Journal of Memory and Language 147 (2026) 104728

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Intermediate decisional and response states in lexical decision: evidence from electromyography and metacognitive confidence ratings

Michele Scaltritti [a,*], Saman Kamari Songhorabadi [a], Simone Sulpizio [b,c]

[a] Dipartimento di Psicologia e Scienze Cognitive - Università degli Studi di Trento, Corso Bettini 31 – 38068 Rovereto (TN), Italy
[b] Dipartimento di Psicologia – Università degli Studi di Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1 – 20126, Milano, Italy
[c] Milan Center for Neuroscience (NeuroMI), Università degli Studi di Milano-Bicocca, Italy

ARTICLE INFO

Keywords:
Lexical decision
Confidence
Metacognition
Motor-response execution

ABSTRACT

The lexical decision paradigm relies on a binary response configuration (word vs. nonword) and a single dichotomous decisional outcome (correct vs. error). The present research used single-trial electromyographic recordings of the responses and metacognitive confidence ratings to gain insight into a) intermediate decisional states and b) potential misalignments between objective and subjective outcomes of the lexical decision process. The results revealed that confidence was high for correct responses, signaling a relatively clear-cut categorization even for more ambiguous stimuli such as low-frequency words and wordlike pseudowords. The dominant factors in shaping metacognitive judgments were stimulus properties and premotor processing time, whereas the contribution of the fluency of the motor responses emerged only under specific conditions. Errors revealed a more complex pattern, with limited conscious detection in the case of ambiguous stimuli. Whereas errors for words were seemingly driven by unresolved decision processes, those for nonwords were mostly determined by lexical competition. Importantly, although these latter errors partially bypassed response control mechanisms, they were more accessible to awareness, pointing to a partial dissociation between error correction and detection. By focusing on the decisional layer of lexical decision, these findings begin to shed light on the specific dynamics that characterize decision-making tasks grounded in memory-based evidence and lexical knowledge.

Introduction

The lexical decision task is an important paradigm for word recognition and decision-making models (e.g., Coltheart et al., 2001; Ratcliff et al., 2004). In its visual variant, participants categorize letter strings as either words or nonwords, typically responding with discrete manual actions such as button presses. Although this binary task exemplifies a wide range of decisional scenarios based on memory and recognition, the two-choice format may constrain the expression of a broader spectrum of decisional states. The present research uses single-trial metacognitive confidence ratings and electromyographic (EMG) measures of response activation to investigate a) decisional outcomes potentially hidden by the binary response and b) the alignment between objective performance (correct vs. error) and subjective metacognitive evaluations.

Different models converge in characterizing the lexical decision task as a process of accumulation of noisy information (evidence), which terminates when a decision threshold is reached (Dufau et al., 2012; Ratcliff et al., 2004; see also Davis, 2010; Norris, 2006; Wagenmakers et al., 2004; Wagenmakers et al., 2008). The crossing of the threshold (i.e., the collection of a pre-determined amount of information) marks the decision point and triggers the corresponding response. Despite theoretical and architectural differences, these models share a dichotomous view of lexical decision performance: responses are either "word" or "nonword".

While this dichotomous perspective has yielded substantial insights, classic distinctions between lexical and decision processes (e.g., Balota & Chumbley, 1984) suggest that lexical decisions may not be fully explained by lexical access alone. Decision processes introduce an additional layer that contributes to response generation, and, in this perspective, the apparent simplicity of a binary classification can mask heterogeneous decisional states that are ultimately collapsed into the two available responses. For instance, ambiguous items—such as low-frequency words or wordlike nonwords—can sometimes elicit correct responses based on guessing or provisional judgments (Diependaele et al., 2012) rather than on a clear categorization in terms of lexical

* Corresponding author at: Dipartimento di Psicologia e Scienze Cognitive, Università degli Studi di Trento, Corso Bettini 31, 38068 – Rovereto (TN), Italy.
E-mail addresses: michele.scaltritti@unitn.it (M. Scaltritti), s.kamarisonghorabadi@unitn.it (S. Kamari Songhorabadi), simone.sulpizio@unimib.it (S. Sulpizio).

https://doi.org/10.1016/j.jml.2025.104728
Received 1 July 2025; Received in revised form 13 December 2025; Accepted 17 December 2025
Available online 24 December 2025

status. Moreover, because nonwords lack stored memory representations (Dufau et al., 2012), responses to these stimuli, even when accurate, may involve greater uncertainty.

These inherent ambiguities may also widen the gap between objective performance (e.g., accuracy) and subjective awareness (e.g., metacognitive sensitivity; Fleming, 2024). For instance, a low-frequency word outside a participant's lexicon (*adipocere*), or a wordlike nonword (*elethant*) may lead to an incorrect response judged as correct. Both intermediate decisional states (e.g., uncertainty, guessing) and misalignments between objective and subjective outcomes can be assessed through metacognitive measures of decisional validity (Balsdon et al., 2020; Desender et al., 2021; Fleming, 2024; Rahnev et al., 2020; Rahnev, 2021). A common method involves asking participants to rate their confidence in the correctness of each decision (Charles & Yeung, 2019; Yeung & Summerfield, 2012). These ratings provide access to finer decisional states beyond binary outcomes, via a subjective assessment of the probability of response correctness.

To this aim, participants in the present study rated their confidence after each lexical decision trial, using a (pseudo)continuous scale ranging from "sure error" to "sure correct". To prevent strategic approaches and probe a heterogeneous sample of decisions, stimuli included different types of items: high- and low-frequency words, pseudowords created by changing a single letter in existing words (one-letter pseudowords), and pseudowords generated by changing two or more letters or by combining legal Italian syllables. Whereas confidence ratings were expected to be highest for high-frequency words, predictions for the other items are less straightforward.

For instance, correct responses to low-frequency words may be associated with reduced confidence, depending on the prevalence of guessing and whether the effortful recognition process influences metacognitive evaluations, potentially leading to residual uncertainty even after successful classification. Further, pseudowords may generally yield more uncertain decisional states, as participants cannot match the stimulus with a stored memory representation. However, one-letter pseudowords may allow the use of verification processes (e.g., Perea et al., 2005). By comparing the stimulus (*elethant*) against activated word representations (*elephant*), participants can detect the mismatch and respond with relatively high confidence—unlike more traditional pseudowords (e.g., *flirp*), whose status may remain uncertain and yield lower confidence.

Finally, confidence ratings help distinguish between qualitatively different errors. For instance, an incorrect "nonword" response to a low-frequency word, rated as "surely correct", suggests that the word is absent from the participant's lexicon and was consequently misclassified as a nonword. Conversely, the same decisional outcome (an incorrect "nonword" response) can reflect momentary slips or guessing when confidence ratings indicate full awareness of the error or complete uncertainty concerning response correctness. For pseudowords, unidentified errors may reflect lexical capture, that is, a misclassification driven by the item's similarity to a specific known word.

Confidence ratings offer one window into decisional states, but lexical decision also produces provisional responses that may provide additional insights into the evolving decisional process. Studies using mouse-tracking have shown that participants can initiate an incorrect response but then abruptly change direction, signaling a change of mind (Barca & Pezzulo, 2015; see also Barca & Pezzulo, 2012). These studies share conceptual parallels with the present work, but EMG recordings offer a more direct measure of partial errors, in the form of covert activations of the incorrect response hand before the final correct response is executed (Eriksen et al., 1985; Hasbroucq et al., 1999). This phenomenon challenges the assumption that motor activity begins only after evidence crosses a final decision threshold (e.g., Dufau et al., 2012; Grainger & Jacobs, 1996; Ratcliff et al., 2004) and reveals temporary decisional states (e.g., Servant et al., 2015) that, despite reaching the effectors, dynamically shift to the opposite option. Partial responses are informative in two ways. First, within perceptual decision-making tasks, trials featuring partial responses have been found to yield higher levels of confidence, compared to those without partial response activations. (Gajdos et al., 2019). Thus, partial responses may represent a cue for metacognitive judgments. Second, partial errors reveal response-control mechanisms (Burle et al., 2002), in the form of corrective mechanisms acting on the unfolding responses. This type of online cognitive control may inform metacognitive awareness about the decisional outcome.

The role of response control in shaping decision confidence can also be understood by examining the time-course of correct and incorrect responses. When plotting accuracy as a function of response time (i.e., conditional accuracy functions; Ollman, 1977), nonwords, especially wordlike ones, show more fast, impulsive errors (Fiora et al., 2026; Scaltritti et al., 2021, Scaltritti, Giacomoni et al. 2023; Grisetto et al., 2025a; 2025b), suggesting a failure in controlled processing and response control. Complementary findings emerge from the time-course of incorrect activations (i.e., partial and full errors jointly considered; conditional incorrect activation functions; Fluchère et al., 2018; Ramdani et al., 2015). Early incorrect activations are in fact frequent for both words and nonwords, but – as indicated by conditional accuracy functions –they are often corrected in word trials, whereas they more often become overt errors for nonwords. Response control thus seems more effective for words, in preventing early incorrect activations from turning into full-blown errors. Whether this asymmetry also influences metacognitive awareness remains unclear. Fast errors that bypass response control, in fact, may be less accessible to awareness, a pattern that could be especially prominent with pseudoword stimuli.

Finally, single-trial EMG recordings allow decomposing the global response time (RT) into two successive chronometric intervals (Botwinick & Thompson, 1966): Premotor time (PMT), the interval between stimulus onset and EMG onset, and motor time (MT), the interval between EMG onset and the completion of the button-press. Relative to the global RT measure, PMT reflects the bulk of the recognition and decision processes (e.g., how quickly information reaches the decision threshold), while MT traces motor-response execution. Traditionally, models of lexical decision and decision-making have treated this latter aspect as a non-decision interval (e.g., Ratcliff et al., 2004). In this view, PMT appears as functionally equivalent to RT in indexing cognitive processing, with MT merely reflecting a non-cognitive constant. However, empirical evidence suggests that cognitive variables can propagate their influence onto motor execution, with both PMT and MT being influenced by the ongoing decisional process (e.g., Scaltritti, Giacomoni et al., 2023; Servant et al., 2021; Weindel et al., 2021). Both intervals may thus contribute to metacognitive confidence and, while shorter RTs are generally associated with higher confidence (Vickers & Packer, 1982; Kiani et al., 2014), this general relationship may be independently driven by the two intervals. Premotor time, capturing the largest portion of decisional dynamics, presumably reflects the primary link between processing difficulty and decision confidence. The role of MT is less straightforward, but there are at least three reasons to expect its involvement in metacognitive confidence.

First, recent models of metacognitive confidence emphasize the involvement of the whole perception–action cycle in confidence formation (e.g., Gajdos et al., 2019; Fleming & Daw, 2017). Participants could access information about the duration of motor-response execution (Pavailler et al., 2025) and use it as an additional cue when evaluating the decisional outcome. Second, a link between confidence and motor-parameters has already been proposed to explain differences in response force between words and nonwords. One hypothesis suggests that the difference in evidence accumulation rates between competing response channels maps onto confidence, which in turn affects acceleration and force after movement initiation. In these studies, the response involved pulling a handle in different directions (Balota & Abrams, 1995). Whether MT, a chronometric measure for discrete button-press responses, relates to confidence in a similar way remains to be tested. If so, differences in MT should correspond to differences in reported confidence. Third, MT also captures response control processes which,

as discussed above, could influence metacognitive evaluations. Longer MTs are in fact observed for incorrect compared to correct responses (e.g., Allain et al., 2004; Weindel et al., 2021), suggesting the activation of inhibitory mechanisms attempting unsuccessfully to halt an erroneous response. It remains an open question whether these control-driven variations in response durations, particularly in error trials, are associated with an enhanced awareness concerning the decisional outcome.

