Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

STROOP EFFECT WITH SIZE WORDS

Elberto A. Plazas

elbertoa.plazasp@konradlorenz.edu.co

Fundación Universitaria Konrad Lorenz

Juan Carlos Forigua-Vargas

Fundación Universitaria Konrad Lorenz

José A. Vélez Marulanda

Fundación Universitaria Konrad Lorenz

Research Article

Keywords: Stroop task, size naming, size words, task conflict, response conflict, vocal reaction time, Spanish

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11455722/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

STROOP EFFECT WITH SIZE WORDS

Elberto A. Plazas[1*]

Juan Carlos Forigua-Vargas[1]

José A. Vélez Marulanda[2]

[1] Faculty of Psychology, Fundación Universitaria Konrad Lorenz, Carrera 9 bis # 62-43, Bogotá, Colombia

[2] Faculty of Mathematics and Engineering, Fundación Universitaria Konrad Lorenz, Carrera 9 bis # 62-43, Bogotá, Colombia

*Corresponding author: Elberto A. Plazas, elbertoa.plazasp@konradlorenz.edu.co

Email addresses: EAP: elbertoa.plazasp@konradlorenz.edu.co; JCFV: juanc.foriguav@konradlorenz.edu.co; JAVM: josea.velezm@konradlorenz.edu.co

Abstract

Background: The Stroop effect has been studied mainly with color as the target dimension. Multistage accounts decompose Stroop interference into task, semantic, and response conflict, and Stroop facilitation into semantic and response facilitation. This exploratory study examined whether standard Stroop effects and their components occur when the target dimension is the physical size of written words.

Methods: In four experiments, Colombian university students who were native Spanish speakers (analyzed samples of 39, 39, 39, and 50 participants) named aloud the size, large or small, of stimuli presented at two letter sizes. Distractors included the size words GRANDE (large) and PEQUEÑO (small), size-related words (Experiment 1), size synonyms and antonyms (e.g., ENORME, DIMINUTO) (Experiment 2), words orthographically and phonologically similar (Experiment 3 and 4) to the response words, and control stimuli (symbol strings, pseudowords, and size-unrelated words). Vocal reaction times were analyzed with generalized linear mixed models (Gamma distribution, log link) and Holm-corrected pairwise comparisons, complemented by delta plots and simulation-based sensitivity analyses. Response types were analyzed with logistic mixed models.

Results: Standard Stroop interference (24 to 37 ms) and standard Stroop facilitation (20 to 36 ms) were observed in all four experiments, and errors were concentrated on incongruent size words. Interference grew with reaction time, whereas facilitation was present from the fastest responses. Response conflict and response facilitation were significant only in Experiment 1. Semantic conflict, semantic facilitation, and task conflict were not detected; size-unrelated words were not named more slowly than symbol strings. In Experiment 4, size-related words that resembled the correct response in form were named faster than words that resembled the opposite response. Each experiment could detect differences of about 29 to 34 ms with 80% power.

Conclusions: Naming the physical size of written words produces standard Stroop interference and facilitation. The absence of task conflict and the dominance of form over meaning are exploratory findings that require confirmation, because the experiments were not preregistered.

Keywords: Stroop task; size naming; size words; task conflict; response conflict; vocal reaction time; Spanish.

Background

The Stroop task is one of the most widely studied paradigms in psychology (e.g., [1, 2]). In the original task, participants were presented with color words in colors that did not match their meaning (e.g., the word "red" written in green ink or the word "blue" written in yellow ink). Participants were then asked to name the color of the word [3, 4]. Reaction times for naming responses are generally longer in these circumstances than when naming colors with no-color words (e.g., "house," "dog") or with a string of letters (e.g., "XXXXX"). This increase in reaction time is also called the "interference Stroop effect." Conversely, when the word's meaning matches its color, reaction times decrease relative to neutral trials (e.g., non-color words or letter strings). This is known as the "facilitation Stroop effect" [5].

Although most research on the Stroop effect has focused on written color words, the effect can be viewed more broadly as a task that pits two properties of a word against each other: first, the material properties of the word as a stimulus, i.e., what is asked to be named; and second, the meaning of the word. The Stroop task is designed so that the material property of the word is the "target" dimension while the meaning is the "distractor" dimension. Consequently, the Stroop effect can be observed in properties other than color. In the case of written words, the Stroop effect has also been observed with respect to the spatial position of the word [6, 7], the typeface of word [8], and big letters composed of small letters [9]. For spoken words, the Stroop effect has been observed in relation to the

gender of the speaker [10] or the pitch of the stimuli [11]. The Stroop effect has also been observed in sign language [12] and in musical notes written on a score [13, 14].

