N_{e/c}$ is sensitive to conflict between the given and the actual correct response, older adults seemed to having had difficulties internally representing the correct response in highly conflicting situation (Yeung et al., 2004). Notably, this effect was error-

specific, that is, we cannot draw conclusions about internal processes for correct responses, as the N_c amplitude did not show a relation to confidence that could have varied with age.

We suggest that the reduced N_e amplitude of low confidence errors with higher age could be related to the observed decrease in metacognitive accuracy in our flanker task. If older adults did not perceive high conflict due to difficulties in forming an accurate internal representation of the correct response, …"

Changes in manuscript (page 25):

"However, the application of linear mixed effects modelling provided us with a powerful tool that can account for varying trial numbers across participants and importantly, the multi-level structure of our data."

2. The measure of metacognitive accuracy, phi, is suboptimal and at times misrepresented:

a. Theoretically it can be shown that phi depends on the accuracy of the first-order response. Demonstrating that the relation between aging and phi holds after controlling for accuracy is relevant, but does not do away with this issue. I still think the findings are sufficiently noteworthy, but this limitation should be made clearer to the reader.

We agree that the task might pose higher demands on older participants compared to younger participants. Thus, it is inherent to the task that both accuracy and confidence might be different depending on the age of the participant, which in turn might have influenced the metacognitive abilities. As we did not opt for adaptive algorithms to adjust task difficulty for each individual participant (which would have added another level of unwanted complexity), we tried to approach this problem by conducting multiple linear regressions. Ideally, we would have chosen to calculate metacognitive efficiency (meta-d'/ d'; Maniscalco and Lau, 2012) instead of phi. While metacognitive efficiency constitutes a commonly used measure that accounts for primary task performance, it is only applicable for two-choice signal detection tasks, and we are not aware of a comparable measure that could be used in a task with four response options.

Within the manuscript, we clarified this potential bias.

Changes in manuscript (page 11):

"It describes the extent to which the distributions of confidence ratings for correct and incorrect trials differ, while still depending on primary task performance and individual biases in confidence judgements …"

b. On the bottom of p. 21 it is written about phi that "a smaller value could either indicate more 'misclassifications', or a general response tendency towards the middle". This seems wrong both conceptually (a measure of metacognitive accuracy should not depend on confidence criteria) and in practice (as I understand it phi does not depend on the precise level of confidence, only on the consistency of its relation to accuracy).

The Reviewer is correct that a high phi coefficient reflects high consistency. However, we aimed to outline cases which could explain a lack of consistency. According to the literature, phi does depend on the level of confidence as it is not accounting for individual biases (Fleming & Frith, 2014). For each participant, phi is calculated as the correlation between a vector containing the accuracy of each trial (i.e. 0, 1) and a vector containing the respective confidence ratings (i.e. 1, 2, 3). Accordingly, the highest correlation would be found if all errors were rated with low confidence and all correct responses with a consistently high confidence. The more variability in this relationship, the lower phi becomes. We assume each participant to have some internal confidence criterion defining the degree of evidence needed to rate a decision as, for instance, 'surely correct'. When this criterion is not reached, they might rate the decision as 'maybe correct', resulting in a lower phi. Importantly, this does not tell us the degree of evidence the person accumulated, but could also be due to a comparably high internal criterion to indicate high certainty.

Both metacognitive accuracy as well as metacognitive bias provide interesting information about a sample. To get an impression of metacognitive bias, which is not distinguishable from metacognitive accuracy using phi, we computed the proportion of ratings across confidence levels for errors and correct responses for each participant (Figure 2). It showed that older adults tended to apply a more conservative rating strategy, which negatively contributed to their phi values.

We recognise that the explanation of phi may have been misleading and we have adjusted it in the manuscript based on the Reviewer's comment.

Changes in manuscript (page 21):

"…Given the nature of Phi, a smaller value could either indicate more undetected errors or correct responses rated as being incorrect, or a generally higher uncertainty (i.e. rating all correct responses as 'maybe correct' will result in a lower Phi value than rating the same number of correct responses as 'surely correct'). Indeed, we observed that older adults used the extreme ends of the confidence scale considerably less often than younger adults."

3. Some of the findings that make this paper interesting are based on a null effect (i.e. adaptive adjustments, the Pe/c). Although the sample size is rather large for an EEG experiment, for correlations (i.e., with age) it is certainly not. As the authors probably know, absence of evidence is not evidence for the absence of an effect. It would be good if

the authors could provide some indication concerning the extent to which their data support these null effects (e.g., using Bayesian statistics).

We agree with the Reviewer that the null effects of age on, for instance, $P_{e/c}$ amplitudes, are an interesting finding in our paper, and we are aware of the difficulty in interpreting null results in a frequentist null-hypothesis testing framework.

Based on the Reviewer's first comment, we are now analysing our data using (generalised) linear mixed effects models. Though model estimates are still assessed in a frequentist approach, the models can far better account for variability in our data (e.g., different numbers of trials per participant, which allows the inclusion of all participants) and the inherent multi-level structure due to individual differences (Lo and Andrews, 2015).

Following the Reviewer's suggestion, we computed additional Bayesian statistical analyses using the package BayesFactor in R (version 0.9.12-4.2; Morey and Rouder, 2018) to assess the extent to which our data support the null effects. The results were added to the supplementary material S6. In order to examine the null effects of age for response caution and $P_{e/c}$ amplitude, we compared the full models including the within-subject factor of interest (accuracy or confidence) and the between-subject factor age to a null model including only the within-subject factor.

For response caution, we tested the hypothesis that response caution is modulated by accuracy, age, and their interaction against the null hypothesis that it was only modulated by accuracy. We found anecdotal evidence in favour of the null hypothesis ($BF_{01} = 2.055$). Comparing the model predicting response caution by the factors confidence and age to the model including only age, we found strong evidence supporting the null hypothesis ($BF_{01} = 238.612$). This suggests that the modulation of response caution was indeed similar across the lifespan.

For the $P_{e/c}$, we assessed evidence for a modulation by age of all trials combined. Here, the model including the interaction term of accuracy and age was around five times more likely given the data than the null model ($BF_{10} = 5.264$). This is mirroring the significant interaction we found in the analysis reported in the manuscript. For the modulation of errors by confidence and age, we found strong evidence against an effect of age on the P_e amplitude ($BF_{01} = 614.830$). The same was true for the age effect on the P_c amplitude of correct responses ($BF_{01} = 1294.515$).

Taken together, these results suggest that our data robustly support the null effects of age on the confidence modulation of response caution and the $P_{e/c}$ amplitude.

4. Whereas confidence is usually studied in situations of perceptual uncertainty (e.g. random dot motion discrimination) error awareness is usually studied in tasks where errors occur not because of uncertainty but because of response conflict, i.e., "fast errors". It would be useful if the authors could provide some indication to what extent errors in the current task are mere conflict or also reflect uncertainty. E.g., based on Figure 4A it

appears as if "surely wrong" trials capture fast trials ("impulsive errors") and "unsure" trials reflect slow trials ("quality errors"). This is now a bit hidden because on p. 16 it only reads that errors are slower than corrects. These findings also seem opposite to the ones by Stahl et al. described on p. 22. Given that the current study makes the explicit link with the confidence literature, such a discussion would be informative for the broader readership.

As correctly recognised by the Reviewer, we explicitly aimed to bridge the gap between confidence and error detection studies. To achieve this, a classical conflict paradigm was used to elicit response conflict while it was simultaneously possible to assess metacognition on a confidence scale. By testing the ability to differentiate the colours without time pressure in a designated colour discrimination task, we intended to exclude errors due to pure perceptual uncertainty. Thus, we believe that the main source of errors was indeed conflict between representations, not perceptual uncertainty. However, of course we cannot rule out that there are other sources of errors in some trials, such as response conflict, memory failures (e.g., forgetting the correct stimulus-response mapping), attention failure, and many more. These, however, should be unsystematic and not the main (i.e. experimentally induced) source of errors.

It is further correct that 'unsure' trials showed the slowest RTs in our study, as can be seen in Figure 4a, and this is mentioned on p. 15 f.:

> *"…with trials associated with the 'unsure' confidence level (815.6 ± 6.7 ms) being considerably slower than trials rated as 'surely correct' (702.1 ± 1.3 ms) or 'surely wrong' (736.6 ± 6.7 ms)."*

In our opinion, this is in line with the findings by Stahl et al. (2020). The authors define slow, signalled errors as 'memory errors', and report lower confidence compared to other response types. Low confidence in their study was defined as low confidence in the detection of the error, or in other words, high uncertainty in the confidence judgement, while we define these trials as 'unsure' and find the same increased RTs. This replicates the findings by Stahl et al. (2020) and extends them to uncertainty in a decision, independent of accuracy. The discrepancies between our results and their results therefore appear to be due to differences in labelling.

We thank the Reviewer for the suggestion to highlight the important point regarding the link between confidence and error detection studies and made the following adjustments in the manuscript.

Changes in manuscript (page 21 f.):

"Interestingly, participants in our study responded slowest in case of uncertainty, i.e. 'unsure' ratings. In contrast, studies on decision confidence typically report increasing RT with decreasing confidence (Kiani et al., 2014; Rahnev et al., 2020; Weidemann and Kahana, 2016). Most of these studies specifically measured confidence in having made a correct decision (i.e. the lowest

confidence indicates guessing, while in our study it indicates high certainty in being incorrect), and typical paradigms in these studies are two-choice signal detection tasks in which the degree of sensory evidence, for instance, perceptual discriminability, is manipulated (Kiani et al., 2014; Moran et al., 2015; Rollwage et al., 2020). In our task, we ensured (using a designated colour discrimination test) that all stimuli were perceptually discriminable without time pressure, and our data showed no signs of age-related differences in stimulus processing (even though it remains possible that some decrease in colour perception, or other untested factors such as attention, working memory, etc., might have contributed to the age-related slowing we observed; see supplementary material S4). Instead, potential sources for errors could be, for instance, stimulus conflict caused by the flankers and the similarity of the stimulus colours, or difficulties in remembering the stimulus response mapping. Using a comparable paradigm, Stahl et al. (2020) found slow errors to be associated with lower confidence than fast, impulsive errors and inferred that those error types should predominantly be caused by weak stimulus-response representations (i.e. due to weak memory traces).