Taken together, these issues seem to fall outside the scope of lexical decision models, which focus on dichotomous responses and envisage a serial transition from decision to response stages. The present study, instead, aimed to provide an empirical assessment of the range of decisional states, integrating behavioral data with single-trial EMG signals (from the muscle responsible for button-press responses) and metacognitive confidence ratings.

The article first provides an analysis of the behavioral performance as a function of stimulus type. In addition to classic RTs, chronometric analyses include measures of PMT and MT (e.g., Scaltritti, Giacomoni et al., 2023). Analyses of response accuracy also included partial errors, as well as variations in the time-course of both accurate responses and incorrect activations, highlighting differences in impulsive response tendencies and in response control as a function of stimulus type (Fiora et al., 2026; Scaltritti et al., 2025). Next, confidence ratings are considered, linking them to the behavioral findings. Other than assessing confidence differences across stimulus types, we examine how confidence varies as a function of response latency (both in terms of PMT and MT) and response accuracy. For accurate responses, PMT is expected to reflect more general decision and lexical access processes, whereas MT may reveal motor-specific contributions to confidence. For errors, PMT should distinguish between fast impulsive errors and slower recognition failures, possibly highlighting differences in error awareness between these qualitatively different types of errors. Specifically, pseudoword errors resulting from uncorrected impulsive activations (i.e., shorter PMTs) may be linked with a decreased awareness. Variations in MT may further capture subtler roles of response control, with longer MTs signaling attempts to inhibit erroneous responses and thereby increasing error awareness.

To enhance the fit of the statistical models, as well as to address different states of awareness concerning errors, separate analyses were conducted for a) the likelihood of errors with full awareness (incorrect responses identified as such with maximal confidence), b) the likelihood to incorrectly identify errors as correct responses (incorrect responses that participants rated as correct with maximal confidence), and c) confidence levels for uncertain errors (incorrect responses with intermediate ratings).

Method

Data availability

Data, materials and scripts are available at: https://osf.io/qhdgy/?view_only=b2a1eb7346224511b0f6bf05ebc24516.

Participants

Sample size estimation was based on recent guidelines in the field (Brysbaert, 2019). Forty-eight Italian native speakers took part in the experiment. Data from two participants were removed because their mean response accuracy fell below the threshold of the 2.5 SDs from the sample mean (i.e., accuracy below .65). The final sample included 46 participants (30 females, M_{age} = 23.30; SD_{age} = 3.25). Participants reported normal or corrected vision, and no history of learning disabilities or neurological issues. After the administration of the Edinburgh Handedness Inventory (Oldfield, 1971), 41 participants were classified as pure right-handers (M = 81.23, SD = 14.74), 1 as pure left-hander (with a handedness score of − 60), 2 as neutral (handedness score = 50), 1 as mixed right-hander (40), and 1 as a mixed left-hander (− 22.2). The study was approved by the ethical committee of the University of Trento (protocol number 2023–064), and participants signed an informed consent document prior to the beginning of the experimental procedures. Participation was compensated with 20€.

Stimuli

One-hundred and fifty high-frequency words, and 150 low-frequency ones were extracted from the PhonItalia database (Goslin et al., 2014). Care was taken to also differentiate high- and low-frequency words across a series of variables measuring orthographic similarity with other items of the lexicon (Table 1), whereas the two categories were comparable in terms of number of letters and syllables.

A set of 150 nonwords was created by changing one letter within existing words (one-letter pseudowords), ensuring they remained orthographically and phonologically legal. The base words used to create these items were different from those selected as high- and low-frequency items, except for two high-frequency words (*oceano* – ocean; valanga – *avalanche*) that were inadvertently used to create one-letter pseudowords (*oceato* and *valanta*). Another set of 150 nonwords was manually created either by combining syllables that are phonotactically and orthographically well-formed in Italian, or by changing multiple letters (two or more) within existing words. The syllables were not drawn from a predefined set; rather, they were selected intuitively by the authors. The resulting strings conformed to Italian phonotactic rules (e.g., legal syllable structure, permissible consonant clusters).

One-letter pseudowords and pseudowords were comparable in terms of all the variables listed in Table 1, except for mean bigram frequency (significantly higher for one-letter pseudowords). Words (high- and low-frequency, jointly considered) and nonwords (one-letter pseudowords and pseudowords, jointly considered) were comparable in terms of all the variables listed in Table 1, except for bigram frequency which was higher for nonwords. Each of the 4 categories of stimuli was further partitioned into 2 subsets, for counterbalancing purposes. Within each stimulus category, the two subsets were comparable for the variables listed in Table 1.

Apparatus and procedure

Participants first completed a questionnaire collecting demographic and health-related information. After installing the EMG electrodes (see EMG Recording and Processing), the experimental procedure began. The experiment was conducted using E-Prime 3 software (version 3.0.3.80) on a laptop. Participants were seated approximately 50 cm away in front of the screen, holding two cylindrical handheld buttons (one in each

Table 1
Psycholinguistic variables for the 4 categories of stimuli.

Variables	HF	LF	1L-PW	PW	t
Frequency (log)	4.43	0.38	–	–	
N. of letters	6.89	6.85	7.04	7.00	1.13
N. of syllables	2.91	2.97	3.07	2.99	1.5
Orthographic N	2.36	1.57	2.21	1.83	0.28
OLD20	2.14	2.32	2.22	2.34	1.03
Bigr. freq. mean	117,931	81,037	124,216	108,709	5.96

Note. HF = high-frequency words; LF = low-frequency words; 1L-PW = one-letter pseudowords; PW = pseudowords; N. of Letters = number of letters; Orthographic N = orthographic neighborhood; OLD20 = orthographic Levenshtein distance to the twenty closest neighbors (Yarkoni et al., 2008); Bigr. freq. mean = mean bigram frequency. Words' variables were extracted from the PhonItalia database (Goslin et al., 2014). For pseudowords, the number of orthographic neighbors and OLD20 were computed with reference to the PhonItalia database using functions from the *vwr* package (Keuleers, 2013) in R. Bigram frequency values were computed using a custom-made script with reference to the same database. Reported *t*-values are from independent samples, two-tailed *t*-test comparing words vs nonwords.

hand) connected to a Blackbox Toolkit module. They were instructed to categorize letter strings as either words or nonwords by pressing the corresponding button with their thumbs and to provide a confidence rating on the accuracy of each response using a dedicated scale. A pedal-button was placed below their dominant foot, and participants were instructed to press it after they provided their rating, to move to the next trial.

The experiment consisted of two main blocks, featuring different stimulus − (word vs pseudoword) response (left- vs right-hand) mappings. This was done to ensure, within each participant, an equal number of responses from the two hands for each category of stimuli. The order of the 2 stimulus–response mapping and the assignment of the 2 sets of stimuli each stimulus–response mapping were counterbalanced across participants. Within each block, a self-terminated break was prompted every 60 trials. Each block was preceded by 16 practice trials. After each trial, participants were asked to assess their confidence concerning response accuracy. Ratings were given on a continuous vertical scale with 100 steps (Rahnev et al., 2020). The two extremes of the scale were marked with tick marks and verbal labels ("sure correct" vs "sure error"). The displacement of the labels along the vertical axes (top vs bottom) was counterbalanced across participants. The scale featured 5 additional equidistant tick marks (without any label), with the 3rd one falling exactly at the middle of the scale (García-Pérez & Alcalá-Quintana, 2023). Participants could move a slider by pressing the same buttons used to perform the response during the lexical decision trials. The left-button commanded an upward movement of the slider, whereas the right-hand commanded a movement in the opposite direction. The position of the slider was updated in cycles of 34 ms. Hence, when participants kept the buttons pressed, they had the impression of a continuous movement of the slider over the scale. Once they reached the desired level, they were instructed to press the pedal-button to terminate the current trial and start the new one. The use of a visual analog scale, rather than a discrete *n*-point scale, was intended to maximize the sensitivity of confidence measurements and to allow for the expression of a potentially wide range of decisional states, without artificially constraining responses into predefined categorical labels.

Stimuli were displayed in 26-pt Courier New font, in black on a gray background (RGB = 190, 190, 190). Each trial started with a fixation cross, with a random duration corresponding to 700, 750, 800, or 850 ms. The stimulus immediately followed and was displayed until response or until the response deadline (1200 ms) was met. After a blank screen lasting 800 ms, the confidence scale was displayed and remained on the screen until the participant pressed the pedal-button. A blank screen of 1 s. served as the inter-trial interval. No feedback was given on responses in any stage of the experiment.

EMG recording and processing

The EMG activity from the flexor pollicis brevis of both hands was acquired via an eego sports system (ANT Neuro) with a 1,000 Hz sampling rate. Two pairs of disposable bipolar electrodes, spaced approximately 2 cm apart, were placed on the thenar eminences of both hands, with the ground electrode placed over the pisiform bone of the right wrist. Prior to electrode placement, isopropyl alcohol and a mild abrasive gel (Nuprep) were applied on the recording sites to optimize signal quality. EMG recordings were monitored in real-time, and participants were instructed to relax whenever noise stemming from tonic activity became apparent. Offline (pre)processing procedures were conducted using EEGLAB (version 14_1_2b; Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) functions in MATLAB (version 2018b, MathWorks Inc., Natick, MA, USA), together with custom scripts.

A 5 Hz (order 2 Butterworth) high-pass filter was first applied on the continuous signal, followed by a 50 Hz notch filter. Epochs from −850 until 1900 ms were segmented, with 0 corresponding to stimulus onset. Within each epoch, the onset of the response-related EMG activity was identified using the integrated profile method (Liu & Liu, 2016). To assist the detection of artifacts and partial responses, a second script identified, within each epoch, windows of EMG activity corresponding to the samples in which the rectified EMG signal exceeded a threshold of 3.5 SDs above the mean activity computed during the pre-stimulus baseline (from −500 to 0 ms). Windows separated by intervals shorter than 25 ms were merged, whereas windows shorter than 50 ms or beginning after the button-press were dropped. Finally, the script highlighted the epochs in which 2 or more windows of activity were detected. All the epochs were inspected and scored as valid only when the EMG onset was accurately detected in correspondence to the window of activity capturing the response EMG burst. This was done to exclude artifacts from signal drift, noise, and false starts. On average, 1 % (SD = 1.64 %) of the trials were rejected following these criteria.

The same procedure was applied to the EMG signal of the non-responding hand, to identify partial errors and partial correct responses. Epochs featuring at least one window of activity within these channels were marked. Each epoch was then visually inspected and partial responses were scored when a) the EMG activation was clustered in a visually clear burst and b) the onset was accurately detected. Partial errors were detected, on average, in 6.42 % of the trials (SD = 4.53 %). Partial correct responses were very few (M = 0.74 %, SD = 0.82 %) and thus not considered.

Lastly, the 2 algorithms for onset detection and for the identification of windows of EMG activity were applied once more to the channel of the responding hand, this time limited to the interval going from stimulus onset until the onset of the final EMG response. The aim was to track hesitations, consisting of covert EMG activations preceding the one leading to the button press. Again, each epoch was screened and, as for partial errors, epochs were marked for hesitations when a) the EMG activation was clustered in a visually clear burst and b) the onset was accurately detected. Hesitations were found, on average, to affect 3.82 % of the trials (SD = 1.92).

Measures

Chronometric measures

The analyses of chronometric measures only included pure-correct responses (i.e., no partial errors or hesitations) for which the EMG onset was correctly identified. For each trial, RTs were divided into PMTs (from stimulus onset until the onset of the EMG activity) and MTs (from the onset of the EMG burst until the button-press). RTs, PMTs, and MTs were separately analyzed.