A variant of the Stroop effect is the semantic Stroop effect, in which words whose meaning is associated with the target dimension are used as distractors. When the target dimension is color, the congruent condition presents words such as "SKY" in blue or "GRASS" in green, while the incongruent condition presents words such as "SKY" in yellow and "GRASS" in red. Klein [15] first documented a semantic gradient ranging from meaningless symbols (e.g., an asterisk) to color-related words, then incongruent color-related words, and finally incongruent color words. Reaction times gradually increase along with this gradient. Several studies have observed interference and facilitation of the semantic Stroop effect with color-related words (e.g., [5, 16–18]).

The Stroop effect is a phenomenon associated with higher-level cognitive control of executive functions. It serves to align thought and behavior in the face of salient environmental distractions or interference [19, 20]. It measures the capacity for cognitive and motor inhibition, specifically the ability to suppress a dominant, habitual response (automatic reading) in order to execute an intentional, less automated response (naming the color) [21–24]. The Stroop task is the paradigmatic test for measuring selective attention. It assesses a subject's ability to focus their attention on a specific aspect of a stimulus (e.g. ink color) while ignoring other aspects [24–28]. The task has also been used to study the automation of word processing and visual text recognition, as well as to analyze whether semantic activation and access to meaning occur involuntarily and continuously upon mere visual exposure to the stimulus [2, 23, 29–31].

The Stroop interference effect is now considered not a single phenomenon, but a composite one involving various types of cognitive conflict during stimulus processing. These conflicts can be evaluated using different tests [16, 20, 22, 32, 33]. According to this multi-stage perspective, there are three distinct levels of conflict. Figure 1 illustrates the relationships between these stages, the conditions under which they are measured, and examples of these conditions. The first stage involves *task conflict*, arising from the exogenous activation of the irrelevant reading task, which interferes with the color-naming task prior to any semantic processing of the word. This is measured by comparing neutral, non-color-associated words against unreadable letter strings or symbols (e.g. 'DOG' in red versus 'XXXX' in red). The second stage is *semantic conflict*, which occurs during the conceptual encoding stage within the amodal semantic network. It is triggered by the simultaneous activation of two incompatible color representations. This is isolated by comparing reaction times for color-associated words that are not part of the direct response set (e.g. 'SKY' in red versus 'DOG' in red). The final stage, *response conflict*, arises during the late phase of motor selection and programming, when the distractor word induces a response belonging to the set of possible responses, thereby competing with the correct response. This is calculated by subtracting the reaction time of associated trials from that of standard incongruent trials (e.g., ‘BLUE’ in red minus ‘SKY’ in red).

From a multi-stage perspective, facilitation is considered a composite phenomenon resulting from the accumulation of benefits across different stages of information processing [20, 32, 34]. Overall Stroop facilitation comprises two key informational components (see Figure 1). First, *semantic facilitation* consists of the speed increase resulting from conceptual congruence or association between the relevant dimension and

the distractor's meaning. This occurs early on when the word activates the corresponding color concept within the amodal semantic network. Semantic facilitation is assessed by subtracting the reaction time for a congruent related word from that for a neutral word (e.g., "DOG" in red minus "SKY" in blue). Second, *response facilitation* represents the additional benefit that occurs during the later stages of response selection and production when the distractor word matches the correct response. Response facilitation is assessed by subtracting the reaction time for a congruent color word from that for a congruent associated word (e.g., "SKY" in blue minus "BLUE" in blue).

One material property of words for which the Stroop effect has not been evaluated is their size. Participants can be asked to name stimuli as large or small while the stimuli are the words “LARGE” and “SMALL” or words semantically related to size. Moreover, the multistage perspective has been investigated predominantly with color as the target dimension, and only a few studies have extended it to other tasks and dimensions, such as picture–word interference [35], spatial Stroop tasks [36], and Stroop-matching tasks [37]. The present study has two objectives. The first is to evaluate standard and semantic Stroop interference and facilitation when the target dimension is the size of written words. The second is to evaluate whether the components of interference and facilitation proposed by the multistage perspective are also present with size as the target dimension. The study is exploratory: the experiments were not preregistered, and each new experiment was motivated by the results of the previous one.

Methods

This section describes the elements that are common to the reported experiments.