As such conclusions could not be drawn from classical error processing studies requiring only a binary error detection rating, our findings provide an important link between those and decision confidence studies. In a typical error processing paradigm that posed higher demands on the older adults (as indicated, for instance, by higher error rates), our results could be interpreted as their impaired metacognitive evaluation (assessed via confidence ratings) being partly related to more frequent memory-related errors, which appear to be more challenging to assess consciously (Maier and Steinhauser, 2017; Stahl et al., 2020)."

Minor points

1. I had a hard time relating the description of p. 20 to what is shown in Figure 6D and 6E. Visually these figures seem to convey the opposite message of the text, which is probably because error trials are noisier then correct trials? Perhaps adding error bars to the ERPs could resolve this confusion.

We adjusted Figure 6 and added error bars to the regression lines to allow an easy inspection of the different noise levels for the reader.

2. Can you also report the partial correlation between age and phi when controlling for overall RT (cf. section 3.1.6)?

Due to the changes in our analysis, we are now computing multiple linear regressions instead of partial correlations. The regression analysis including the factor of RT yielded no significant interaction between age and RT (p = .455), suggesting that the age-related decline in metacognitive accuracy was not linked to the general slowing of responses in older age.

3. Figure 2 could be even more informative if the y-axis labels were not computed separately for error and correct, but based on all trials.

We gladly provide the suggested figure here, but would like to point out our reasons for choosing to keep the separate plots for errors and correct trials in the manuscript.

Figure R2 below shows the proportions for all combinations of rating levels and errors and correct responses on a joint y-axis. This means that the values for the leftmost column depict the proportions of errors rated as ‘surely wrong’ of all responses. Thus, it depends both on the ratio of ‘surely wrong’ ratings as well as on the ratio of errors. Therefore, proportions for certain confidence levels for errors become very small if the participant did not make many errors overall. This complexity will get lost in a figure like the one requested, and we are worried that it might hide rather than illustrate important aspects of the distribution. For instance, in the revised Figure 2 plotting only errors, it is clearly visible that younger adults rate a higher proportion of their errors as ‘surely wrong’ compared to older adults, and this is not possible to visualise in the figure below.

Moreover, we think that it is reasonable to separate errors and correct trials from a theoretical perspective: errors have often been shown to be processed differently than correct responses, and the same is true for our electrophysiological findings. Therefore, they could also be differently related to confidence. In sum, we hope the Reviewer agrees that our figures are indeed informative.

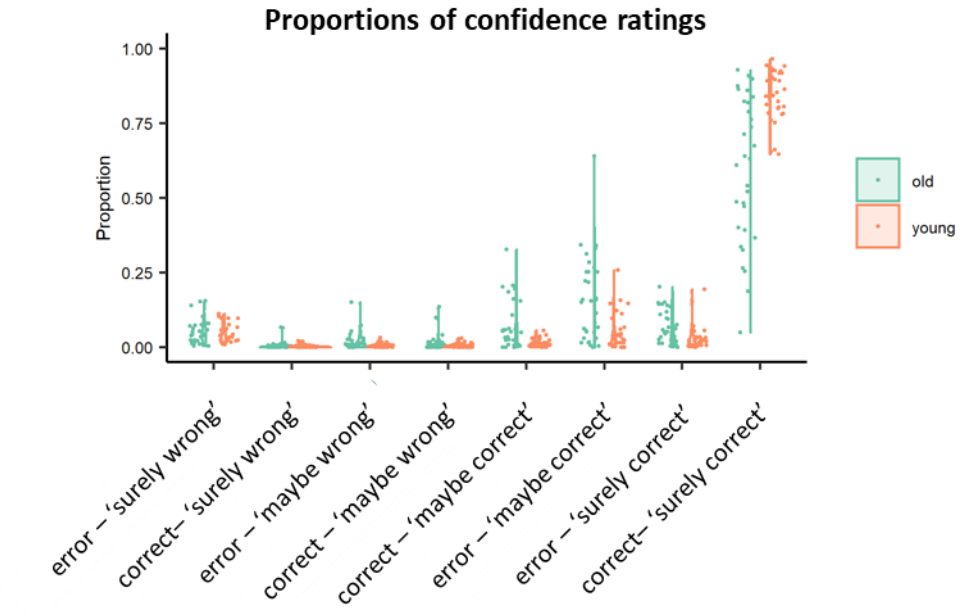


Figure R2. Proportions of all combinations of rating levels and errors and correct responses. Dots represent the individual proportion of the particular accuracy-confidence combination. A median split by age (*Mdn* = 46) was conducted for illustration purposes. Older adults are shown in green, younger adults in orange.

4. On the bottom of p. 12 it reads that error trials "surely wrong" will be labelled as "low confidence", yet in the figures it still read "surely wrong" for error trials. Please clarify.

We apologise for the confusion, but we also note that this re-categorisation of trials is now obsolete. Due to the changes in our analysis, we are now modelling our data at the single trial level, accounting for imbalances in trial numbers across participants. This means, we do not collapse across confidence levels anymore, because we are now able to use three confidence levels for all analyses. The revised manuscript makes no more reference to the labels mentioned by the Reviewer.

5. I am wondering whether it could have been confusing for participants that the same buttons were used for choices and confidence. For example, are there strong 'priming' effects from the choice buttons to the confidence buttons?

We thank the Reviewer for posing this relevant question. When designing the experiment, we aimed to avoid strong overlap between processes of decision making and confidence judgement by adding an inter-judgement interval of 800 ms between the initial response and the display of the confidence scale. Using the same keys for the primary and secondary response is common practice in the field, and in comparison to similar studies, our interval is rather long (e.g., Boldt

and Yeung, 2015; Palmer et al., 2014; Wessel et al., 2018). We assume that it would be even more confusing if the participants had to switch the keys within a trial.

Based on the Reviewer's comment, we assessed whether we could observe a 'priming' effect of the first response on the confidence judgement in our data. Generally, the flanker task was designed in a way that the target colours, and thus the correct response keys, were counterbalanced across trials. In reality, of course, the actual response fingers were not perfectly counterbalanced (due to errors). Considering the confidence judgments, the majority of trials in our study were rated as 'surely correct', corresponding to pressing the fourth response key. To visualise the relationship between used finger in the initial response and the second rating, we below plotted the individual proportion of the four confidence levels associated with the initial response for each of the four response keys (Figure R3). Note that confidence level 1 was always mapped to the left middle finger, level 2 to the left index finger, level 3 to the right index finger and level 4 to the right middle finger. As can be seen from the figure below, no 'priming' effect is visible in the data. For instance, when the primary decision was made with the left middle finger, the decision was not rated as 'surely wrong' more often (Confidence 1).

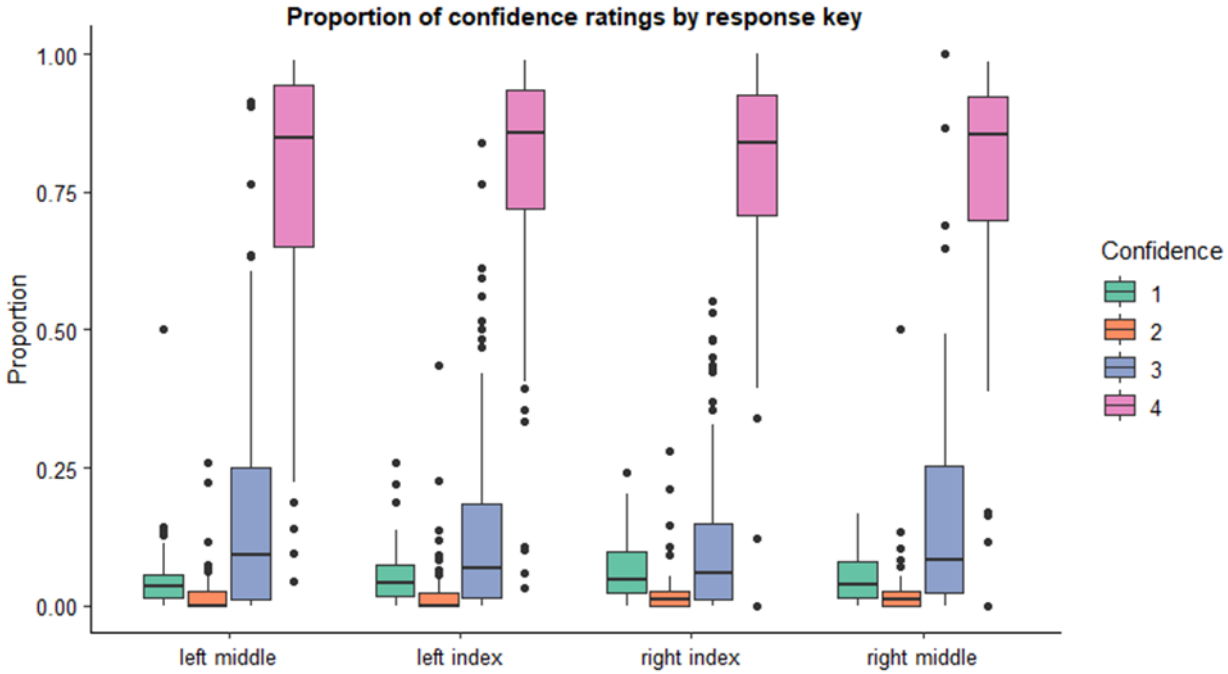


Figure R3. Proportion of each confidence rating (1-4, see legend) for each response finger of the initial decision (x-axis). Vertical lines depict the median. The lower and upper hinges reflect the first and third quartiles and the whiskers extend to the lowest/largest value or max. 1.5 times the inter-quartile range from the hinge. Dots represent individual data points lying outside the range of the whiskers.

6. Please report in the participant section that participants were not color-blind.

We added the suggested changes in the manuscript.

Changes in manuscript (page 6):

"Inclusion criteria were right-handedness according to the Edinburgh Handedness Inventory (EDI; Oldfield, 1971), fluency in German, (corrected-to-) normal visual acuity, no colour-blindness and no history of neurological or psychiatric diseases."

7. Can you please insert an additional approach to dissociate corrects from errors in Figure 4 (e.g., different symbols)? As a colorblind person I was unable to perceive the difference between these colors.

We regret not considering this potential barrier before. We have now updated all figures and ensured to use different line types, which are discriminable independent of the colours.

8. Top p. 20, please don't write "marginally missed the significance threshold". Non-significant effects are uniformly distributed so p=.06 is as informative as p=.99.

We acknowledge this valid point. Due to the changes in our analysis, we do not make this statement anymore when reporting the revised results.

9. Top p. 25, "it is not metacognitive accuracy per se, but rather the perceived confidence that shapes future behavior" feels confusing to me. Metacognitive accuracy cannot drive behavior because it is just a conceptual term to describe the relation between confidence and accuracy. Perhaps it would be more correct to say that perceived confidence shapes future behavior irrespective of metacognitive accuracy.