Response accuracy

Analyses of response accuracy included all trials with a valid EMG onset, irrespective of the presence of partial EMG activation in the non-responding hand. Trials in which participants failed to deliver their response within the deadline were instead discarded (2.52 %).

Conditional accuracy functions were used to assess variations in accuracy as a function of latency and were computed by partitioning trials (for each participant and condition) into five quantiles as a function of EMG onset. Quantiles were treated as a fixed effect in the analyses. Conditional incorrect activations functions were computed in the same way, but with the additional inclusion of partial errors, to measure the variations in the proportion of incorrect EMG activations as a function of their latency.

The analysis of partial errors focused on correct responses and thus assessed variations in the likelihood of covert incorrect responses as a function of Stimulus Category. Finally, correction likelihood was assessed on trials featuring overt or covert error responses (i.e., full and partial errors), with the aim to estimate the probability with which an incorrect EMG activation is successfully corrected.

Confidence ratings

Original ratings collected by the software ranged over a scale from −50 ("surely error") to 50 ("surely correct"). For the analyses, scores

were converted to a scale from 0 to 1. Consistent with theoretical frameworks integrating decision confidence and error monitoring (e.g., Yeung & Summerfield, 2012), the scale indexes the subjective probability of correctness, via a monotonic (pseudo)continuum: 0 represents maximal certainty that the response was an error (i.e., full error awareness), 0.5 represents maximal uncertainty (guessing), and 1 represents maximal certainty that the response was correct. Confidence ratings for correct and incorrect responses were separately analyzed. For both accurate and incorrect responses, the analyses included only those trials featuring a valid detection of the response-related EMG burst, irrespective of the presence of partial EMG activations across both channels.

In terms of confidence ratings, error trials displayed a somewhat bimodal distribution. Out of the total 3960 error trials considered in the analyses, 1264 were rated as "surely error" (34.25 %), 631 as "surely correct" (17.10 %). The peculiar distribution, other than presenting suboptimal fit when addressed with a single linear model, arguably highlights the presence of qualitatively different classes of errors. Trials with incorrect responses were thus classified as either errors with full awareness (i.e., with a confidence rating corresponding to "surely error"), misidentified errors (i.e., with a confidence rating corresponding to "surely correct"), or an uncertain error (all other cases). Different complementary analyses were conducted on the three types of incorrect responses (see below).

Statistical analyses

Data were analyzed using linear mixed-effects models or generalized mixed-effects models in the case of dichotomous dependent variables. For the latter, the default settings were modified to allow an increased number of iterations (2^5) and the *bobyqa* algorithm was implemented during the second-stage model optimization.

The significance of fixed effects was assessed via likelihood ratio tests, comparing models with and without the fixed term of interest. A fixed effect was considered significant if its removal led to a significant reduction in explained variance. At this stage, the random-effects structure was restricted to random intercepts for participants and items. Once significant effects were identified, the maximal random-effects structure was implemented, including random slopes (and correlations) for all the significant fixed effects (Barr et al., 2013). In case of convergence issues, the random-effects structure was progressively simplified by removing, in order of priority: random effects associated with zero variance, correlation parameters, and finally the random slopes associated with the smallest variance components.

In all the analyses, the experimental manipulation of Stimulus Category was assessed as a single fixed effect consisting of four levels (high-frequency, low-frequency, one-letter pseudoword, pseudoword). For conditional accuracy/incorrect activations functions, fixed effects included Stimulus Category, Quantiles (of the PMT), as well as their interaction. In modeling the Quantile variable, second order polynomials were assessed to accommodate non-linear effects.

For the analysis of response confidence in correct trials, the predictors included Stimulus Category, the chronometric measures of PMT and MT (both partitioned into 5 quantiles) and the Partial Error factor (i.e., whether the trial contained a partial activation from the non-response hand). The starting model also included the two-way interactions between Stimulus Category and the other predictors.

For confidence ratings in error trials, three complementary analyses were conducted. The first analysis focused on errors with full awareness. Generalized mixed-effects models featuring a logit link function were used to assess the likelihood of this type of errors (coded as 1) against all others (coded as 0). The same logic was applied to the analyses of misidentified errors (errors rated as correct responses with maximal levels of confidence were coded as 1, others as 0). Due to the relatively low number of observations in these two analyses, particularly within certain categories of stimuli, models were conducted on the proportion of events (errors with full awareness/misidentified errors) weighed by the number of events observed.[1] For uncertain errors, linear mixed-effects models were used, and confidence was treated as a continuous dependent variable. Other than Stimulus Category, these models also included PMT and MT (again, partitioned into 5 quantiles) as predictors, together with their interactions with Stimulus Category. Partial correct responses were not considered due to the very low number of occurrences.

For all models of response confidence, fixed terms that failed to reach conventional significance ($p < .05$) were dropped. The surviving terms were included in the random-effect structure and, if in this passage any fixed term turned to be non-significant, both the fixed and the random terms in question were excluded from the model.

All pairwise comparisons were conducted on estimated marginal means, using a Bonferroni correction to prevent Type I error inflation. The analyses were conducted using the packages *lme4* (version 1.1–35; Bates et al., 2015), *afex* (1.4–1; Singman et al., 2021), and *emmeans* (1.10.5) in R (version 4.3.0; R Core Team, 2015). Figures were made through the *ggplot2* package (version 3.5.1; Wickham, 2016).

Results

Chronometric measures

Results for chronometric measures are summarized in Fig. 1, and pairwise comparisons for all models are reported in Table 2.

Response times displayed a significant effect of Stimulus Category, $\chi^2(3) = 528.02$, $p < .001$. Pairwise comparisons revealed faster RTs for high-frequency words compared to all other categories of stimuli. Low-frequency produced faster RTs compared to one-letter pseudowords but not compared to pseudowords. No significant differences in RTs were observed between one-letter pseudowords and pseudowords. Results on PMTs closely mirrored those on RTs, with a significant effect of Stimulus Category, $\chi^2(3) = 516.48$, $p < .001$, followed by the same pattern of significant differences in terms of pairwise comparisons between categories of stimuli (Table 2). Motor times revealed a quite different pattern. The effect of Stimulus Category was still significant, $\chi^2(3) = 126.62$, $p < .001$. There was no significant difference in MTs across high- and low-frequency words (Scaltritti, Giacomoni et al., 2023). However, both displayed shorter MTs compared to one-letter pseudowords as well as pseudowords. Finally, there was no difference in MTs across one-letter pseudowords and pseudowords.

Response accuracy

The effect of Stimulus Category on response accuracy was significant, $\chi^2(3) = 313.18$, $p < .001$. Responses were more accurate for high-frequency words (*Est.* = .98, *SE* = .003), compared to all other stimuli (all *ps* < .01). Low-frequency words, instead, yield the lowest accuracy (*Est.* = .76, *SE* = .034), compared to all other stimuli (all *ps* < .001). Finally, accuracy was lower for one-letter pseudowords (*Est.* = .92, *SE* = .012), compared to pseudowords (*Est.* = .96, *SE* = .006; $p < .001$).

Conditional accuracy functions revealed a significant interaction between Stimulus Category and Quantiles, $\chi^2(3) = 317.87$, $p < .001$, and second order polynomials improved the model's fit, $\chi^2(4) = 98.14$, $p < .001$. As visible in Fig. 2, the interaction is mainly driven by the increased likelihood of fast errors for non-lexical items. A significant Stimulus Category by Quantiles interaction also emerged within conditional inaccurate activations functions, $\chi^2(3) = 399.68$, $p < .001$, with a better fit when fitting the Quantile variable using second order orthogonal polynomials, $\chi^2(4) = 282.26$, $p < .001$. Interestingly, an

[1] The results, in terms of significant fixed effects, were qualitatively similar to those obtained with the standard approach, albeit the fit with the empirical data was improved.

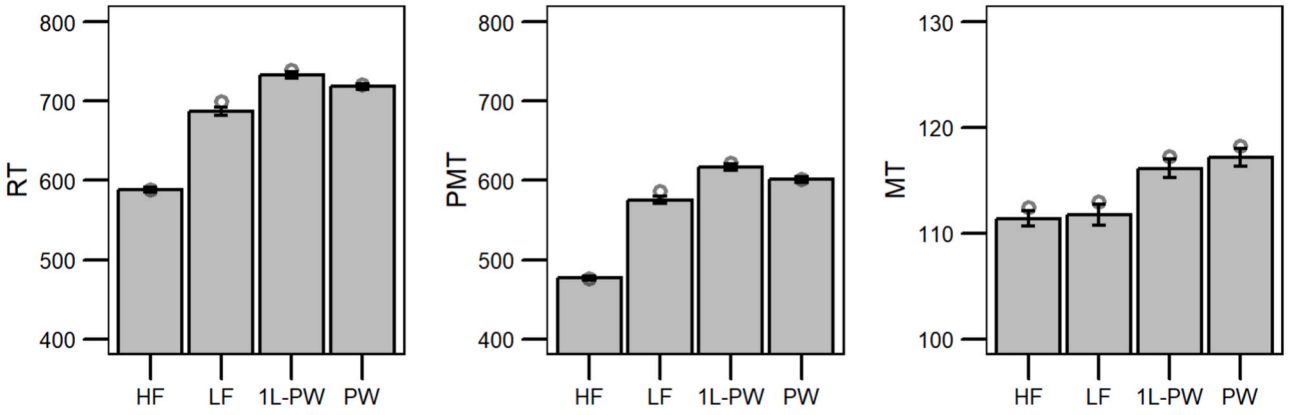

Fig. 1. Results for reaction time (RT), premotor time (PMT), and motor time (MT). Empty circles represent estimated marginal means. Error bars display 95% confidence intervals, adjusted for within-participants variables following Morey (2008). HF = high-frequency words; LF = low-frequency words; 1L-PW = one-letter pseudowords; PW = pseudowords.

Table 2
Pairwise comparisons between different stimulus categories for chronometric measures (estimated marginal means; Bonferroni correction for multiple comparisons).

	RT				PMT				MT			
Comparison	Est.	SE	z	p	Est.	SE	z	p	Est.	SE	z	p
HF − LF	−111	6.88	−16.18	<.001	−110	6.80	−16.19	<.001	−0.54	0.77	−0.71	1
HF − 1L-PW	−151	7.86	−19.20	<.001	−146	7.82	−18.68	<.001	−4.78	0.87	−5.47	<.001
HF − PW	−131	8.00	−16.48	<.001	−126	7.91	−15.93	<.001	−5.78	0.94	−6.13	<.001
LF − 1L-PW	−40	8.77	−4.53	<.001	−36	8.74	−4.13	<.001	−4.24	1.02	−4.16	<.001
LF − PW	−21	8.89	−2.31	.12	−16	8.82	−1.81	.42	−5.23	1.08	−4.86	<.001
1L-PW − PW	19	9.67	1.98	.29	20	9.63	2.09	.22	−0.99	1.16	−0.86	1

Note. RT = reaction time; PMT = premotor time; MT = motor time; Est. = estimate; SE = standard error; HF = high-frequency words; LF = low-frequency words; 1L-PW = one-letter pseudowords; PW = pseudowords.

increased rate of fast incorrect activations seems discernible even for high-frequency words, albeit to a much lesser degree compared to non-lexical stimulus categories. The distribution of incorrect activations for low-frequency words seems instead more homogeneous across different quantiles. When jointly considering conditional accuracy/incorrect activations functions, it seems that whereas fast incorrect activations for high-frequency are corrected before issuing the final response, corrective mechanisms can only partially compensate for the increased rate of impulsive incorrect activations yielded by one-letter pseudowords and pseudowords.