Participants

The participants in this study were students at a private university in Bogotá, Colombia. Ninety percent of them belonged to the psychology program. Before the experiment began, they read and signed an informed consent form approved by the institution's Bioethics Committee. Participants under 18 years of age could take part in the study only if they brought a consent form signed by their parents. They received monetary compensation in Colombian pesos, which was approximately equivalent to US $5.60 at the time of the study, for completing the task.

Design

All experiments had an intra-subject design in which participants were exposed to seven conditions. The dependent variable was reaction time (RT) to vocally naming the size of the stimuli. The response set consisted of only two words in Spanish: "grande" (large) and "pequeño" (small). The independent variable that defined each experiment's conditions was the relationship between the semantic properties of the distractor and the size of the stimulus. All experiments included incongruent and congruent size-word conditions. The control condition of no size-related words was presented in all experiments. The first three experiments also included control conditions of symbol strings and pseudowords.

The estimates of facilitation and interference depend on the baseline used [38, 39]. Following Augustinova et al. [32], standard interference was computed against symbol strings, which carry no readable information, and standard facilitation against size-unrelated words, which share lexical status with the size words but have no size meaning. Pseudowords are intermediate controls that have a linguistic form but no meaning. We

156 therefore also report both effects against every available baseline (symbol strings,
157 pseudowords, and size-unrelated words).

158 Sample sizes were not determined by an a priori power analysis. The number of
159 participants was set to match or exceed that of within-subjects vocal Stroop studies with a
160 similar number of trials per condition (e.g., Experiment 2 in [32]; [34, 40]). These studies
161 are cited as references for the design, not for the method of analysis. To establish which
162 effects our samples could detect, we ran a simulation-based sensitivity analysis [41]. For
163 each experiment, we simulated data from the final fitted model (see Data Analysis),
164 keeping the actual number of participants, items, and valid trials, and set a difference of δ
165 ms between the incongruent size-word condition and the baseline (δ = 15, 25, 35, and 45
166 ms). Each simulated data set (150 per value of δ) was analyzed with the same model and
167 the same Holm correction over the 21 pairwise comparisons (α = .05). Power was the
168 proportion of simulations in which the difference was significant, and the minimum
169 difference detectable with 80% power was obtained by linear interpolation. The simulations
170 were run in R with *glmmTMB*, and the code is available on the Open Science Framework
171 (OSF). The results are reported at the beginning of the Results section of each experiment.

172 **Instruments and Stimuli**

173 The experiments were conducted in a psychology laboratory. The data were
174 collected individually. Participants sat in front of a computer monitor approximately 60 cm
175 away. The computer monitor was 23 inches, with a refresh rate of 60 Hz. The experimental
176 task was programmed in PsychoPy version 2024.2.1 [42]. The stimuli were symbol strings,
177 pseudowords, and words in black Arial font, all caps, over a white background. The stimuli
178 appeared in the center of the screen with a letter height of 100 pixels (large) or 20 pixels

(small); their width therefore varied with the number of characters. Table 1 lists the stimuli of each condition in each experiment, the size in which each one was presented, and the number of presentations per participant. Participants used a headset with a microphone to record their vocal responses, which were later analyzed. Table 1 also gives the Jaro-Winkler orthographic distance between each stimulus and the two response words.

Procedure

After signing the informed consent form, each participant was seated in front of a computer and read the following instructions in Spanish:

> In each trial, an object will be presented on the screen. It may be a symbol string or a word. The objects that will be presented are two sizes: LARGE or SMALL. For each object, say aloud and as quickly as possible GRANDE (large) or PEQUEÑO (small) according to its size. Remember to respond as quickly as possible because this is a reaction time experiment. Press the SPACE key when you are ready to start.

Before starting, the experimenter left the laboratory and left the participants alone until they finished the experiment. The experimental task was designed to present each stimulus successively without requiring a response from the participant to advance to the next trial. The only response participants had to provide was the vocal response to the size of the stimuli.

There was no practice block. The first ten trials of each participant served as familiarization and were excluded from all analyses. Because the task did not require a key press to advance from one trial to the next, engagement was assessed from the responses themselves. Participants with less than 70% correct responses across the 168 trials were

excluded: one in Experiment 1, three in Experiment 2, one in Experiment 3, and four in Experiment 4. Among the remaining participants, the mean percentage of correct responses ranged from 94.5% to 96.4% across experiments, and no participant fell below 71%.