We agree with the Reviewer and are thankful for this suggestion. Indeed, this is the message we tried to convey, so we adjusted this part accordingly.

Changes in manuscript (page 25):

"This suggests that it is the perceived confidence that shapes future behaviour, irrespective of metacognitive accuracy:…"

Reviewer #3:

This study by Overhoff and colleagues investigates neural correlates of metacognition across the adult lifespan. The authors recorded scalp EEG as a sample of participants with a broad age range performed a challenging flanker task, and assessed effects of age on behaviour and established measures of performance monitoring/metacognition (the error negativity and error positivity event-related potentials). The findings are interesting and nuanced, suggesting that while aging produces a clear decrease in metacognitive accuracy on this task, this is accompanied by only subtle effects on the associated neural correlates (mainly the error negativity).

This is a well-conducted, thorough study that fills an important gap in the literature on cognitive aging. Specifically, while a number of studies have looked into effects of aging on neural correlates of categorical error detection, the present study is the first to assess these neural signatures using graded confidence judgments. This is an important next step, since graded confidence judgments are potentially more informative for identifying the nature of possible age-related changes in metacognitive ability. I also found the paper to be well written, with a good introduction to the relevant literature and a considered discussion of a nuanced set of results.

I have only one major comment that I feel needs to be addressed, in addition to a small number of minor points.

My major comment pertains to the way in which the error positivity (Pe) component was defined. The authors state early on (P.3) that the Pe typically peaks 250 ms after the response; and accordingly, they measure the Pe as the peak amplitude from 150-350 ms and - importantly - over electrode Cz. While this definition may be motivated by some older studies that the authors cite, it is not consistent with what I believe to be the majority (or even all) of the studies that have motivated the emerging view, which the authors appear to subscribe to, that the Pe reflects the accumulation of error evidence. Specifically, each of the following papers have measured the Pe from more posterior cites (typically centred on electrode Pz): Steinhauser & Yeung (2010, Journal of Neuroscience), Boldt & Yeung (2015, Journal of Neuroscience), Murphy et al. (2015, eLife), Desender et al. (2019, eLife). This applies even when the CSD transform was used (Murphy et al., 2015), as it was presently.

The Pe also tends to peak - and importantly, exhibit sensitivity to confidence (Boldt & Yeung, 2015) - toward the end and after the window analyzed here. And please note that these two concerns - about latency and topographic localization - may be especially acute for older adult cohorts: If the generally slower decision-making with increased age generalizes to implicit confidence judgments then one would expect this to translate into an even later Pe latency for older adults; and there is some evidence that the Pe may be even more posteriorly distributed in older adults compared to young adults (Harty et al., 2017, Neuroimage).

In sum, I am concerned that the current analysis of the putative Pe is in fact missing a sizeable part of this signal. At the extreme, it's even possible that what the authors are presently analysing may be the same neural signal that generates the error negativity: a fronto-central low-frequency oscillation phase-locked to the response (e.g. Yeung et al., 2007, Psychophysiology). I would very much welcome further analysis of the signal approximately over electrode Pz, and shifted later in time (from the present draft it is not possible to assess what might be going on there, since the plotted topographies in Figure 5 are restricted to earlier time-points, and plotted ERPs are restricted to fronto-central electrodes).

but this is far from universally true) and anterior (latter especially for CSD-transformed data). Consistent with some studies but not the majority (or even all) of those that have made a convincing case for the Pe as an evidence accumulation signal. Indeed, not totally clear whether Pe as defined here is really a distinct signal from Ne (in the sense that both may reflect a fronto-central theta). Would like to see an analysis of later, more parietal signal.

We thank the Reviewer for raising this important issue. The definition of the $P_{e/c}$ varies largely between studies, and to date, there is still no agreement on how to measure the component (Boldt and Yeung, 2015). This is probably related to the fact that it is a late, slow potential that often does not show a clear peak at all (Clayson et al., 2013).

We acknowledge that many studies are using more posterior scalp locations to extract the P_e amplitude, though, as the Reviewer mentions, this vastly varies. Likewise, the Cz and CPz electrodes are also commonly used to assess the $P_{e/c}$ (e.g., Endrass et al., 2012; Nieuwenhuis et al., 2001; Siswandari et al., 2019; Steinhauser and Yeung, 2010). Our decision to extract the P_e amplitude from electrode Cz was primarily based on the visual inspection of the topographical maps and grand average waveforms at the candidate electrodes. As Figure R4 shows, the only distinct P_e-like deflection can be seen at electrode Cz.

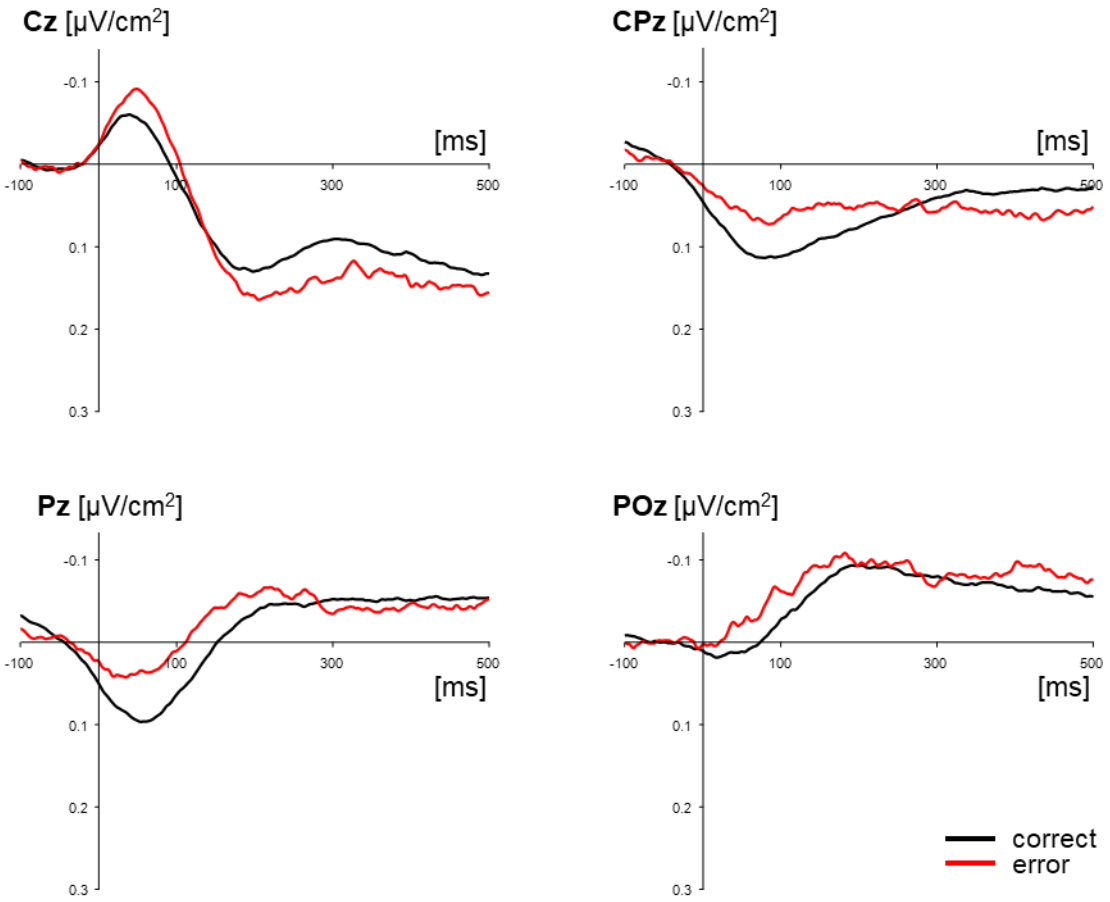


Figure R4. $P_{e/c}$ for errors and correct responses after current source density transformation at four different electrode sides (Cz, CPz, Pz, POz).

Concerning the timing of the $P_{e/c}$, two points should be noted with regard to the present study: First, many of the studies mentioned by the Reviewer are using a difference measure between errors and correct responses for the $P_{e/c}$ (Boldt and Yeung, 2015; Desender et al., 2019; Steinhauser and Yeung, 2010). Results from Steinhauser and Yeung (2010) suggest that this could result in slightly later peak amplitudes, as can also be assumed from our data when using the difference wave (see Figure R5 below). Second, compared to mean amplitudes, the measure of adaptive means ($\pm$ 50 ms around the peak) we are now using due to major changes in our analysis (i.e. all behavioural and electrophysiological analyses are now conducted at the single-trial level using linear mixed effects models; please see comment 1 by Reviewer #2) is less dependent on the particular time window (Clayson et al., 2013; Luck and Gaspelin, 2017). Nevertheless, the time window we are using to identify the peak ranges from 150 to 350 ms, which includes the time windows of most of the studies mentioned before.

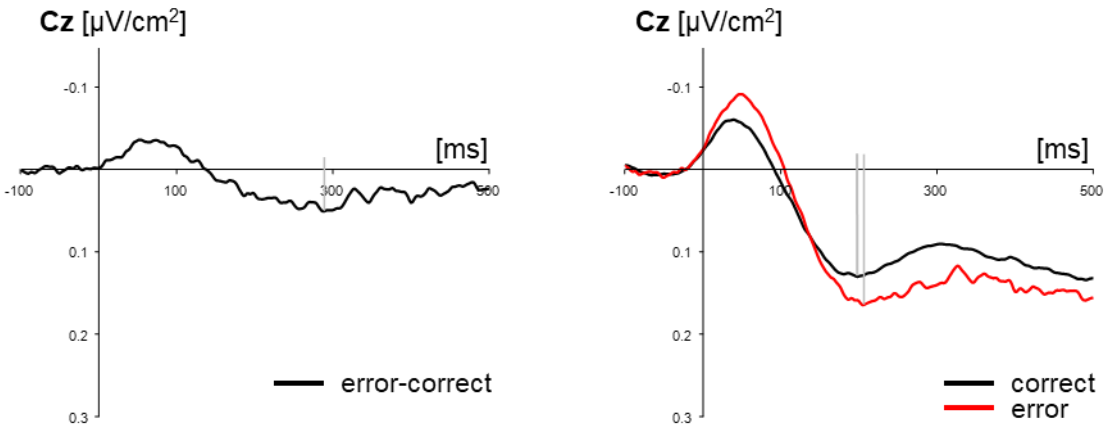


Figure R5. $P_{e/c}$ for the difference (error-correct; left) and errors and correct responses separately (right) after current source density transformation. Grey, vertical lines point to the peak latencies.