This interpretation is confirmed by the analysis of partial errors and correction likelihood. The significant effect of Stimulus Category for partial errors, $\chi^2(3) = 103.37$, $p < .001$, reveals that partial incorrect responses occur mainly for low-frequency words and one-letter pseudowords, whereas their occurrence across high-frequency words and pseudowords is comparable (Table 3). In terms of correction likelihood (effect of Stimulus Category, $\chi^2(3) = 168.56$, $p < .001$), high-frequency words however reveal a significant advantage over all the other categories of stimuli. Interestingly, corrections are also more likely to occur for one-letter pseudowords and pseudowords, compared to low-frequency words, likely due to the possibility to partially overcome fast errors stemming from lexical capture. Results are summarized in Fig. 2, whereas pairwise comparisons are listed in Table 3.

Response confidence

Fig. 3 provides a qualitative description of the distribution of confidence ratings. Specifically, it reports the total number of responses rated with each possible confidence score, separately for different stimulus categories (columns) and for correct and error responses (first and second row, respectively).

Most of the correct responses were identified with maximal levels of confidence. The distribution of confidence ratings for errors appears rather different. Whereas many of them were fully identified (confidence rating = 0), there is more variability compared to correct responses, signaling a greater degree of uncertainty in error awareness. Importantly, a substantial number of errors for the more ambiguous stimuli (low-frequency words, one-letter pseudowords and pseudowords) were misidentified as correct responses (confidence = 1). These are errors that entirely escape awareness and are actually considered as correct responses.

Accurate responses

Descriptively, participants were mostly aware of correct responses (Fig. 3). These were identified with maximal confidence (rating = 1) in more than 69 % of the trials. Guessing, i.e., trials in which confidence ratings fell between .45 and .55, were few (0.66 %), and the same was true for correct responses identified as “surely error” (0.36 %). Even when using a looser criterion (ratings < .50) to identify correct trials misidentified as errors, the frequency of occurrence remained negligible (1.39 %).

Considering all the correct responses, the 2 ways interactions Stimulus Category by PMT Quantile ($\chi^2(3) = 28.05$, $p < .001$), Stimulus Category by MT Quantile ($\chi^2(3) = 14.56$, $p = .002$), and Stimulus Category by Partial Error ($\chi^2(3) = 9.54$, $p = .02$) were significant. The inclusion of second order polynomials in fitting the PMT Quantile ($\chi^2(4) = 51.89$, $p < .001$) and the MT Quantile ($\chi^2(4) = 11.66$, $p = .02$) variables improved goodness of fit. With the inclusion of random slopes, however, the interaction between Stimulus Category and Partial Error turned to non-significant ($\chi^2(3) = 2.35$, $p = .50$), whereas the simple effect of Partial Error remained significant ($\chi^2(1) = 16.88$, $p < .001$).

As visible in Fig. 4, response confidence for correct responses decreased with increasing PMTs, and although the pattern seems reliable for all the different types of stimuli (at least in its linear component, see Table 4), it seemed emphasized for non-lexical items (one-letter

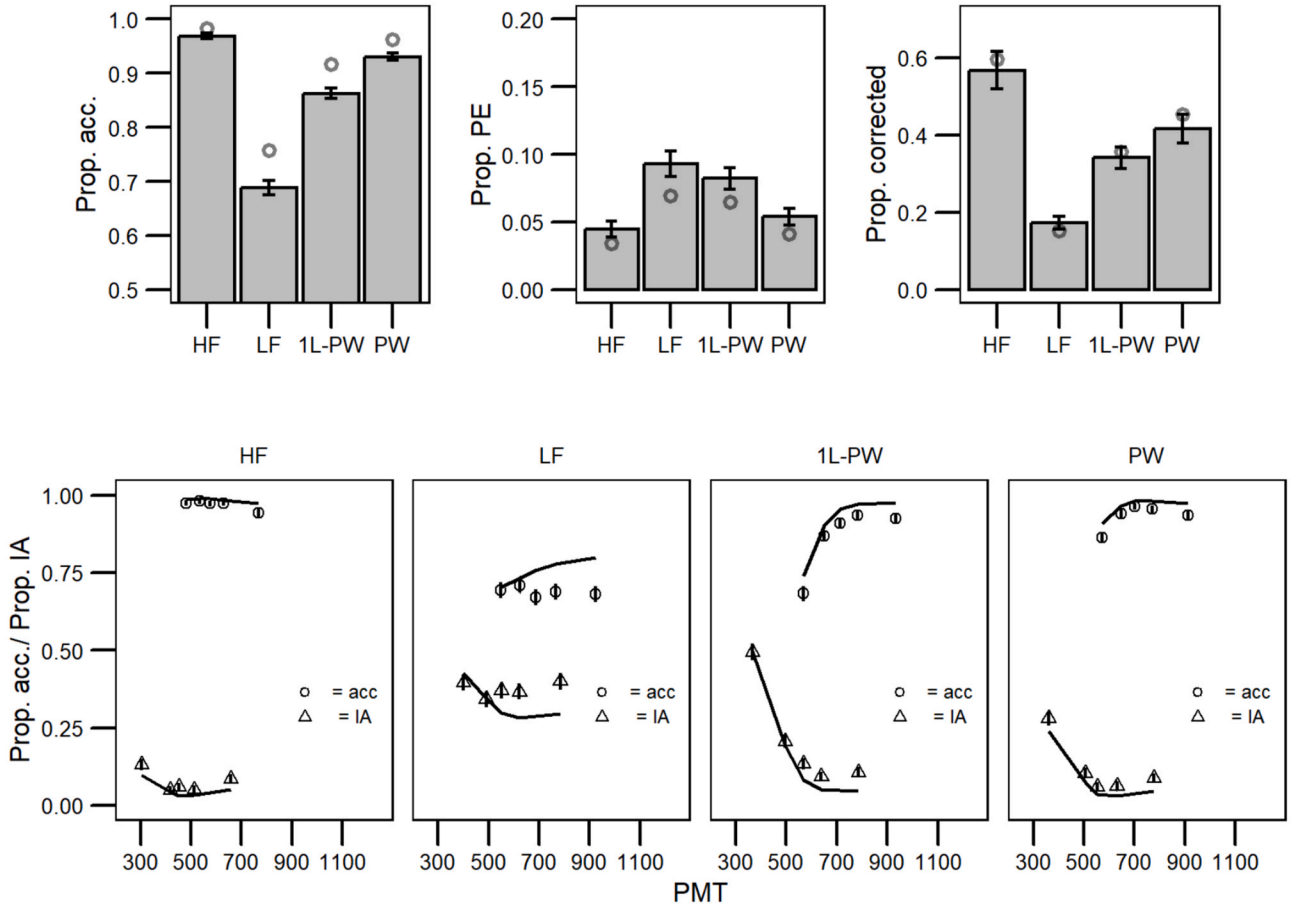

Fig. 2. Results for accuracy measures. First row: Results for accuracy (Prop. acc. = proportion of accurate responses), partial errors (Prop. PE = proportion of partial errors), and correction likelihood (Prop. corrected). Empty circles represent estimated marginal means. Second row: Conditional accuracy functions (Prop. acc; circles) and conditional inaccurate activations functions (Prop. IA = proportion of incorrect activations; triangles) for each stimulus category. Error bars display 95 % confidence intervals, adjusted for within-participants variables following Morey (2008). Lines represent estimated marginal means. HF = high-frequency words; LF = low-frequency words; 1L-PW = one-letter pseudowords; PW = pseudowords.

Table 3
Pairwise comparisons between different stimulus categories for accuracy measures (estimated marginal means; Bonferroni correction for multiple comparisons).

	Accuracy				Partial Errors				Correction likelihood			
Comparison	Est.	SE	z	p	Est.	SE	z	p	Est.	SE	z	p
HF − LF	.22	.03	6.83	<.001	−.03	.01	−4.56	<.001	.44	.04	11.93	<.001
HF − 1L-PW	.07	.01	6.19	<.001	−.03	.01	−5.25	<.001	.24	.04	5.74	<.001
HF − PW	.02	.005	3.70	.001	−.01	.004	−1.80	.43	.14	.04	3.25	.007
LF − 1L-PW	−.16	.03	−5.00	<.001	.004	.008	0.62	1	−.20	.03	−6.15	<.001
LF − PW	−.20	.03	−6.34	<.001	.03	.01	3.65	.002	−.30	.04	−8.20	<.001
1L-PW − PW	−.05	.01	−4.37	<.001	.02	.01	4.04	<.001	−.10	.04	−2.36	.11

Note. Est. = estimate; SE = standard error; HF = high-frequency words; LF = low-frequency words; 1L-PW = one-letter pseudowords; PW = pseudowords.

pseudowords and pseudowords; see Fig. 4 and Table 4). For low-frequency words, the non-linear component seems emphasized, suggesting a drop in confidence for slowest responses (Fig. 4). An inverse relationship was observed between confidence ratings and MTs; however, tests of the simple slopes confirmed this effect was significantly different from zero only for low-frequency words (Est. = -0.005, SE = .001, z = -4.56, $p < .001$; all other $ps > .12$; see Fig. 4; Table 4). Finally, confidence ratings were lower when trials included partial errors, compared to pure-correct responses (Table 4, main effect of partial error, P. Err.).

Another insight we can gather from this analysis concerns potential links between the differences across stimulus types in response execution and in response confidence. Possibly, the slower MTs for non-lexical items are due to residual uncertainty for pseudo-items with no representation in memory. This hypothesis is not supported by the data. Pairwise comparisons between stimulus categories in terms of confidence ratings (Table 5) reveal higher ratings for high-frequency words, compared to all other categories. Low-frequency words yielded ratings comparable to those observed for both one-letter pseudowords and pseudowords. No significant difference in terms of confidence surfaced between one-letter pseudowords and pseudowords (Fig. 4). Paired with the results from chronometric measures, we thus observe a) significant differences in terms of MTs, without corresponding effects on confidence ratings (e.g., low-frequency vs. one-letter pseudowords and pseudowords), as well as b) comparable MTs despite differences in terms of response confidence (e.g., high- vs low-frequency words). Confidence also does not seem to fully parallel the pattern of differences highlighted at the premotor level, albeit dissociations are less clear. In particular,