Each stimulus was presented for 4 s, with an intertrial interval of 0.5 s. Although this presentation time is longer than is typical in Stroop studies, we did not consider that it affected participants' RTs. Naming responses were completed well before the stimulus disappeared (99% of correct responses occurred within 1,390 ms), so the remaining display time followed the response and could not affect RT on that trial. Its only consequence was a response–stimulus interval of about 3 to 4 s. Response–stimulus intervals between 2 and 5 s do not modulate Stroop interference [43], and only short intervals reduce some of its components [16]. Moreover, the timing was identical across conditions, so it cannot account for differences between them.

Each experiment consisted of three blocks of trials. Each block presented eight stimuli for each of the seven conditions, for a total of 56 trials per block and 168 trials per experiment. Half of the trials presented large stimuli, and the other half presented small stimuli. Trials were randomly presented in each block. The experiment lasted about 13 minutes for each participant.

Data Analysis

Responses were classified as correct (the size of the stimulus), errors (the opposite size), other responses (a word outside the response set or an unrecognizable response), or missing responses. Their frequencies by condition are reported in the Results of each experiment. These analyses included all trials after the first ten of each participant. Because responses were repeated within participants and items, each response type was analyzed at

the trial level with a logistic mixed model (that response type vs. all others), with condition as the fixed effect and random intercepts for participants and items, fitted with *glmmTMB* [44]. When the model with item intercepts did not converge, only participant intercepts were kept. The effect of condition was tested with a likelihood-ratio test against a model without condition. When this test was significant, pairwise comparisons between conditions are reported as odds ratios (OR) with 95% confidence intervals (CIs) and Holm correction.

In every experiment, the RTs of the conditions were compared. Vocal responses were recorded at 44.1 kHz during the 2 s following stimulus onset and were analyzed offline with custom Python scripts (available on OSF). Response onset was determined with Praat routines run through the Parselmouth library [45]: starting 350 ms after the beginning of the recording, onset was defined as the first point at which intensity reached 40 dB and was followed by at least 40 ms of continuous voicing (F0 > 60 Hz). When this criterion failed, onset was taken as the start of the response word estimated by automatic speech alignment with WhisperX [46]. RT was the time from the beginning of the recording to response onset; because recording started immediately before the screen refresh that displayed the stimulus, RTs include a delay of at most one refresh cycle (16.7 ms), which is constant across conditions. Each response was transcribed with the Google speech recognition service for Spanish, through the SpeechRecognition library, and matched to the two response words by approximate string matching (similarity ≥ 85%). Recordings without recognizable speech were classified as missing responses when their mean energy was below −35 dB, and as unrecognized responses otherwise. All automatic transcriptions were subsequently reviewed by the research team. Only correct responses were analyzed,

after excluding the first ten trials of each participant, which served to learn the two stimulus sizes, and the few correct responses whose onset could not be detected.

RTs of correct responses were analyzed at the trial level with generalized linear mixed models (GLMMs) with a Gamma distribution and a log link [47]. The models were fitted with the *glmmTMB* package [44] in R (version 4.5.0) [48]. This approach models the positive skew of RTs without transforming them. It uses every valid trial while taking into account that trials are nested within participants and items, and it handles the small differences in the number of trials per condition left by the exclusions. Condition (seven levels) was the fixed effect. Its overall effect was tested with a Wald χ^2 test using the *car* package [49]. The package versions used were glmmTMB 1.1.14, emmeans 2.0.2, DHARMa 0.4.7, and car 3.1.5.

For the random effects, we started from the maximal structure justified by the design [50]: random intercepts and condition slopes for participants, and random intercepts for items. Slopes were included because participants differ not only in overall speed but also in the size of their Stroop effects. Item intercepts were included because the words differ in length, frequency, and form. This model did not converge in Experiments 1 to 3 and, in Experiment 4, it fit worse than the simpler structures (higher Akaike and Bayesian information criteria, AIC and BIC; Additional file 1: Table S1), so we simplified it as recommended by Matuschek et al. [51]. The condition slopes were replaced by a random intercept for each participant-by-condition combination, which captures individual differences in the effects of condition with a single variance parameter. The final model was:

glmmTMB(latency ~ condition + (1 | participant) + (1 | participant:condition) + (1 | item), family = Gamma(link = "log"), data = data)

In Experiment 2, the participant-by-condition variance was estimated at zero, so that term was dropped. Model fit was checked with simulated quantile residuals using the *DHARMa* package [52]. Dispersion was adequate (1.07–1.12), and the proportion of outliers was within the expected range. The Kolmogorov–Smirnov tests detected small deviations from uniformity (D = .040–.050), which are expected with several thousand observations (Additional file 1: Table S6). Because the study was exploratory, we did not define a subset of primary contrasts. All 21 pairwise comparisons between conditions were computed in each experiment with the *emmeans* package [53] and corrected with the Holm [54] method. Each difference is reported in milliseconds, back-transformed from the log scale, with its standard error (SE) and 95% CI.