Lastly, we agree with the Reviewer that age potentially affects the topography and the latency of the $P_{e/c}$. This could have gone unnoticed in our study, as we assessed effects of age as gradual changes across the lifespan. To test for this possible confound, we split the group by age (median split) and show the topographical maps for younger and older adults in three time windows from 150 to 450 ms below. Figure R6 illustrates that the positive deflection is slightly more spatially expanded in the older age group, but no clear differences in scalp topography can be seen.

To statically test for differences in latency between the two age groups, we correlated the peak latency of the P_e and P_c with age. We computed the individual peak latencies at electrode Cz in three time windows (150 - 350 ms, 150 - 550 ms, 200 - 400 ms) separately for errors and correct responses and correlated them with age, but we did not find any significant association between the latency and age (150 - 350 ms, error: $r(62) = -.096$, $p = .450$, correct: $r(62) = .037$, $p = .774$; 150 - 550 ms, error: $r(62) = -.138$, $p = .279$, correct: $r(62) = .196$, $p = .121$; 200 - 400 ms, error: $r(62) = -.138$, $p = .279$, correct: $r(62) = .047$, $p = .711$).

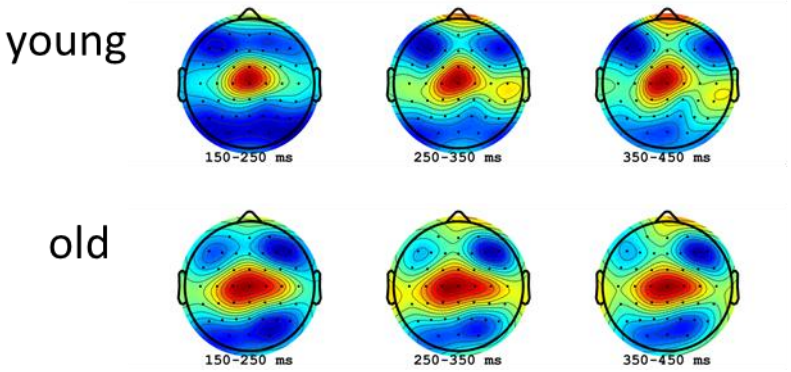


Figure R6. Topographical maps of errors after current source density transformation for three different time windows. The sample was split into younger and older adults via median split *(Mdn* = 46).

In sum, we are aware of the different approaches to measure the $P_{e/c}$ and that many studies use different electrode sites and time windows. However, the visual inspection of our data revealed a clear $P_{e/c}$ component at electrode Cz around 200 - 250 ms that can similarly be found in the literature. Our results are further robust to variations in the exact time window to calculate the mean. Moreover, our task differs from other studies with regards to its complexity. Stahl et al. (2020) used a similar design and found spatially and temporally highly comparable $P_{e/c}$ components to the ones in our study.

Minor:

*** I found uses of the terms "higher confidence" and "lower confidence" to be confusing at times. If I understand correctly, the authors use "higher confidence" synonymously with ratings of "surely correct"; and "lower confidence" synonymously with ratings of "surely error". The confusion comes from the fact that both "surely correct" and "surely error" are in fact instances of high confidence - but in different things (the former that the preceding response was correct; the latter that it was incorrect). Low confidence, by contrast, is captured by the two categories of "unsure" ratings. I recommend the authors take a pass through the manuscript and clarify uses of the terms "higher confidence" and "lower confidence" accordingly.**

Due to the changes in our analysis we are now able to use three levels of confidence and thus, we do not need to collapse confidence levels anymore. The respective old labels have been removed from the manuscript. Moreover, we also tried to clarify the use of 'high confidence' and 'low confidence' throughout the manuscript by emphasising the discrimination between confidence and certainty in the relevant sections. We hope that this solves the issue and avoids any confusion.

Changes in manuscript (page 18):

"…the N_e amplitudes of low confidence errors (i.e. rated as 'surely wrong') was decreased with older age…"

Changes in manuscript (page 20):

"…$N_{e/c}$ and $P_{e/c}$ amplitudes declined with higher confidence in having made a correct response, which was specifically observed for trials with response errors."

Changes in manuscript (page 24):

"…low confidence (reflecting a belief that an error had been committed) or uncertainty about a decision were associated with higher response caution…"

* **P.2: the citation of Desender et al., 2018 seems to be missing from the reference list.**

We thank the Reviewer for this observation. It should have been 'Desender et al., 2019b', which is included in the reference list. The in-text citation was updated accordingly.

Changes in manuscript (page 2):

"…because it guides our present and future behaviour (Desender et al., 2019b; Rabbitt, 1966)."

* **P.11, typo: "…For this analysis, only to pairs of consecutive trials…"**

Changes in manuscript (page 11):

"For this analysis, only pairs of two consecutive valid trials were included."

* **P.14, typo: "…processing speed as assesses by…"**

Changes in manuscript (page 13):

"The average score for sustained attention and processing speed as assessed by the d2-test…"

* **P.19, last line, suspected typo: "…a significant interaction between confidence and age…". Should "confidence" here in fact read "accuracy"?**

This is correct, it was meant to be 'accuracy'.

Changes in manuscript (page 19):

"There was no effect of age, but a significant interaction between accuracy and age…"

References

Boldt, A., Yeung, N., 2015. Shared neural markers of decision confidence and error detection. J. Neurosci. 35, 3478–3484. https://doi.org/10.1523/JNEUROSCI.0797-14.2015

Clayson, P.E., Baldwin, S.A., Larson, M.J., 2013. How does noise affect amplitude and latency measurement of event-related potentials (ERPs)? A methodological critique and simulation study. Psychophysiology 50, 174–186. https://doi.org/10.1111/psyp.12001

Desender, K., Murphy, P.R., Boldt, A., Verguts, T., Yeung, N., 2019. A Postdecisional Neural Marker of Confidence Predicts Information-Seeking in Decision-Making. J. Neurosci. 39, 3309–3319. https://doi.org/10.1101/433276

Endrass, T., Schreiber, M., Kathmann, N., 2012. Speeding up older adults: Age-effects on error processing in speed and accuracy conditions. Biol. Psychol. 89, 426–432. https://doi.org/10.1016/j.biopsycho.2011.12.005

Fleming, S.M., Frith, C.D., 2014. The Cognitive Neuroscience of Metacognition, The Cognitive Neuroscience of Metacognition. Springer. https://doi.org/10.1007/978-3-642-45190-4

Kayser, J., Tenke, C.E., 2015. On the benefits of using surface Laplacian (Current Source Density) methodology in electrophysiology. J. Psychophysiol. 97, 171–173. https://doi.org/10.1016/j.ijpsycho.2015.06.001

Kiani, R., Corthell, L., Shadlen, M.N., 2014. Choice certainty is informed by both evidence and decision time. Neuron 84, 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

Lo, S., Andrews, S., 2015. To transform or not to transform: using generalized linear mixed models to analyse reaction time data. Front. Psychol. 6, 1–16. https://doi.org/10.3389/fpsyg.2015.01171

Luck, S.J., 2014. An introduction to the event-related potential technique. MIT press.

Luck, S.J., Gaspelin, N., 2017. How to get statistically significant effects in any ERP experiment (and why you shouldn't). Psychophysiology 54, 146–157. https://doi.org/10.1111/psyp.12639

Maier, M.E., Steinhauser, M., 2017. Working memory load impairs the evaluation of behavioral errors in the medial frontal cortex. Psychophysiology 54, 1472–1482. https://doi.org/10.1111/psyp.12899

Maniscalco, B., Lau, H., 2012. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious. Cogn. 21, 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Moran, R., Teodorescu, A.R., Usher, M., 2015. Post choice information integration as a causal determinant of confidence: Novel data and a computational account. Cogn. Psychol. 78, 99–147. https://doi.org/10.1016/j.cogpsych.2015.01.002

Morey, R.D., Rouder, J.N., 2018. BayesFactor: Computation of Bayes Factors for Common Designs.

Murphy, P.R., Robertson, I.H., Harty, S.S., O'Connell, R.G., O'Connell, R.G., 2015. Neural evidence accumulation persists after choice to inform metacognitive judgments. Elife 4, 1–23. https://doi.org/10.7554/eLife.11946

Nieuwenhuis, S., Richard Ridderinkhof, K., Blom, J., Band, G.P.H., Kok, A., 2001. Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. Psychophysiology 38, 752–760. https://doi.org/10.1017/S0048577201001111

Oldfield, R.C., 1971. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia 9, 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Palmer, E.C., David, A.S., Fleming, S.M., 2014. Effects of age on metacognitive efficiency. Conscious. Cogn. 28, 151–160. https://doi.org/10.1016/j.concog.2014.06.007

Rabbitt, P.M.A., 1966. Errors and error correction in choice-response tasks. Exp. Psychol. 71, 264–272. https://doi.org/10.1037/h0022853

Rahnev, D., Desender, K., Lee, A.L.F., Adler, W.T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L.Y., Balcı, F., Bang, J.W., Bègue, I., Birney, D.P., Brady, T.F., Calder-Travis, J., Chetverikov, A., Clark, T.K., Davranche, K., Denison, R.N., Dildine, T.C., Double, K.S., Duyan, Y.A., Faivre, N., Fallow, K., Filevich, E., Gajdos, T., Gallagher, R.M., de Gardelle, V., Gherman, S., Haddara, N., Hainguerlot, M., Hsu, T.Y., Hu, X., Iturrate, I., Jaquiery, M., Kantner, J., Koculak, M., Konishi, M., Koß, C., Kvam, P.D., Kwok, S.C., Lebreton, M., Lempert, K.M., Ming Lo, C., Luo, L., Maniscalco, B., Martin, A., Massoni, S., Matthews, J., Mazancieux, A., Merfeld, D.M., O'Hora, D., Palser, E.R., Paulewicz, B., Pereira, M., Peters, C., Philiastides, M.G., Pfuhl, G., Prieto, F., Rausch, M., Recht, S., Reyes, G., Rouault, M., Sackur, J., Sadeghi, S., Samaha, J., Seow, T.X.F., Shekhar, M., Sherman, M.T., Siedlecka, M., Skóra, Z., Song, C., Soto, D., Sun, S., van Boxtel, J.J.A., Wang, S., Weidemann, C.T., Weindel, G., Wierzchoń, M., Xu, X., Ye, Q., Yeon, J., Zou, F., Zylberberg, A., 2020. The Confidence Database. Nat. Hum. Behav. 4, 317–325. https://doi.org/10.1038/s41562-019-0813-1