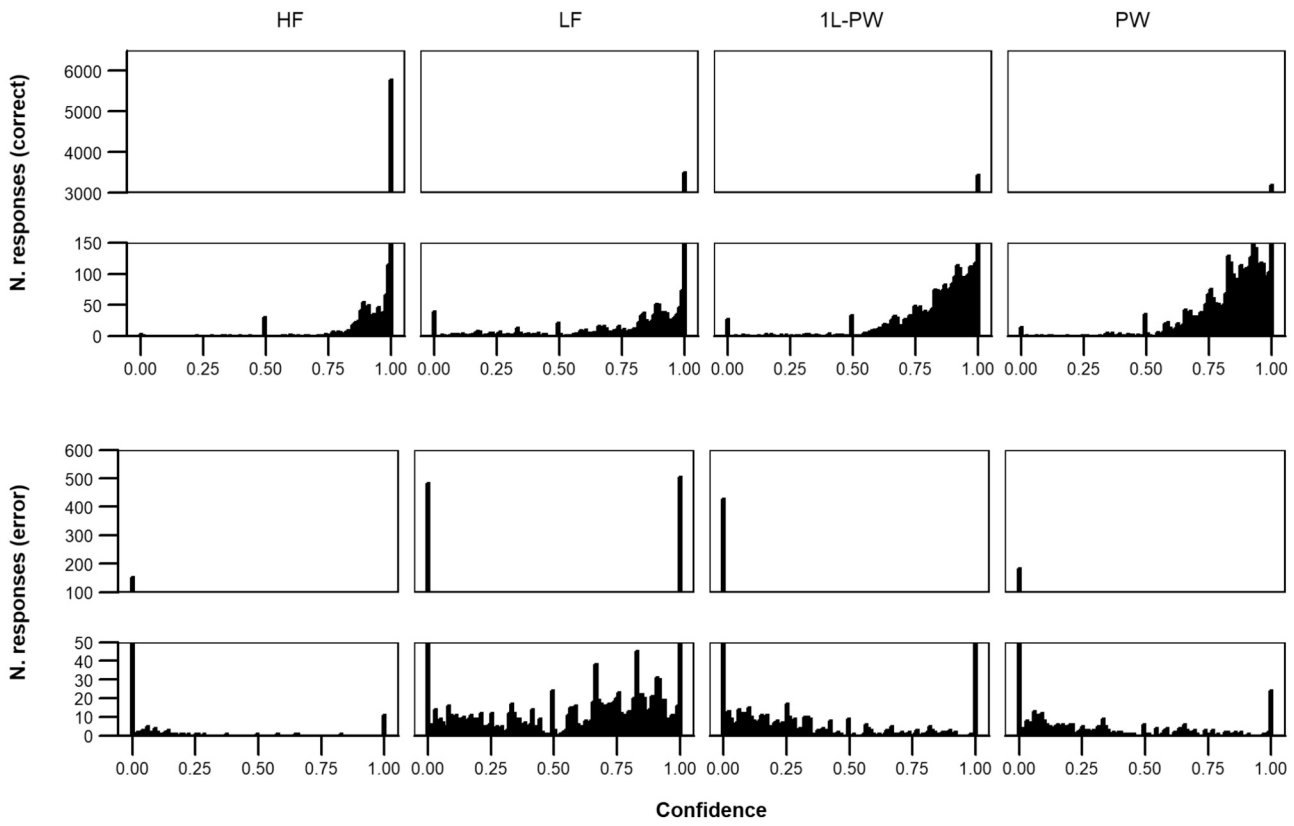

Fig. 3. Distribution of confidence ratings for correct and error responses as a function of stimulus types. HF = high-frequency words; LF = low-frequency words; 1L-PW = one-letter pseudowords; PW = pseudowords.

whereas PMT is significantly shorter for low-frequency words compared to one-letter pseudowords, the two types of stimuli yield comparable confidence ratings.

Error responses

Qualitatively, 34.15 % of the errors were identified as such with maximal confidence. A substantial fraction of error trials revealed various degrees of uncertainty (48.96 %). Interestingly, 17.15 % of error trials were misidentified as correct responses (Fig. 3, second row). Fig. 5 provides a more detailed descriptive summary of confidence ratings for errors.

Concerning errors identified with maximal confidence (i.e., rating = 0), their likelihood was reliably influenced only by the simple effects of Stimulus Category, $\chi^2(3) = 266.10$, $p < .001$, and PMT Quantile, $\chi^2(1) = 63.88$, $p < .001$. Errors were more likely to be identified with maximal confidence in case of high-frequency words compared to all other classes of stimuli. Also, errors for low-frequency words were less likely to be identified with maximal confidence, compared to all other stimuli (Table 6). For all stimulus categories, the likelihood of detecting errors with maximal confidence decreased with increasing PMTs (*Est.* = -0.050, *SE* = 0.006, $z = -8.21$, $p < .001$; Fig. 6).

Confidence for uncertain errors (i.e., ratings between.01 and.99) was again modulated by simple effects of Stimulus Category, $\chi^2(3) = 219.87$, $p < .001$, and PMT Quantile, $\chi^2(1) = 28.24$, $p < .001$. Confidence ratings were significantly higher for low-frequency words compared to all other stimulus categories (Table 6), signaling an inherent uncertainty about errors involving these stimuli. Additionally, confidence ratings increased as a function of PMTs (*Est* = 0.025, *SE* = 0.004, $t = 5.56$, $p < .001$), signaling an enhanced uncertainty concerning the status of incorrect responses with increasing latencies (Fig. 6).

The likelihood of misidentification of an error (incorrect response identified as correct with maximal confidence, rating = 1) was instead modulated by a significant Stimulus Category by MT Quantile interaction, $\chi^2(3) = 11.89$, $p < .001$, and the model was significantly improved when including second order polynomials to fit the MT Quantile variable, $\chi^2(4) = 10.51$, $p = .03$. Given the modest fit offered by the model (Fig. 6, last row), results need to be taken with caution. Only the slope estimated for one-letter pseudowords was significantly different from 0 (*Est.* = -0.026, *SE* = .009, $z = -2.86$, $p = .004$; all other *ps* > .11), and none of the pairwise comparisons between the slopes of the different stimulus categories reached significance (all *ps* > .05). This is possibly due to the low absolute number of observations falling in this category (626). Beyond the interaction with MT, the general phenomenon of error responses considered as correct remains noticeable, at least in relative terms (17.15 % of the total errors) and magnified in case of low-frequency words (Fig. 3).

General discussion

By integrating behavioral measures with single trial EMG recordings and confidence ratings, the present study aimed to assess intermediate decisional and response states in lexical decision and, relatedly, the degree of overlap between objective and subjective decisional outcomes (e.g., Fleming, 2024).

Alignment between correct decisions and metacognitive judgments

For correct responses, the results mitigated our concerns about the artificial constraints imposed by the two-choice configuration on decisional states. Confidence ratings were consistently high (>.90) across all categories of stimuli (Fig. 4, third row), and often associated with maximal levels of confidence. Even uncertain responses clustered within the upper end of the confidence scale. Guessing and misidentifications

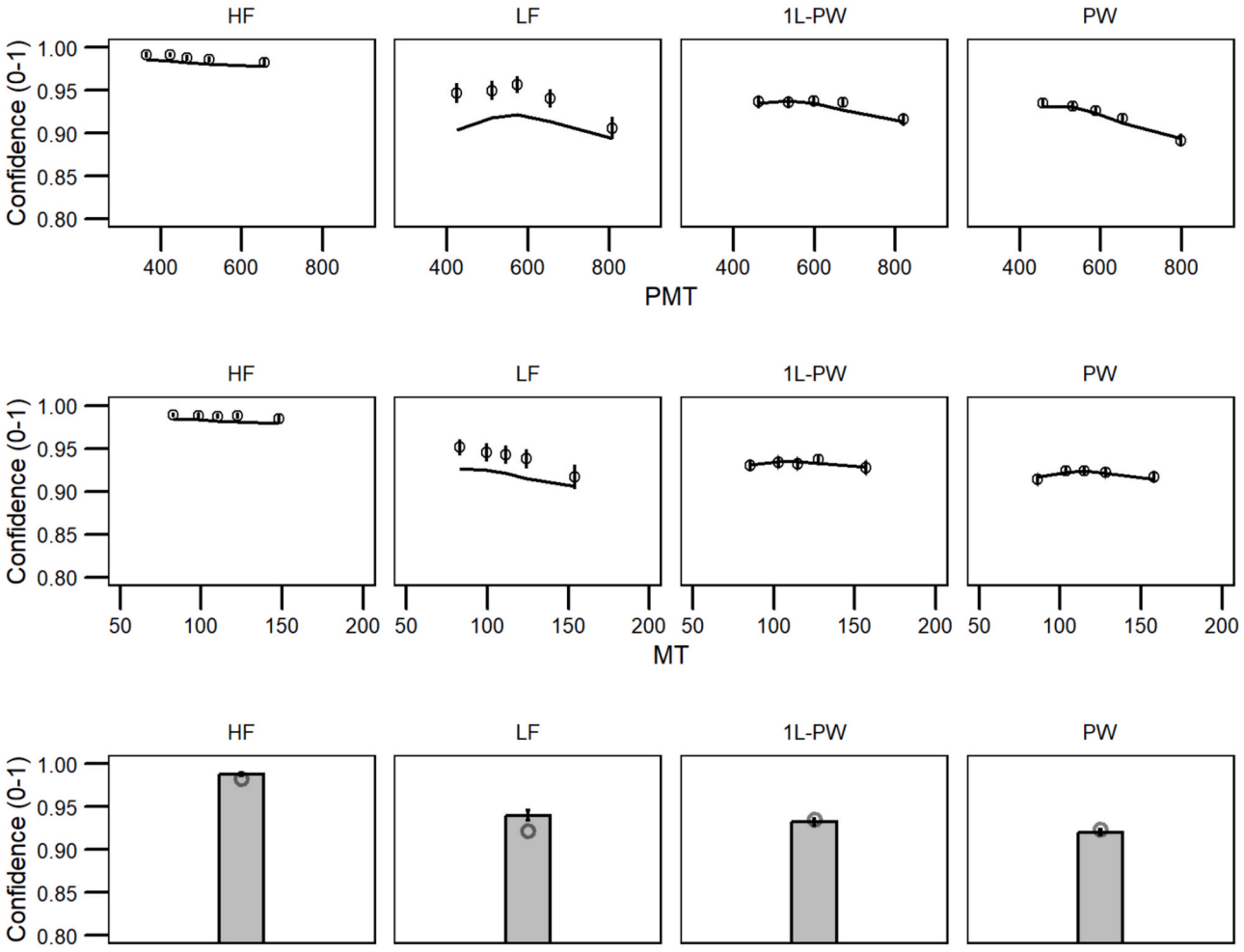

Fig. 4. Results of analyses of response confidence for correct responses. First and second row: points represent empirical scores and lines the estimated effect. Third row, empty circles represent estimated marginal means. Across figures, error bars display 95 % confidence intervals, adjusted for within-participants variables following Morey (2008). HF = high-frequency words; LF = low-frequency words; 1L-PW = one-letter pseudowords; PW = pseudowords.

Table 4
Parameters of the fixed effects for the linear mixed-effects (LME) model on confidence ratings for correct responses.

Fixed effect	Estimate	SE	t	p
Intercept	0.99	0.00	178.86	<.001
Stim. Type: LF	−0.07	0.01	−8.93	<.001
Stim. Type: 1L-PW	−0.05	0.01	−5.05	<.001
Stim. Type: PW	−0.07	0.01	−5.67	<.001
PMT (lin.)	−0.42	0.21	−2.03	0.043
PMT (quad.)	−0.02	0.20	−0.11	0.909
MT (lin.)	−0.31	0.20	−1.55	0.121
MT (quad.)	−0.02	0.20	−0.12	0.903
P. Err.	−0.01	0.00	−4.11	<.001
Stim. Type: LF x PMT (lin.)	−0.11	0.33	−0.34	0.736
Stim. Type: 1L-PW x PMT (lin.)	−0.73	0.31	−2.32	0.021
Stim. Type: PW x PMT (lin.)	−1.54	0.30	−5.17	<.001
Stim. Type: LF x PMT (quad.)	−1.41	0.31	−4.60	<.001
Stim. Type: 1L-PW x PMT (quad.)	−0.65	0.29	−2.26	0.024
Stim. Type: PW x PMT (quad.)	−0.67	0.28	−2.39	0.017
Stim. Type: LF x MT (lin.)	−0.80	0.31	−2.57	0.010
Stim. Type: 1L-PW x MT (lin.)	0.16	0.29	0.54	0.588
Stim. Type: PW x MT (lin.)	0.20	0.28	0.70	0.483
Stim. Type: LF x MT (quad.)	−0.28	0.31	−0.91	0.363
Stim. Type: 1L-PW x MT (quad.)	−0.32	0.29	−1.10	0.269
Stim. Type: PW x MT (quad.)	−0.50	0.28	−1.77	0.076

Note. Est. = estimate; SE = standard error; HF = high-frequency words; LF = low-frequency words; 1L-PW = one-letter nonwords; PW = pseudowords; lin. = linear; quad. = quadratic.

Table 5
Pairwise comparisons of confidence ratings for correct responses across stimulus categories (estimated marginal means; Bonferroni correction for multiple comparisons).