To check that the conclusions did not depend on the modeling approach, we also ran a repeated-measures analysis of variance (ANOVA) on each participant's mean RT per condition, with the Greenhouse–Geisser correction. Additional file 1 also includes these participant means (Table S8). The ANOVA showed an effect of condition in every experiment (all $ps < .001$, $\eta^2p = .24–.30$), consistent with the GLMM. The full results are in Additional file 1: Table S7.

To examine how interference and facilitation change across the RT distribution, we computed delta plots [55]. For each participant and condition, we obtained the .1, .3, .5, .7, and .9 quantiles of correct RTs and averaged them across participants (Vincentizing) [56]. At each quantile, the effect (the difference between two conditions) was plotted against the mean RT of those two conditions (see [57]).

Generative AI (Claude, Anthropic) was used to assist in writing the analysis and figure scripts, drafting and editing parts of the manuscript, and formatting the references. The authors reviewed and verified all outputs and take full responsibility for the content of this article.

Experiment 1

The objective of the first experiment is to evaluate standard Stroop interference and facilitation using size words. The experiment will also evaluate semantic Stroop interference and facilitation using size-related words.

Methods

Participants

40 college students, 25 female (62.5%) and 15 male (37.5%), aged between 16 and 30 years (M = 18.7; SD = 2.2). One participant was excluded for having less than 70% correct responses (see Procedure), so the analyses included 39 participants.

Design and Stimuli

Participants were exposed to seven conditions (Table 1 lists all stimuli with their English translations). Three were control conditions: symbol strings, pseudowords, and words unrelated to size. The other four combined two types of words with two size relations. Size-related words name objects that are large (MONTAÑA, BALLENA) or small (HORMIGA, SEMILLA), and size words are the response words themselves (GRANDE, PEQUEÑO). Both types were presented either in the size that matched their meaning (congruent) or in the opposite size (incongruent). Control stimuli were presented equally often in both sizes.

Results

With 39 participants, Experiment 1 had 80% power to detect differences of about 29 ms between conditions, with Holm correction over the 21 comparisons. Power was .17 for a difference of 15 ms, .68 for 25 ms, and .96 for 35 ms.

Of the 6,162 trials analyzed, 5,940 (96.4%) were correct, 75 (1.2%) were errors, 39 (0.6%) were other responses, and 108 (1.8%) were missing responses (Table 2). Condition affected the proportion of errors, $\chi^2(6) = 68.53$, $p < .001$, and of correct responses, $\chi^2(6) = 49.87$, $p < .001$. It did not affect the proportion of missing responses, $\chi^2(6) = 10.21$, $p = .116$, or of other responses, $\chi^2(6) = 2.34$, $p = .886$. Incongruent size words produced more errors (42) than every other condition (ORs = 5.5–29.9, all ps < .001), and the remaining conditions did not differ from each other. Accordingly, correct responses were less frequent with incongruent size words than with symbol strings, pseudowords, size-unrelated words, congruent size-related words, and congruent size words (ORs = 0.22–0.42, ps ≤ .003). They were also less frequent with incongruent size-related words than with congruent size words (OR = 0.41, $p = .048$).

The RT analysis included 5,937 correct responses from 39 participants; three correct responses were excluded because their onset could not be detected. Condition had a reliable effect on RTs, Wald $\chi^2(6) = 100.60$, $p < .001$. Figure 2 shows the estimated marginal means with their 95% confidence intervals, and Additional file 1: Table S2 shows the model parameters. From fastest to slowest, the conditions were congruent size words (M = 649.0 ms, 95% CI [616.3, 683.5]), congruent size-related words (M = 669.9, [636.1, 705.5]), incongruent size-related words (M = 678.5, [644.3, 714.6]), size-unrelated words (M = 680.0, [645.7, 716.2]), pseudowords (M = 688.9, [654.2, 725.5]), symbol strings (M =

695.4, [660.4, 732.4]), and incongruent size words (M = 719.0, [682.7, 757.3]). In the following paragraphs, p values are Holm-corrected across the 21 pairwise comparisons, and confidence intervals are unadjusted.