Rollwage, M., Loosen, A., Hauser, T.U., Moran, R., Dolan, R.J., Fleming, S.M., 2020. Confidence drives a neural confirmation bias. Nat. Commun. 11. https://doi.org/10.1038/s41467-020-16278-6

Siswandari, Y., Bode, S., Stahl, J., 2019. Performance monitoring beyond choice tasks: The time course of force execution monitoring investigated by event-related potentials and multivariate pattern analysis. Neuroimage 197, 544–556. https://doi.org/10.1016/j.neuroimage.2019.05.006

Stahl, J., Mattes, A., Hundrieser, M., Kummer, K., Mück, M., Niessen, E., Porth, E., Siswandari, Y., Wolters, P., Dummel, S., 2020. Neural correlates of error detection during complex response selection: Introduction of a novel eight-alternative response task. Biol. Psychol. 156. https://doi.org/10.1016/j.biopsycho.2020.107969

Steinhauser, M., Yeung, N., 2010. Decision processes in human performance monitoring. J. Neurosci. 30, 15643–15653. https://doi.org/10.1523/JNEUROSCI.1899-10.2010

Weidemann, C.T., Kahana, M.J., 2016. Assessing recognition memory using confidence ratings and response times. R. Soc. Open Sci. 3, 150670. https://doi.org/10.1098/rsos.150670

Wessel, J.R., Dolan, K.A., Hollingworth, A., 2018. A blunted phasic autonomic response to errors indexes age-related deficits in error awareness. Neurobiol. Aging 71, 13–20. https://doi.org/10.1016/j.neurobiolaging.2018.06.019

Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., Price, A.L., 2014. Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. https://doi.org/10.1038/ng.2876

Yeung, N., Botvinick, M.M., Cohen, J.D., 2004. The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol. Rev. 111, 931–959. https://doi.org/10.1037/0033-295X.111.4.931

AGEING AND METACOGNITION

Verification for the manuscript

"Neural correlates of metacognition across the adult lifespan"

1. All authors declare that they have no potential conflicts of interest to disclose.
2. This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation – Project-ID 431549029 – SFB 1451; GRF and PHW), and the Australian Research Council (Discovery Project Grant; DP160103353; SB).
3. The data and analyses contained in this manuscript have not been submitted elsewhere and will not be submitted elsewhere while under consideration at *Neurobiology of Aging*.
4. The study was approved by the ethics committee of the German Psychological Society (DGPs) and conformed to the Declaration of Helsinki.
5. All authors have reviewed the contents of the manuscript being submitted, approved its contents and validated the accuracy of the data.

AGEING AND METACOGNITION

CRediT Authorship Contribution Statement

Helen Overhoff: Conceptualization, Methodology, Software, Investigation, Data Curation, Formal analysis, Writing - Original Draft, Visualization, Writing - Review & Editing. **Yiu Hong Ko**: Conceptualization, Writing - Review & Editing. **Daniel Feuerriegel**: Methodology, Writing – review & editing. **Gereon R. Fink**: Resources, Writing - Review & Editing. **Jutta Stahl**: Conceptualization, Methodology, Software, Writing - Review & Editing, Supervision. **Peter H. Weiss**: Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision. **Stefan Bode**: Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision. **Eva Niessen**: Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Supervision.