Comparison	Confidence			
	Est.	SE	z	p
HF − LF	.06	0.01	8.80	<.001
HF − 1L-PW	.05	0.01	5.07	<.001
HF − PW	.06	0.01	5.71	<.001
LF − 1L-PW	−.01	0.01	−1.47	1
LF − PW	−.002	0.01	−0.36	1
1L-PW − PW	.01	0.01	0.88	1

Note. Est. = estimate; SE = standard error; HF = high-frequency words; LF = low-frequency words; 1L-PW = one-letter pseudowords; PW = pseudowords.

(correct responses rated as errors) were rare, even for more ambiguous stimuli such as low-frequency words and nonwords (Fig. 3, first row). These findings suggest a strong alignment between subjective confidence and objective correctness in lexical decisions.

Prior studies have examined inconsistencies in lexical decisions across repeated presentations, highlighting how internal noise can undermine choice stability, especially for very low-frequency words (Diependaele et al., 2012). While these past findings suggest that guessing plays a significant role when responding to low-frequency words, our data show that, when correctly identified, responses are

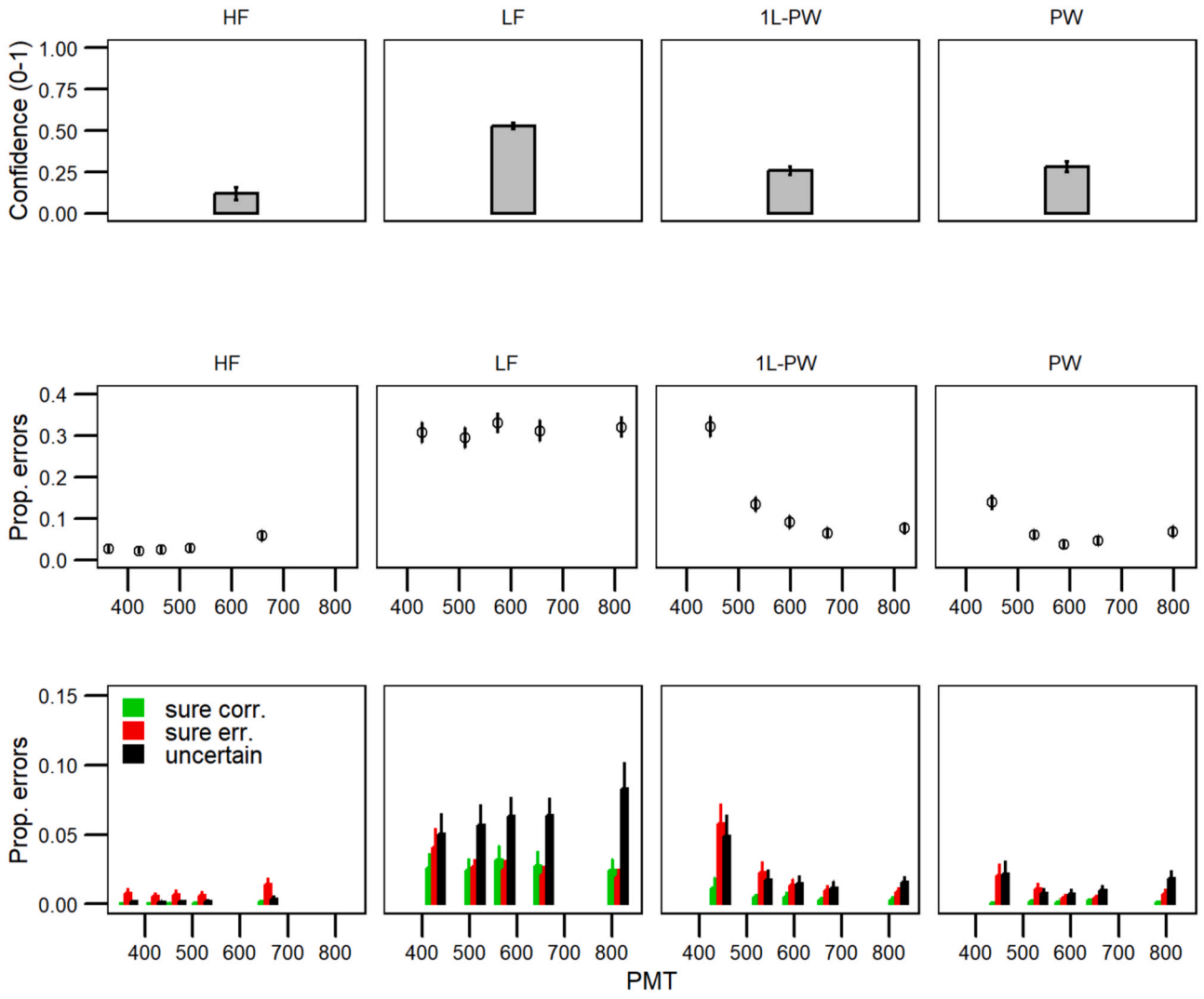

Fig. 5. Descriptive summary of response confidence for error responses. First row: overall average confidence as a function of stimulus type. Second row: Proportion of errors as a function of stimulus type and PMT quantile. Third row: data of the second row are broken down for different types of error (sure corr. = errors rated as correct responses with maximal confidence levels; sure err. = errors identified with maximal confidence; uncertain = errors rated between 0.01 and 0.99).

Table 6
Pairwise comparisons between different stimulus categories for a) the probability of errors with full awareness, b) confidence ratings for uncertain errors (Bonferroni correction for multiple comparisons).

	Prob. errors with full awareness				Confidence for uncertain errors			
Comparison	Est.	SE	z	p	Est.	SE	t	p
HF − LF	0.60	0.04	14.99	<.001	−0.37	0.04	−8.40	<.001
HF − 1L-PW	0.41	0.04	10.12	<.001	−0.13	0.04	−2.95	.020
HF − PW	0.49	0.04	11.11	<.001	−0.14	0.04	−3.01	.017
LF − 1L-PW	−0.19	0.04	−5.19	<.001	0.24	0.03	9.02	<.001
LF − PW	−0.11	0.04	−2.79	.032	0.23	0.03	7.59	<.001
1L-PW − PW	0.08	0.035	2.19	.172	−0.01	0.029	−0.34	1

Note. Est. = estimate; SE = standard error; HF = high-frequency words; LF = low-frequency words; 1L–PW = one-letter nonwords; PW = pseudowords.

typically associated with high confidence, even for more ambiguous stimuli. Correct responses, which are often the focus of lexical decision experiments, seem to reflect clear-cut categorizations accompanied by high confidence. The binary response format, therefore, does not appear to substantially distort the underlying decisional state, given the relatively categorical nature of word versus nonword judgments. Additional exploratory analyses (Supplementary Materials) confirmed this conclusion even when controlling for potential carry-over phenomena stemming from the previous trial, such as potential lowering of confidence levels after errors, reduced confidence for unfamiliar/ambiguous items following familiar ones, or confidence inflation after unfamiliar items.

Correct responses further revealed systematic links between measures of performance and the metacognitive assessment of decision outcomes. An important role in shaping confidence was played by chronometric measures. Premotor time (PMT), which encompasses much of the response latency and underlying decisional processing, inversely correlated with confidence, particularly for ambiguous stimuli like low-frequency words and nonwords. This indicates that more

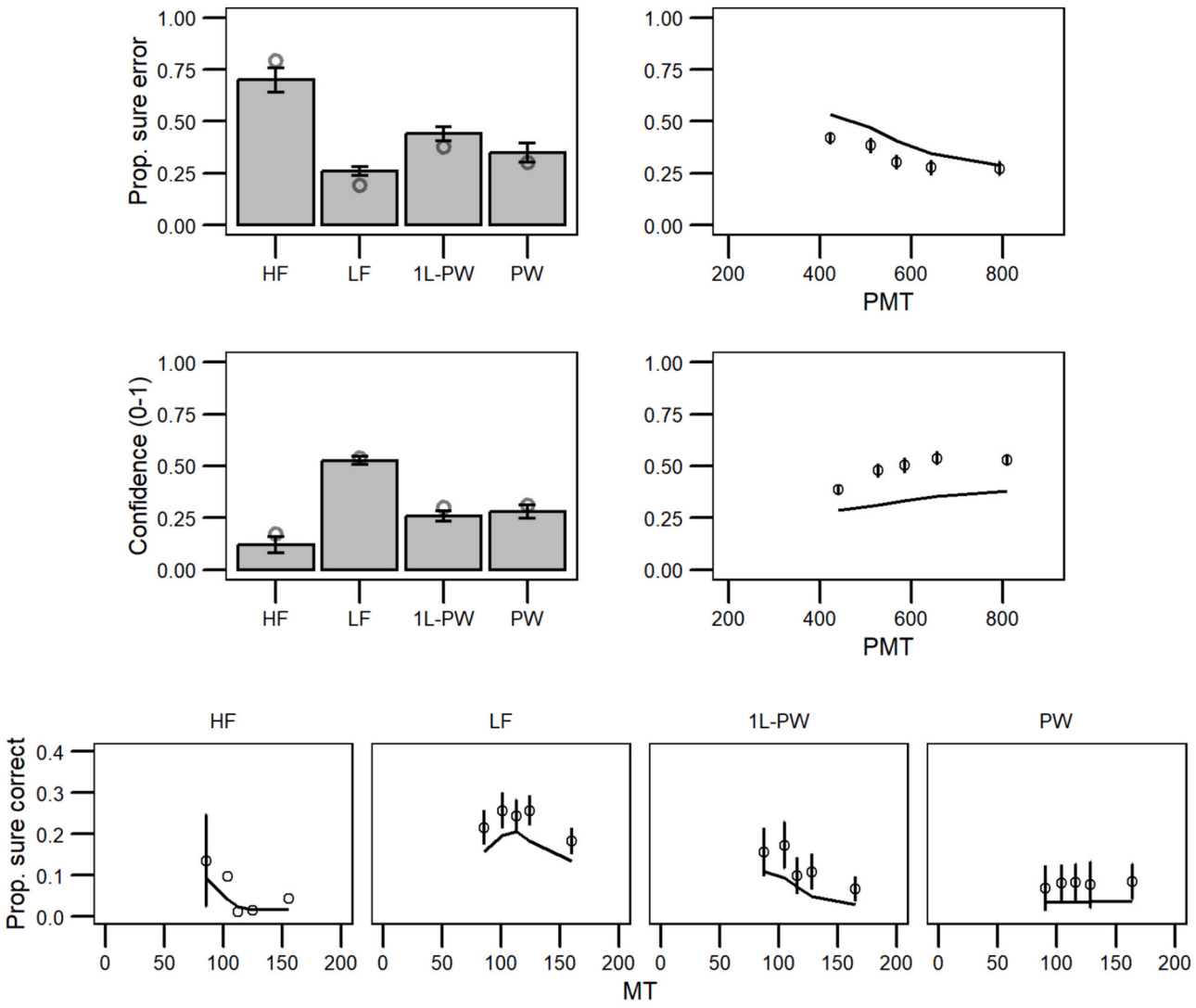

Fig. 6. Results of analyses of response confidence for error responses. First row: Effects of Stimulus Type and PMT Quantile on the likelihood of identifying errors with maximal confidence (= 0). Second row: Effects of Stimulus Type and PMT Quantile on confidence ratings for uncertain errors (confidence = .01-.99). Third row: Stimulus Type by MT Quantile for the likelihood of rating errors as correct responses with maximal confidence (= 1). Within bar plots, empty circles represent estimated marginal means. Within scatterplots, points represent empirical scores and lines the estimated effect. Error bars display 95 % confidence intervals, adjusted for within-participants variables following Morey (2008). HF = high-frequency words; LF = low-frequency words; 1L-PW = one-letter pseudowords; PW = pseudowords.

effortful, slower recognition correlates with reduced confidence, a pattern consistent with findings from perceptual decision-making (Kiani et al., 2014).

Motor fluency and confidence

In terms of motor response execution, chronometric measures of the lexical decision performance align with prior findings (Scaltritti, Giacomoni et al., 2023; Scaltritti, Greatti et al., 2023; Scaltritti et al., 2025; Kamari Songhorabadi et al., 2025). Notably, motor execution was selectively modulated by lexicality but not by lexical frequency (Fiora et al., 2026; Scaltritti, Giacomoni et al., 2023). This dissociation suggests that MT does not merely reflect a spillover of pre-motor decision processes, but rather a categorical distinction between responses driven by stored representations (words) versus those for stimuli without long-term memory entries (nonwords).[2] Numerically, the lexicality effect on MTs was small (~5 ms). While non-trivial relative to the short duration of the motor interval (~115 ms), this represents a fraction of the total effect observed in global RTs (~85 ms), indicating that the major part of processing differences is resolved within the premotor interval.

Motor response duration influenced confidence only in the case of low-frequency words, where longer MTs were associated with lower confidence. This limited involvement of motor-response duration in

[2] A novel aspect concerns the comparison between low-frequency and pseudoword stimuli: despite comparable PMTs, MTs were longer for the non-lexical category. Stimuli taking comparable processing times at the premotor level may thus display a difference in terms of response execution, further pointing towards the presence of multiple decisional components across the two intervals (e.g., Kamari Songhorabadi et al., 2025; Scaltritti, Greatti et al., 2023).

shaping confidence suggests that metacognitive evaluations are flexible, potentially integrating different cues depending on the perceived difficulty of the prior decision. One such cue may be fluency, defined as the subjective ease or difficulty experienced during processing (Oppenheimer, 2008)—in this case, the fluency of motor execution. Whereas consistent with the general notion that the whole perception–action cycle is involved in confidence formation (e.g., Fleming & Daw, 2017), the fluency-based interpretation contrasts with the view that greater action monitoring involvement enhances response confidence (e.g., Sanchez et al., 2024). This view seemingly implies that longer MTs, which have been associated to the engagement of online response control (e.g., Allain et al., 2004), should be associated with increased confidence. However, the opposite pattern emerged. Further support for the fluency hypothesis comes from the effects of partial errors on confidence ratings. Although partial responses have been associated with increased confidence in perceptual decisions (Gajdos et al., 2019), our data showed that transient incorrect activations, despite triggering corrective control, reduced confidence (independently from Stimulus Type). The data thus seem more consistent with the idea that motor fluency—rather than the degree of engagement of online response control—predicts confidence in lexical decision, at least in those instances in which motor parameters are considered among the cues driving confidence formation.

The metacognitive profile of errors

Although overall accuracy was high, with errors occurring mainly for low-frequency words, these relatively rare incorrect responses exhibited a markedly different metacognitive profile, with a substantial reduction in the alignment between subjective and objective decisional outcomes. Importantly, qualitative differences across stimulus categories, in both temporal dynamics and metacognitive accessibility, suggested the underlying presence of distinct generative processes. Metacognitive awareness concerning errors was strongly shaped by stimulus category. Errors on high-frequency words were detected with high confidence or limited uncertainty, with negligible misidentifications (incorrect responses rated as correct). In contrast, for low-frequency words, only about one quarter of errors were recognized with maximal confidence. More uncertain ratings were distributed across the entire scale, and a substantial fraction of errors were fully misidentified as correct responses (Fig. 3, second row). This pattern likely reflects the convolution of different types of errors: slips (i.e., errors with full awareness), not-known items treated as nonwords (i.e., errors misidentified as correct responses), and unresolved decisional states featuring different levels of uncertainty. Concerning the latter, it is worth noting that ratings in the guessing range were not particularly frequent, and uncertain responses were spread across the entire range of the confidence scale, signaling a wide spectrum of non-fully resolved decisional states (Fig. 3, second row, second column).

While additional analyses failed to reveal specific features of low-frequency words that could be associated with misidentified errors (Supplementary Materials), comparisons between low-frequency and non-lexical items further clarify the different dynamics underlying incorrect responses. Nonword errors, on average, were associated with reduced uncertainty (Fig. 5, first row; Fig. 6, second row), a higher proportion of fully aware identifications (Fig. 6, first row), and fewer cases in which errors were misjudged as correct responses. In short, nonword errors appeared to be more accessible to awareness. One possibility is that errors on low-frequency words often arise from decisional uncertainty, reflected in the increased variability of confidence ratings, or from a lack of lexical knowledge, as suggested by the frequent cases in which erroneous responses were judged as correct with high confidence. In contrast, nonword errors may stem from competition with activated lexical entries. Time-course analyses of incorrect response activations and accuracy support this interpretation (Fig. 2, second row). For pseudowords, the rate of incorrect activations was much higher in faster latencies, pointing to lexical capture phenomena triggered by the co-activation of lexical entries (e.g., a pseudoword like *elethant* activating the word *elephant*; Fiora et al., 2026; Grisetto et al., 2025a; 2025b; Scaltritti et al., 2021, Scaltritti, Giacomoni 2023; Scaltritti et al., 2025). Online corrective mechanisms could only partially suppress these fast error-tendencies, as conditional accuracy functions showed that a sizeable proportion of fast incorrect activations turned to overt errors. By contrast, low-frequency words displayed a flatter profile in both incorrect activations and conditional accuracy.

From a chronometric perspective, error awareness was primarily driven by the premotor component of the response latency. There was in fact an inverse relationship between PMT and error awareness: the faster the incorrect response, the more confident participants were to detect the error. Although this pattern is consistent with findings from perceptual decision-making tasks (Kiani et al., 2014), the dynamics observed in lexical decision diverge in one important way. In perceptual tasks, the error rate tends to increase with longer latencies. In contrast, nonwords display the highest error rates within the fastest latency range. Albeit fast nonword errors (as discussed above) are likely to reflect lexical captures bypassing corrective mechanisms, they do not necessarily escape awareness, pointing to a dissociation between error correction and detection (Rabbitt, 2002). In the present dataset, one possibility is that participants became aware of (some) fast errors at a point where it was too late to interrupt execution. This seems consistent with the finding that the error-rate in lexical decision is greatly reduced when a delay is introduced between stimulus presentation and response (Romero-Ortells et al., 2024).

By contrast, the contribution of motor response duration to error trials was comparatively limited. Overall, MT did not significantly influence the likelihood of full error awareness or confidence ratings for uncertain errors. As prolonged MTs have been linked to online attempts by the executive control system to inhibit erroneous responses (Allain et al., 2004), the absence of a MT effect may suggest that inhibitory interventions are not consistently accessible to metacognitive evaluation. An involvement of MT was indeed observed only under highly specific conditions: error trials in which participants incorrectly judged their response as correct and the stimulus was a one-letter pseudoword. Model fit was modest (Fig. 6, last row), and results should therefore be interpreted with caution. Nevertheless, the data indicate that longer MTs may reduce such misidentifications. One possibility is that more extreme response-control interventions, while not sufficient to alter confidence judgments more broadly, may act as a cue against full misidentification of an error. This contribution, however, seems limited to highly ambiguous stimuli, such as pseudowords differing in just one letter from existing words.

Theoretical Implications

In summary, confidence ratings for correct responses clustered at high or maximal levels, reflecting clear discrimination between lexical and non-lexical items. Furthermore, motor fluency influenced confidence, albeit this influence was clearer just for more difficult decisions such as those concerning low-frequency words. Regarding errors, the alignment between subjective and objective outcomes was weaker, especially in ambiguous categories. Confidence ratings suggest distinct generative mechanisms: while errors on unfamiliar words may reflect unresolved decisional states or lack of lexical knowledge, nonword errors appear driven by lexical competition yielding impulsive responses that are only partially mitigated by online control. Notably, the enhanced awareness observed for these fast errors points to an at least partial dissociation between error detection and correction in lexical decisions.

Taken together, these findings underscore the interplay between lexical decision-making, motor-response execution, and metacognitive evaluation. These data do not necessarily imply that models of lexical decision should be modified to incorporate motor-execution or post-

response confidence, as these dynamics fall outside the strict scope of word recognition. However, the analysis of confidence ratings and motor output offers additional information concerning the decisional outcome beyond the task's traditional boundaries. The evidence accumulation framework currently central to the field provides interpretative suggestions. Indeed, parallel work in perceptual decision-making has extended evidence-accumulation models in two directions: to account for motor dynamics beyond response initiation, and to explain confidence judgments formation.

For instance, the gated cascaded diffusion model (e.g., Achard et al., 2025; Dendauw et al., 2024; Servant et al., 2021) allows evidence accumulation to continue beyond the initial decision boundary, thereby shaping motor activity. This model accommodates both behavioral and EMG data, providing a quantitative account that includes motor processes typically subsumed under the "non-decision time" label. In parallel, some models of metacognitive confidence suggest that evidence accumulation may continue even after the decision boundary is crossed, either via continued sensory sampling, or via an internal reassessment of decision accuracy (e.g., Desender et al., 2021; Pleskac & Busemeyer, 2010; Yeung & Summerfield, 2012; see also Fleming, 2024). While targeting different outcomes, motor execution versus confidence, both approaches converge on a shared principle: evidence-accumulation continues beyond the decision threshold. This convergence prompts a series of questions. Are post-threshold signals unitary or distinct for motor and confidence processes? If distinct, when and how are they integrated? Our findings may offer some preliminary insights.

On one hand, motor parameters cannot be fully reduced to confidence. For instance, low-frequency words and pseudowords differ in motor time (MT) yet elicit similar confidence ratings. Conversely, high- and low-frequency words differ in confidence despite comparable MTs. These dissociations suggest that motor execution is not just a readout of the same evidence used to form metacognitive judgments. Furthermore, MT appears to selectively capture qualitative distinctions between lexical and non-lexical items. This implies that, in memory-based decisions, the motor system may be sensitive to the categorical distinction between items with versus without a representation in memory, rather than to the graded quality of lexical evidence (Fiora et al., 2026; Scaltritti, Giacomoni et al., 2023).

On the other hand, there were influences of motor parameters on confidence, albeit limited to specific conditions. For correct responses, whereas partial errors were generally associated with reduced confidence, the influence of response duration was limited to low-frequency words. Motor and metacognitive signals, while distinguishable, are likely integrated at some point, particularly under specific circumstances (i.e., for the most difficult stimulus category in our experiment). In this view, confidence formation may rely on multiple cues, including motor fluency, as part of a wider metacognitive system.

When situating these EMG-based findings within the broader context of (lexical) decisional dynamics, a caveat regarding the comparisons with related behavioral paradigms is necessary. Specifically, merging insights derived from prolonged movement tasks (e.g., mouse-tracking; pulling a lever) with those from discrete button-press tasks is not trivial. The cognitive and motor constraints of generating prolonged movement trajectories fundamentally differ from those involved in rapid, discrete button presses. Prolonged response durations allow more time for corrective mechanisms to intervene (Ramdani et al., 2021; Spieser et al., 2017) and may generally enhance any chance of overlap between decisional and motor stages. Furthermore, while the EMG approach relies on a bimanual two-choice setup, other paradigms have relied on unimanual responses (e.g., mouse movement), complicating the comparisons. Finally, EMG provides a direct measure of muscular onset, whereas mouse-tracking (or lever movement) analyses rely on changes in movement kinematics as indices of decisional processes. Ultimately, the specific weight of these factors requires empirical investigation through direct cross-modal comparisons.