Figure 1

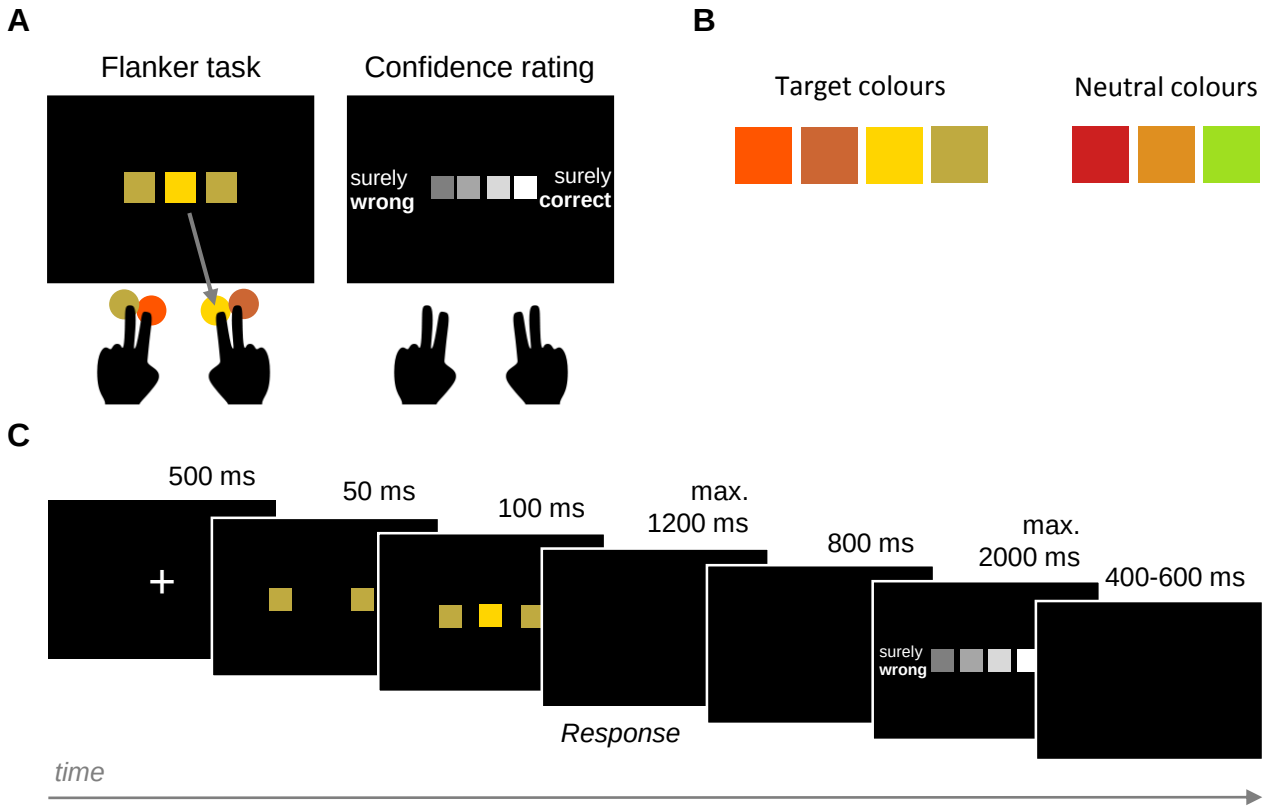


Figure 2

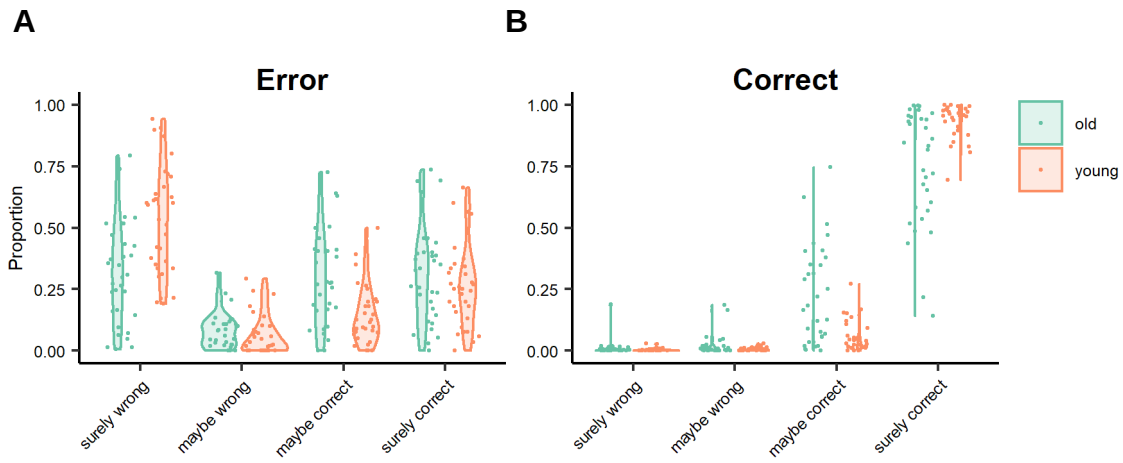


Figure 3

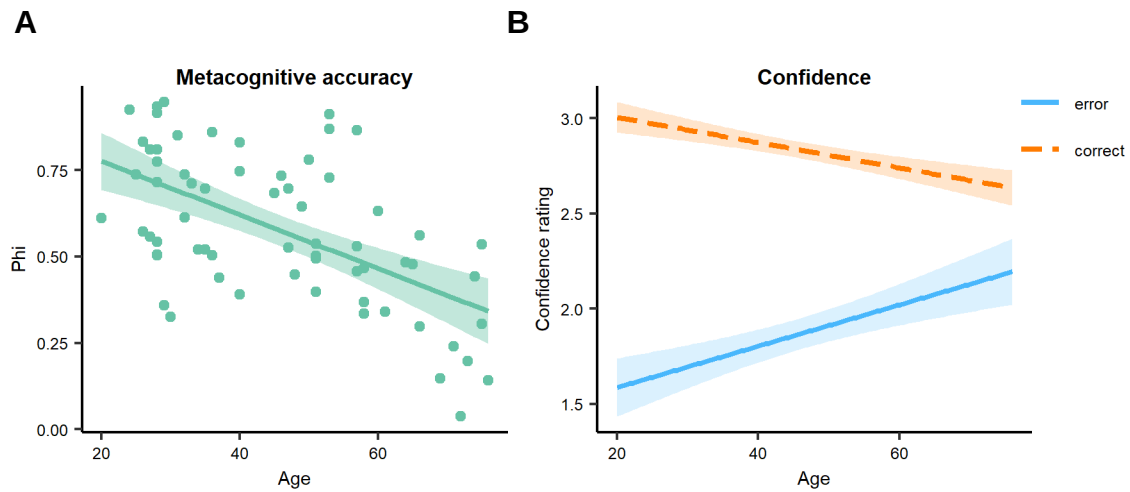


Figure 4

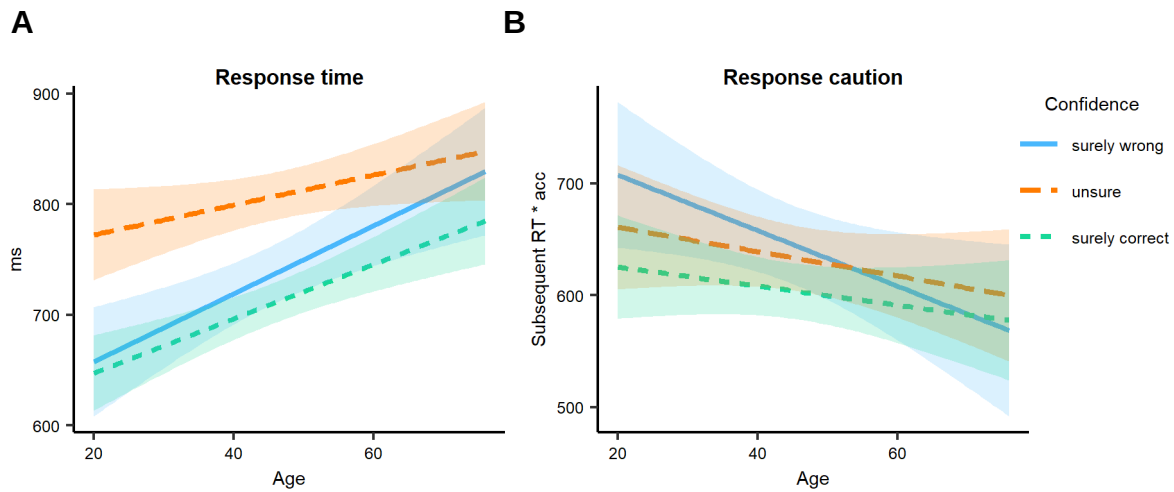


Figure 5

Click here to access/download;Figure(s);Overhoff_metacognition-ageing_FIG5_NbA.pptx

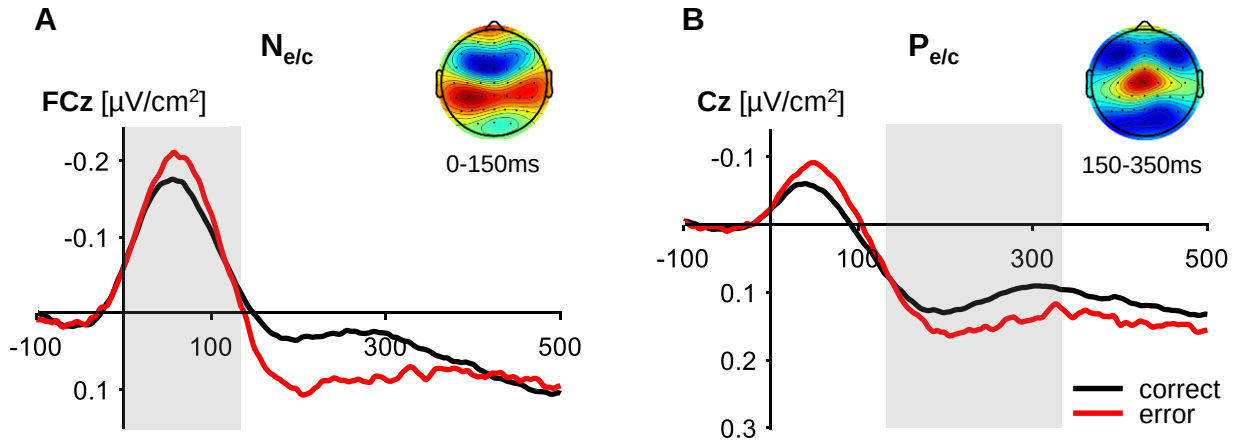
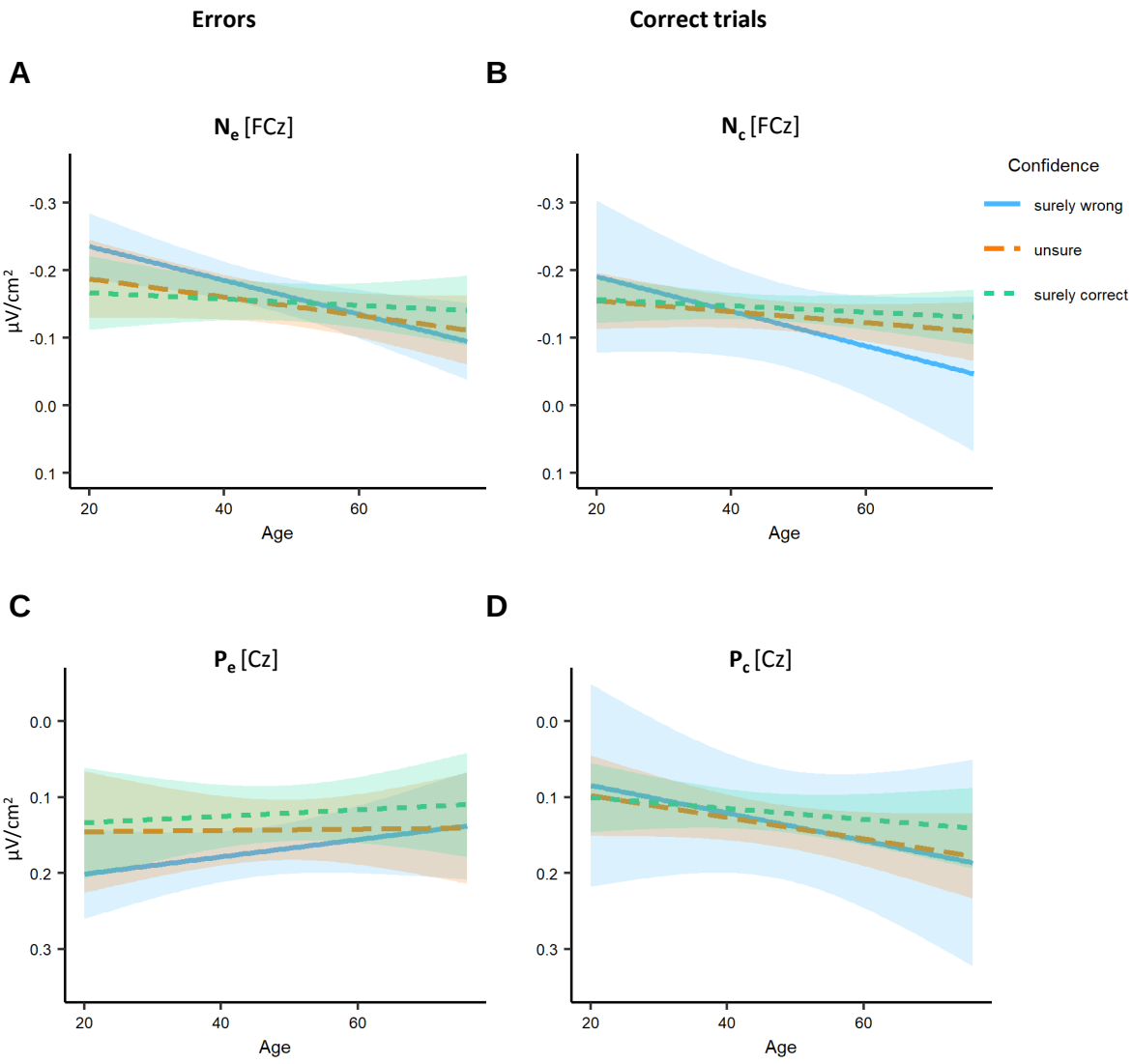


Figure 6

Click here to access/download;Figure(s);REV_Overhoff_metacognitio



Neural correlates of metacognition across the adult lifespan: Supplementary Material

Helen Overhoff[a,b,c*], Yiu Hong Ko[a,b,c], Daniel Feuerriegel[b], Gereon R. Fink[a,d], Jutta Stahl[c], Peter H. Weiss[a,d], Stefan Bode[b], & Eva Niessen[c]

[a]Cognitive Neuroscience, Institute of Neuroscience and Medicine (INM-3), Research Centre Jülich, Leo-Brandt-Str. 5, 52425 Jülich, Germany

[b]Melbourne School of Psychological Sciences, University of Melbourne, Parkville Campus, Parkville 3010, Victoria, Australia

[c]Department of Individual Differences and Psychological Assessment, University of Cologne, Pohligstr. 1, 50969 Cologne, Germany

[d]Department of Neurology, Faculty of Medicine, University Hospital Cologne, University of Cologne, Kerpener Str. 62, 50937 Cologne, Germany

*Corresponding author. Cognitive Neuroscience, Institute of Neuroscience and Medicine (INM-3), Research Centre Jülich, Leo-Brandt-Str. 5, 52425 Juelich, Germany.

Email: h.overhoff@fz-juelich.de

Supplementary Material List of Contents

S1. Mixed effects regression model structures and coefficients for behavioural analyses

For the analysis of behavioural parameters, data were analysed using linear and generalised linear mixed effects models. We always used the between-subjects factor age as the regressor. The within-subject factor of interest was either accuracy (error, correct) or (pooled) confidence (3 levels). We fitted random intercepts for participants and, if possible, random slopes by participant for the within-subject factor of interest. For linear mixed models, *F* statistics are reported, and degrees of freedom were estimated by Satterthwaite's approximation, and for generalised linear mixed models, X^2 statistics are reported.

Significant effects of accuracy or confidence were followed up by pairwise comparisons between error and correct trials or across confidence levels using paired-samples *t*-tests for linear mixed models and *Z*-tests for generalised linear models. Significant interactions were followed up by (generalised) linear mixed regressions, separately for each level of a given within-subject factor to assess potential effects of age. These follow-up tests were chosen because our main interest was in the differential relations between accuracy, confidence, and behaviour across the lifespan rather than between the levels. Post-hoc test results were compared against Holm-corrected alpha levels to account for multiple comparisons.

Error rate (ER)

For the analysis of the error rate, we fitted a generalised linear mixed effects model (binomial family, logit function) testing for effects of confidence and age on accuracy. The variable of age was centred and scaled.

```
Accuracy ~ confidence * age + (1 | sbj)
```

Analysis of Deviance Table with Wald tests.

Predictor	*df*	X^2	*p*
Confidence	2	2200.020	**<.001**
Age	1	4.704	**.030**
Confidence*Age	2	168.125	**<.001**

Regression coefficients for the predictor of age

	Estimate	Std. Error	*z* Ratio
Age	0.37	0.169	2.169

Post-hoc test of contrasts between confidence levels.

Contrast	Estimate	Std. Error	*z* Ratio	*p*
Low – medium	-3.88	0.071	28.872	**<.001**
Low – high	-5.93	0.182	-10.103	**<.001**
Medium – high	-2.05	0.147	-26.395	**<.001**

```
Accuracy [low conf] ~ age + (1 | sbj)

Accuracy [medium conf] ~ age + (1 | sbj)

Accuracy [high conf] ~ age + (1 | sbj)
```

Post-hoc Analysis of Deviance Table with Wald tests and regression coefficients for the predictor of age for each level of confidence.

Predictor	*df*	X^2	Estimate	Std. Error	*p*
Age [low conf]	1	0.898	0.01	0.012	.343
Age [medium conf]	1	3.467	-0.01	0.007	.063
Age [high conf]	1	37.664	-0.05	0.009	**<.001**

Response time (RT)

For the analysis of RTs, we fitted linear mixed effects models testing for effects of accuracy and age or confidence and age on RTs, respectively.

```
RT ~ accuracy * age + (accuracy | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Accuracy	1	61.6	5.572	**.021**
Age	1	62.9	17.358	**<.001**
Accuracy*Age	1	56.5	2.846	.097

Regression coefficients for the predictor of age

	Estimate	Std. Error	*t* Ratio
Age	2.22	0.734	3.023

Post-hoc test of contrasts between error and correct.

Contrast	Estimate	Std. Error	*t* Ratio	*p*
Error – correct	17.9	7.15	2.498	**.015**

```
RT ~ confidence * age + (confidence | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Confidence	2	61.8	27.291	**<.001**
Age	1	63.7	13.305	**<.001**
Confidence*Age	2	56.6	5.187	**.009**

Regression coefficients for the predictor of age

	Estimate	Std. Error	*t* Ratio
Age	3.07	0.845	3.637

Post-hoc test of contrasts between confidence levels.

Contrast	Estimate	Std. Error	*t* Ratio	*p*
Low – medium	-72.3	10.40	-6.952	**<.001**
Low – high	25.3	9.48	2.673	**.010**
Medium – high	97.6	6.74	14.496	**<.001**

```
RT [low conf] ~ age + (1 | sbj)

RT [medium conf] ~ age + (1 | sbj)

RT [high conf] ~ age + (1 | sbj)
```

Post-hoc Analysis of Deviance Table with Wald tests and regression coefficients for the predictor of age for each level of confidence.

Predictor	*Num df*	*Den df*	*F*	Estimate	Std. Error	*p*
Age [low conf]	1	68.4	13.592	3.19	0.866	**<.001**
Age [medium conf]	1	53.9	3.634	1.38	0.725	.062
Age [high conf]	1	62.4	18.358	2.47	0.578	**<.001**

Confidence

For the analysis of confidence, we fitted a linear mixed effects model testing for the effects of accuracy and age on confidence ratings.

```
Confidence ~ accuracy * age + (accuracy | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Accuracy	1	63.4	162.928	**<.001**
Age	1	62.9	2.078	.154
Accuracy*Age	1	62.4	37.361	**<.001**

Post-hoc test of contrasts between error and correct.

Contrast	Estimate	Std. Error	*t* Ratio	*p*
Error – correct	-0.986	0.046	-21.252	**<.001**

```
Confidence [error] ~ age + (1 | sbj)

Confidence [correct] ~ age + (1 | sbj)
```

Post-hoc Analysis of Deviance Table with Wald tests and regression coefficients for the predictor of age for error and correct trials.

Predictor	*Num df*	*Den df*	*F*	Estimate	Std. Error	*p*
Age [error]	1	68.4	17.977	0.01	0.003	**<.001**
Age [correct]	1	53.9	23.816	-0.01	0.001	**<.001**

Behavioural adaptation

For the analysis of behavioural adjustments, we fitted linear mixed effects models testing for effects of accuracy and age or confidence and age on response caution, respectively.

```
Response caution ~ accuracy * age + (accuracy | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Accuracy	1	55.9	12.366	**<.001**
Age	1	62.4	2.141	.148
Accuracy*Age	1	43.9	6.709	**.013**

Post-hoc test of contrasts between error and correct.

Contrast	Estimate	Std. Error	*t* Ratio	*p*
Error – correct	27.6	7.26	3.808	**<.001**

```
Response caution [error] ~ age + (1 | sbj)

Response caution [correct] ~ age + (1 | sbj)
```

Post-hoc Analysis of Deviance Table with Wald tests and regression coefficients for the predictor of age for error and correct trials.

Predictor	*Num df*	*Den df*	*F*	Estimate	Std. Error	*p*
Age [error]	1	60.5	3.836	-1.82	45.978	.055
Age [correct]	1	61.9	0.584	-0.60	0.782	.448

```
Response caution ~ confidence * age + (confidence | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Confidence	2	54.9	7.306	**.002**
Age	1	63.5	2.935	.092
Confidence*Age	2	50.4	2.837	.068

Post-hoc test of contrasts between confidence levels.

Contrast	Estimate	Std. Error	*t* Ratio	*p*
Low – medium	13.0	13.42	0.969	.336
Low – high	42.8	10.67	4.013	**<.001**
Medium – high	29.8	8.78	3.399	**.002**

S2. Mixed effects regression model structures and coefficients for electrophysiological analyses

For the analysis of electrophysiological parameters, data were analysed fitting the same models and performing the same post-hoc tests as for the behavioural data

$N_{e/c}$ amplitudes

For the analysis of the $N_{e/c}$, we fitted a linear mixed effects model testing for effects of accuracy and age on the mean amplitude including all trials.

```
Ne/c ~ accuracy * age + (accuracy | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Accuracy	1	38.6	9.054	**.005**
Age	1	31.3	3.484	.067
Accuracy*Age	1	31.8	5.472	**.026**

Post-hoc test of contrasts between error and correct.

Contrast	Estimate	Std. Error	*t* Ratio	*p*
Error – correct	-0.020	0.007	-2.842	**.007**

```
Ne [error] ~ age + (1 | sbj)

Nc [correct] ~ age + (1 | sbj)
```

Post-hoc Analysis of Deviance Table with Wald tests and regression coefficients for the predictor of age for error and correct trials.

Predictor	*Num df*	*Den df*	*F*	Estimate	Std. Error	*p*
Age [error]	1	55.4	5.030	0.00	0.001	**.029**
Age [correct]	1	62.6	1.124	0.00	0.001	.293

For the analysis of the N_e of errors, we fitted a linear mixed effects model testing for effects of confidence and age on the mean amplitude including only errors.

```
Ne ~ confidence * age + (1 | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Confidence	2	2706.4	4.007	**.018**
Age	1	57.4	4.068	**.048**
Confidence*Age	2	2731.5	3.662	**.026**

Regression coefficients for the predictor of age

	Estimate	Std. Error	*t* Ratio
Age	0.00	0.000	3.000

Post-hoc test of contrasts between confidence levels.

Contrast	Estimate	Std. Error	*t* Ratio	*p*
Low – medium	-0.010	0.013	1.696	.180
Low – high	42.8	10.67	2.795	**.016**
Medium – high	29.8	8.78	0.908	**.**364

```
Ne [low conf] ~ age + (1 | sbj)

Ne [medium conf] ~ age + (1 | sbj)

Ne [high conf] ~ age + (1 | sbj)
```

Post-hoc Analysis of Deviance Table with Wald tests and regression coefficients for the predictor of age for each level of confidence.

Predictor	*Num df*	*Den df*	*F*	Estimate	Std. Error	*p*
Age [low conf]	1	58.1	9.735	0.00	0.001	**.003**
Age [medium conf]	1	37.0	2.394	0.00	0.039	.130
Age [high conf]	1	43.9	0.517	0.00	0.001	.476

For the analysis of the N_c of correct responses, we fitted a linear mixed effects model testing for effects of confidence and age on the mean amplitude including only correct trials.

```
Nc ~ confidence * age + (1 | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Confidence	2	16678.6	0.428	.652
Age	1	220.9	2.573	.110
Confidence*Age	2	16565.6	1.145	.318

$P_{e/c}$ amplitudes

For the analysis of the $P_{e/c}$, we fitted a linear mixed effects model testing for effects of accuracy and age on the mean amplitude including all trials.

```
Pe/c ~ accuracy * age + (accuracy | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Accuracy	1	55.3	10.378	**.002**
Age	1	62.5	0.025	.876
Accuracy*Age	1	49.2	6.443	**.014**

Post-hoc test of contrasts between error and correct.

Contrast	Estimate	Std. Error	*t* Ratio	*p*
Error – correct	0.031	0.011	2.799	**.007**

```
Pe [error] ~ age + (1 | sbj)

Pc [correct] ~ age + (1 | sbj)
```

Post-hoc Analysis of Deviance Table with Wald tests and regression coefficients for the predictor of age for error and correct trials.

Predictor	*Num df*	*Den df*	*F*	Estimate	Std. Error	*p*
Age [error]	1	58.5	0.976	-0.00	0.001	.328
Age [correct]	1	63.5	1.562	0.00	0.000	.219

For the analysis of the P_e of errors, we fitted a linear mixed effects model testing for effects of confidence and age on the mean amplitude including only errors.