Importantly, our findings also point to domain-specific features that may limit the generalization of perceptual decision models to memory-based tasks (e.g., Dendauw et al., 2024; Ratcliff et al., 2016). For example, while perceptual and lexical decision tasks exhibit opposing latency profiles—with errors being typically slow in the former but fast in the latter (for nonwords)— both domains exhibit increased error awareness for faster errors, despite these subsets of fast errors likely stemming from different generative processes. Similarly, while partial EMG activations increase confidence in perceptual tasks (Gajdos et al., 2019), here partial errors were associated with reduced confidence.

Such discrepancies suggest that while evidence accumulation may be a domain-general principle, its implementation possibly adapts to task structure and informational constraints. Perceptual decisions typically involve mapping ambiguous but externally defined stimuli onto given states of the perceptual environment that can be univocally defined (e.g., left vs right). In contrast, lexical decisions rely on internally generated, memory-based signals and (uniquely) require rejecting items for which no representation exists (nonwords). The challenge of making "no" decisions in this context has drawn theoretical attention (Dufau et al., 2012; Norris, 2009) and underscores the conceptual complexity of the task. Indeed, lexical decision has often raised concerns related to the addition of a decision layer that may obscure the underlying processes of word recognition and lexical access (Balota & Chumbley, 1984; Ratcliff et al., 2004). However, this complication may offer complementary advantages. Most notably, a window into more complex decisional scenarios that rely on long-term memory and require judgments in the absence of evidence, an underexplored but ecologically relevant cognitive challenge.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Michele Scaltritti: Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Saman Kamari Songhorabadi:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Data curation. **Simone Sulpizio:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by the European Union – Next Generation EU – PRIN 2022 PNRR (DD 1409—14/09/22) – PNRR – M4 – C2 – INV1.1 – PRIN – Functional characterization of decisional components in motor responses for young and older adults – grant number [2022-NAZ-0671/PER] – CUP [E53D23019540001].

We are grateful to Costanza Bigolin and Margherita Cardellini for their help in conducting the experiment.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.

org/10.1016/j.jml.2025.104728.

Data availability

Data, materials and scripts are available at: https://osf.io/qhdgy.

References

Achard, J. A., Gajdos Preuss, T., & Servant, M. (2025). Extending continuous flow models of immediate decision reports to delayed decision reports. *Journal of Experimental Psychology: General, 154*, 1583–1610.

Allain, S., Carbonnell, L., Burle, B., Hasbroucq, T., & Vidal, F. (2004). On-line executive control: An electromyographic study. *Psychophysiology, 41*(1), 113–116.

Balota, D. A., & Abrams, R. A. (1995). Mental chronometry: Beyond onset latencies in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1289.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 340–357.

Balsdon, T., Wyart, V., & Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature Communications, 11*, 1753.

Barca, L., & Pezzulo, G. (2012). Unfolding visual lexical decision in time. *PloS one, 7*, Article e35932.

Barca, L., & Pezzulo, G. (2015). Tracking second thoughts: Continuous and discrete revision processes during visual lexical decision. *PLoS One, 10*, Article e0116193.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Botwinick, J., & Thompson, L. W. (1966). Premotor and motor components of reaction time. *Journal of Experimental Psychology, 71*, 9–15.

Burle, B., Possamaï, C. A., Vidal, F., Bonnet, M., & Hasbroucq, T. (2002). Executive control in the Simon effect: An electromyographic and distributional analysis. *Psychological Research, 66*, 324–336.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*(1), 16.

Charles, L., & Yeung, N. (2019). Dynamic sources of evidence supporting confidence judgments and error detection. *Journal of Experimental Psychology: Human Perception and Performance, 45*, 39–52.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204.

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review, 117*, 713–758.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9–21.

Dendauw, E., Evans, N. J., Logan, G. D., Haffen, E., Bennabi, D., Gajdos, T., & Servant, M. (2024). The gated cascade diffusion model: An integrated theory of decision making, motor preparation, and motor execution. *Psychological Review, 131*(4), 825–857.

Desender, K., Donner, T. H., & Verguts, T. (2021). Dynamic expressions of confidence within an evidence accumulation framework. *Cognition, 207*, Article 104522.

Diependaele, K., Brysbaert, M., & Neri, P. (2012). How noisy is lexical decision? *Frontiers in Psychology, 3*, 348.

Dufau, S., Grainger, J., & Ziegler, J. C. (2012). How to say "no" to a nonword: A leaky competing accumulator model of lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1117–1128.

Eriksen, C. W., Coles, M. G., Morris, L. R., & O'hara, W. P. (1985). An electromyographic examination of response competition. *Bulletin of the Psychonomic Society, 23*, 165–168.

Fiora, E., Scaltritti, M., & Sulpizio, S. (2026). Recognizing the unknown: Motor-response execution reflects the availability of positive evidence during recognition. *Acta Psychologica, 262*, Article 106068.

Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology, 75*, 241–268.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124*, 91–127.

Fluchère, F., Burle, B., Vidal, F., van den Wildenberg, W., Witjas, T., Eusebio, A., & Hasbroucq, T. (2018). Subthalamic nucleus stimulation, dopaminergic treatment and impulsivity in Parkinson's disease. *Neuropsychologia, 117*, 167–177.

Gajdos, T., Fleming, S. M., Saez Garcia, M., Weindel, G., & Davranche, K. (2019). Revealing subthreshold motor contributions to perceptual confidence. *Neuroscience of Consciousness, niz001*.

García-Pérez, M. A., & Alcalá-Quintana, R. (2023). Accuracy and precision of responses to visual analog scales: Inter-and intra-individual variability. *Behavior Research Methods, 55*, 4369–4381.

Goslin, J., Galluzzi, C., & Romani, C. (2014). Phonitalia: A phonological lexicon for Italian. *Behavior Research Methods, 46*(3), 872–886.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103*, 518–565.

Grisetto, F., Roger, C., & Mahe, G. (2025a). New insights into visual word recognition: Analyzing error distribution in typical readers. *Journal of Cognition, 8*, 29.

Grisetto, F., Roger, C., & Mahé, G. (2025b). Error dynamics as a marker of reading efficiency development: Insights from lexical decision performance in young readers. *Journal of Experimental Child Psychology, 260*, Article 106347.

Hasbroucq, T., Possamaï, C. A., Bonnet, M., & Vidal, F. (1999). Effect of the irrelevant location of the response signal on choice reaction time: An electromyographic study in humans. *Psychophysiology, 36*, 522–526.

Kamari Songhorabadi, S., Sulpizio, S., & Scaltritti, M. (2025). Dissociating premotor and motor components of response times: Evidence of independent decisional effects during motor-response execution. *Psychonomic Bulletin & Review, 32, 1890–1900, 32*, 1890–1900. https://doi.org/10.3758/s13423-025-02663-z

Keuleers, E. (2013). *vwr: Useful functions for visual word recognition research* (R Package Version 0.3.0) [Computer software]. https://CRAN.R-project.org/package=vwr.

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron, 84*, 1329–1342.

Liu, J., & Liu, Q. (2016). Use of the integrated profile for voluntary muscle activity detection using EMG signals with spurious background spikes: A study with incomplete spinal cord injury. *Biomedical Signal Processing and Control, 24*, 19–24.

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience, 8*, 213.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau. *Tutorials in Quantitative Methods for Psychology, 4*(2), 61–64.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*, 327–357.

Norris, D. (2009). Putting it all together: A unified account of word recognition and reaction-time distributions. *Psychological Review, 116*, 207–219.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*(1), 97–113.

Ollman, R. T. (1977). Choice reaction time and the problem of distinguishing task effects from strategy effects. In S. Domic (Ed.), *Attention & Performance VI* (pp. 99–113). Hillsdale, NJ: Erlbaum.

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*, 237–241.

Pavailler, N., Gevers, W., & Burle, B. (2025). Temporal metacognition: Direct readout or mental construct? The case of introspective reaction time. *Journal of Experimental Psychology: General, 154*, 1122–1148.

Perea, M., Rosa, E., & Gómez, C. (2005). The frequency effect for pseudowords in the lexical decision task. *Perception & Psychophysics, 67*, 301–314.

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

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/.

Rabbitt, P. M. A. (2002). Consciousness is slower than you think. *Quarterly Journal of Experimental Psychology A, 55*, 1081–1092.

Rahnev, D. (2021). Visual metacognition: Measures, models, and neural correlates. *American Psychologist, 76*, 1445–1453.

Rahnev, D., Desender, K., Lee, A. L., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., & Zylberberg, A. (2020). The confidence database. *Nature Human Behaviour, 4*, 317–325.

Ramdani, C., Carbonnell, L., Vidal, F., Béranger, C., Dagher, A., & Hasbroucq, T. (2015). Dopamine precursors depletion impairs impulse control in healthy volunteers. *Psychopharmacology, 232*, 477–487.

Ramdani, C., Sagui, E., Schmid, B., Castagna, O., Davranche, K., Vidal, F., & Hasbroucq, T. (2021). Action monitoring fails when motor execution is too fast: No time for correction. *Journal of Systems and Integrative. Neuroscience, 7*.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20*, 260–281.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review, 111*, 159–182.

Romero-Ortells, I., Baciero, A., Marcet, A., Perea, M., & Gómez, P. (2024). A stringent test of visuospatial position uncertainty accounts of letter position coding. *Language, Cognition and Neuroscience, 39*, 1278–1290.

Sanchez, R., Courant, A., Desantis, A., & Gajdos, T. (2024). Making precise movements increases confidence in perceptual decisions. *Cognition, 249*, Article 105832.

Scaltritti, M., Giacomoni, F., Job, R., & Sulpizio, S. (2023). Redefining the decisional components of motor responses: Evidence from lexical and object decision tasks. *Journal of Experimental Psychology: Human Perception and Performance, 49*, 835–851.

Scaltritti, M., Greatti, E., & Sulpizio, S. (2023). Electrophysiological evidence of discontinuities in the propagation of lexical decision processes across the motor hierarchy. *Neuropsychologia, 108630*.

Scaltritti, M., Greatti, E., & Sulpizio, S. (2025). Decisional components of motor responses are not related to online response control: Evidence from lexical decision and speed-accuracy tradeoff manipulations. *Memory & Cognition, 53*, 911–925.

Scaltritti, M., Job, R., & Sulpizio, S. (2021). Selective suppression of taboo information in visual word recognition: Evidence for cognitive control on semantics. *Journal of Experimental Psychology: Human Perception and Performance, 47*, 934–945.

Servant, M., Logan, G. D., Gajdos, T., & Evans, N. J. (2021). An integrated theory of deciding and acting. *Journal of Experimental Psychology: General, 150*, 2435–2454.

Singman, H., Bolker, B., Westfal, J., Aust, F., & Ben-Shachar, M. S. (2021). *afex: Analysis of Factorial Experiments*. R package version 0.28-1. https://CRAN.R-project.org/package=afex.

Servant, M., White, C., Montagnini, A., & Burle, B. (2015). Using covert response activation to test latent assumptions of formal decision-making models in humans. *Journal of Neuroscience, 35*(28), 10371–10385.

Spieser, L., Servant, M., Hasbroucq, T., & Burle, B. (2017). Beyond decision! Motor contribution to speed–accuracy trade-off in decision-making. *Psychonomic Bulletin & Review, 24*, 950–956.
Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica, 50*, 179–197.
Wagenmakers, E. J., Steyvers, M., Raaijmakers, J. G., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology, 48*, 332–367.
Wagenmakers, E. J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language, 58*, 140–159.
Weindel, G., Anders, R., Alario, F., & Burle, B. (2021). Assessing model-based inferences in decision making with single-trial response time decomposition. *Journal of Experimental Psychology: General, 150*(8), 1528–1555.
Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer.
Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971–979.
Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*, 1310–2133.