```
Pe ~ confidence * age + (confidence | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Confidence	2	57.5	0.810	.450
Age	1	59.2	0.425	.517
Confidence*Age	2	40.6	0.294	.747

For the analysis of the P_c of correct responses, we fitted a linear mixed effects model testing for effects of confidence and age on the mean amplitude including only correct trials.

```
Pc ~ confidence * age + (1 | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Confidence	2	16709	0.313	.732
Age	1	181	1.740	.189
Confidence*Age	2	16650	1.364	.256

S3. Modulation of ERPs by confidence, independent of accuracy

In our main analysis, we fitted linear mixed effects models to the $N_{e/c}$ and $P_{e/c}$ amplitudes of all trials with the within-subject factor accuracy and the between-subject factor age. The amplitudes of both ERPs were larger for errors than correct responses, and the $N_{e/c}$ amplitude decreased with age for errors. As both components have further been shown to be sensitive to variations in confidence (Boldt & Yeung, 2015), we additionally computed the $N_{e/c}$ and $P_{e/c}$ amplitudes in relation to reported confidence for errors and correct trials combined (three levels: 'surely wrong', 'unsure', 'surely correct'). Here, we provide the results for the linear mixed effects regression analyses including confidence instead of accuracy as the within-subject factor.

$N_{e/c}$ amplitude

For the analysis of the $N_{e/c}$, we fitted a linear mixed effects model testing for effects of confidence and age on the mean amplitude including all trials.

$N_{e/c}$ ~ confidence * age + (confidence | sbj)

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Confidence	2	47.2	7.057	**.002**
Age	1	60.6	4.703	**.034**
Confidence*Age	2	40.4	4.989	**.012**

Regression coefficients for the predictor of age

	Estimate	Std. Error	*t* Ratio
Age	0.00	0.001	2.960

Post-hoc test of contrasts between confidence levels.

Contrast	Estimate	Std. Error	*t* Ratio	*p*
Low – medium	-0.030	0.011	-2.811	**.022**
Low – high	-0.026	0.011	-2.449	**.038**
Medium – high	-0.004	0.007	0.649	**.**519

$N_{e/c}$ [low conf] ~ age + (1 | sbj)

$N_{e/c}$ [medium conf] ~ age + (1 | sbj)

$N_{e/c}$ [high conf] ~ age + (1 | sbj)

Post-hoc Analysis of Deviance Table with Wald tests and regression coefficients for the predictor of age for each level of confidence.

Predictor	*Num df*	*Den df*	*F*	Estimate	Std. Error	*p*
Age [low conf]	1	58.1	11.353	0.00	0.001	**.001**
Age [medium conf]	1	45.2	0.453	0.00	0.001	.505
Age [high conf]	1	64.1	0.355	0.00	0.001	.553

$P_{e/c}$ amplitude

For the analysis of the $P_{e/c}$, we fitted a linear mixed effects model testing for effects of confidence and age on the mean amplitude including all trials.

```
Pe/c ~ confidence * age + (confidence | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Confidence	2	52.1	3.817	**.028**
Age	1	63.1	0.076	.783
Confidence*Age	2	47.4	1.452	.244

Post-hoc test of contrasts between confidence levels.

Contrast	Estimate	Std. Error	*t* Ratio	*p*
Low – medium	0.037	0.018	2.091	.083
Low – high	0.053	0.015	3.581	**.002**
Medium – high	0.016	0.010	1.660	**.**103

Together, these results replicate previous findings of a confidence-related modulation of the $N_{e/c}$ and $P_{e/c}$ amplitudes (Boldt & Yeung, 2015; Scheffers & Coles, 2000). Moreover, we provide evidence, for the first time, that the $N_{e/c}$, but not the $P_{e/c}$ was differently modulated by ageing across confidence levels. Notably, it has to be considered that the percentage of errors within each confidence level varied substantially between participants and across the lifespan. However, the same holds for the opposite conclusion, that is, a potential modulation of the ERP amplitudes by accuracy is always inherently connected to confidence (e.g., Fleming et al., 2012). Therefore, the more robust analysis, in our opinion, is the separate examination of correct and incorrect trials, which we report in the main article.

S4. Stimulus-related ERPs of conflict processing

When investigating age-related alterations in neural correlates of response evaluation, the interval between stimulus presentation and response is also informative – in particular the N2 and the P300 components of the ERP. These have been related to error processing as indexing stimulus conflict monitoring (N2) and error-related attention reallocation (P300; Groom & Cragg, 2015; Polich, 2007; Yeung & Cohen, 2006). Research has shown a decline of both components in older age (Korsch et al., 2016; Lucci et al., 2013). Assessing the modulation of these components by age allowed us to draw conclusions about the specificity of potential modulations of the $N_{e/c}$ and $P_{e/c}$ in our metacognitive task. We, therefore, additionally computed the N2 and the P300 components using the stimulus-locked data.

The epochs were cut at 1,500 ms after target stimulus presentation, and the preprocessing was equivalent to the response-locked data. The N2 was quantified as the mean amplitude around the negative peak latency (± 50 ms) of the grand-average ERP in the time window from 150 to 300 ms at Cz, and the P300 around the positive peak latency (± 50 ms) of the grand-average ERP in the time window from 200 to 500 ms at POz (Groom & Cragg, 2015; Klawohn et al., 2020; Polich, 2007). The latencies were retrieved for errors and correct responses, respectively.

N2 amplitudes

For the analysis of the N2, we fitted a linear mixed effects model testing for effects of accuracy and age on the mean amplitude including all trials.

```
N2 ~ accuracy * age + (accuracy | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Accuracy	1	2188.4	0.001	.974
Age	1	58.3	1.586	.220
Accuracy*Age	1	1487.4	0.005	.942

P300 amplitudes

For the analysis of the P300, we fitted a linear mixed effects model testing for effects of accuracy and age on the mean amplitude including all trials.

```
P300 ~ accuracy * age + (accuracy | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Accuracy	1	89.6	0.012	.914
Age	1	60.8	0.281	.598
Accuracy*Age	1	67.0	0.184	.670

The results suggests that both the monitoring of stimulus conflict and the attention-related evaluation of conflict were comparable across the lifespan. This means that age-related differences in early stimulus-related conflict monitoring do not account for subsequent modulations of response processing.

S5. Electrophysiological analyses using untransformed ERP data

In addition to the analysis of the CSD-transformed ERP data reported in the manuscript, we computed the same analyses for the analysis of accuracy using untransformed raw data.

$N_{e/c}$ amplitudes

For the analysis of the $N_{e/c}$, we fitted a linear mixed effects model testing for effects of accuracy and age on the mean amplitude including all trials.

```
Ne/c ~ accuracy * age + (accuracy | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Accuracy	1	48.4	23.326	**<.001**
Age	1	59.3	0.297	.588
Accuracy*Age	1	41.4	10.754	**.002**

Post-hoc test of contrasts between error and correct.

Contrast	Estimate	Std. Error	*t* Ratio	*p*
Error – correct	-1.46	0.257	-5.698	**<.001**

```
Ne [error] ~ age + (1 | sbj)

Nc [correct] ~ age + (1 | sbj)
```

Post-hoc Analysis of Deviance Table with Wald tests for error and correct trials.

Predictor	*Num df*	*Den df*	*F*	*p*
Age [error]	1	51.9	2.728	.105
Age [correct]	1	62.6	1.424	.237

$P_{e/c}$ amplitudes

For the analysis of the $N_{e/c}$, we fitted a linear mixed effects model testing for effects of accuracy and age on the mean amplitude including all trials.

```
Pe/c ~ accuracy * age + (accuracy | sbj)
```

Analysis of Variance Table with Satterthwaite's method.

Predictor	*Num df*	*Den df*	*F*	*p*
Accuracy	1	60.6	5.338	**.024**
Age	1	61.4	0.595	.443
Accuracy*Age	1	54.2	3.143	.082

Post-hoc test of contrasts between error and correct.

Contrast	Estimate	Std. Error	*t* Ratio	*p*
Error – correct	0.746	0.351	2.124	**.038**

In sum, the pattern of results for the raw ERP data was very similar to the results for the CSD-transformed data. The only discrepancy was that two effects did not become significant using the raw data, namely the effect of age in the post-hoc analysis for the N_e of errors, and the interaction between age and accuracy for the $P_{e/c}$ amplitude. As this effect was not large in our reported analysis either, the increased noise in the raw data might have concealed the effect in this analysis.

S6. Bayesian statistics for reported null-effects

In our manuscript, we are reporting frequentist statistics for all analyses in order to keep the statistical framework consistent. However, we computed additional analyses reporting Bayes factors for all null findings, because they constitute an essential part of our conclusion (i.e., null effects of age for response caution and $P_{e/c}$ amplitude).

We ran Bayesian statistical analyses using the package BayesFactor in R (version 0.9.12-4.2; Morey and Rouder, 2018) to assess the extent to which our data support the null effects. In order to examine the null effects of age for response caution and $P_{e/c}$ amplitude, we compared the full models including the within-subject factor of interest (accuracy or confidence) and the between-subject factor age to a null model including only the within-subject factor.

For response caution, we tested the hypothesis that response caution is modulated by accuracy, age, and their interaction against the null hypothesis that it was only modulated by accuracy. We found anecdotal evidence in favour of the null hypothesis (BF_{01} = 2.055). Comparing the model predicting response caution by the factors confidence and age to the model including only age, we found strong evidence supporting the null hypothesis (BF_{01} = 238.612). This suggests that the modulation of response caution was indeed similar across the lifespan.

For the $P_{e/c}$, we assessed evidence for a modulation by age of all trials combined. Here, the model including the interaction term of accuracy and age was around five times more likely given the data than the null model (BF_{10} = 5.264). This is mirroring the significant interaction we found in the analysis reported in the manuscript. For the modulation of errors by confidence and age, we found strong evidence against an effect of age on the Pe amplitude (BF_{01} = 614.830). The same was true for the age effect on the Pc amplitude of correct responses (BF_{01} = 1294.515).

Taken together, these results suggest that our data robustly support the null effects of age on the confidence modulation of response caution and the $P_{e/c}$ amplitude.

S7 Supplementary material reference list

Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *The Journal of Neuroscience*, *35*(8), 3478–3484. https://doi.org/10.1523/JNEUROSCI.0797-14.2015

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280–1286. https://doi.org/10.1098/rstb.2012.0021

Groom, M. J., & Cragg, L. (2015). Differential modulation of the N2 and P3 event-related potentials by response conflict and inhibition. *Brain and Cognition*, *97*, 1–9. https://doi.org/10.1016/j.bandc.2015.04.004

Klawohn, J., Santopetro, N. J., Meyer, A., & Hajcak, G. (2020). Reduced P300 in depression: Evidence from a flanker task and impact on ERN, CRN, and Pe. *Psychophysiology*, *57*(4), 1–11. https://doi.org/10.1111/psyp.13520

Korsch, M., Frühholz, S., & Herrmann, M. (2016). Conflict-specific aging effects mainly manifest in early information processing stages-an ERP study with different conflict types. *Frontiers in Aging Neuroscience*, *8*, 1–12. https://doi.org/10.3389/fnagi.2016.00053

Lucci, G., Berchicci, M., Spinelli, D., Taddei, F., & Di Russo, F. (2013). The Effects of Aging on Conflict Detection. *PLoS ONE*, *8*(2). https://doi.org/10.1371/journal.pone.0056566

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs.* (R package version 0.9.12-4.2). https://cran.r-project.org/package=BayesFactor

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Scheffers, M. K., & Coles, M. G. H. (2000). Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 141–151. https://doi.org/10.1037/0096-1523.26.1.141

Yeung, N., & Cohen, J. D. (2006). The impact of cognitive deficits on conflict monitoring predictable dissociations between the error-related negativity and N2. *Psychological Science*, *17*(2), 164–171. https://doi.org/10.1111/j.1467-9280.2006.01680.x