Standard Stroop interference was observed: incongruent size words were named more slowly than symbol strings (Δ = 23.6 ms, SE = 7.8, z = 3.02, p = .028, 95% CI [8.3, 38.9]). Interference was also present against the other two baselines: incongruent size words were slower than pseudowords (Δ = 30.1 ms, SE = 7.8, z = 3.86, p = .001, 95% CI [14.8, 45.3]) and size-unrelated words (Δ = 39.0 ms, SE = 7.8, z = 5.00, p < .001, 95% CI [23.7, 54.2]; Table 3). Of its three components, only response conflict was significant: incongruent size words were slower than incongruent size-related words (Δ = 40.5 ms, SE = 7.8, z = 5.18, p < .001, 95% CI [25.2, 55.8]). Semantic conflict, the difference between incongruent size-related words and size-unrelated words, was not significant (Δ = −1.5 ms, SE = 7.4, z = −0.20, p = .990, 95% CI [−16.1, 13.1]). Task conflict was not significant either, and its direction was opposite to that expected: size-unrelated words were named faster than symbol strings (Δ = −15.4 ms, SE = 7.5, z = −2.05, p = .282, 95% CI [−30.1, −0.7]).

Standard Stroop facilitation was observed: congruent size words were named faster than size-unrelated words (Δ = 31.0 ms, SE = 7.3, z = 4.25, p < .001, 95% CI [16.7, 45.3]). Facilitation was also present against pseudowords (Δ = 39.9 ms, SE = 7.3, z = 5.43, p < .001, 95% CI [25.5, 54.3]) and symbol strings (Δ = 46.4 ms, SE = 7.4, z = 6.27, p < .001, 95% CI [31.9, 60.9]; Table 3), so the standard effects did not depend on the choice of baseline. Response facilitation was significant: congruent size words were faster than congruent size-related words (Δ = 20.8 ms, SE = 7.2, z = 2.89, p = .039, 95% CI [6.7,

35.0]). Semantic facilitation, the difference between size-unrelated words and congruent size-related words, was not significant (Δ = 10.1 ms, SE = 7.4, z = 1.37, p = .990, 95% CI [−4.3, 24.6]).

Figure 3 shows the delta plots of standard interference and facilitation. Interference was absent in the fastest responses (−8.0 ms at the .1 quantile, 95% CI [−27.1, 11.1]) and increased steadily across the distribution, reaching 57.5 ms at the .9 quantile ([16.9, 98.1]). Facilitation, in contrast, was already present in the fastest responses (26.1 ms at the .1 quantile, [11.0, 41.2]) and remained roughly constant between 36 and 40 ms in the remaining quantiles. Thus, the interference produced by incongruent size words emerged only in slower responses, whereas the benefit of congruent size words affected responses of all speeds.

The size-related words did not produce semantic effects. As reported above, incongruent size-related words were not slower than size-unrelated words, and congruent size-related words were not faster than them. Compared with the other controls, congruent size-related words were faster than symbol strings (Δ = 25.5 ms, SE = 7.5, z = 3.42, p = .008, 95% CI [10.9, 40.2]) but not than pseudowords (Δ = 19.0 ms, SE = 7.4, z = 2.56, p = .094, 95% CI [4.5, 33.6]), and incongruent size-related words did not differ from symbol strings (Δ = 16.9 ms, SE = 7.5, z = 2.25, p = .198, 95% CI [2.2, 31.6]) or pseudowords (Δ = 10.4 ms, SE = 7.5, z = 1.39, p = .990, 95% CI [−4.3, 25.0]). The two size-related conditions did not differ from each other (Δ = 8.6 ms, SE = 7.4, z = 1.17, p = .990, 95% CI [−5.9, 23.1]). Both differed from the standard condition of the opposite relation: incongruent size-related words were slower than congruent size words (Δ = 29.5 ms, SE = 7.3, z = 4.04, p < .001, 95% CI [15.2, 43.8]), and congruent size-related words were faster than incongruent

size words (Δ = 49.1 ms, SE = 7.8, z = 6.32, $p < .001$, 95% CI [33.9, 64.3]). However, they also differed from the standard condition with the same relation, as shown by the significant response conflict and response facilitation. Thus, size-related words behaved like the control conditions: their RTs fell between those of the two size-word conditions and did not depend on whether their meaning matched the presented size. The difference between congruent size-related words and symbol strings probably reflects the faster naming of words relative to symbol strings, because size-unrelated words were also numerically faster than symbol strings, rather than a semantic effect.

Experiment 1 showed that the Stroop effect occurs when the target dimension is the size of words: incongruent size words slowed size naming, and congruent size words speeded it. Both effects were accompanied by response-level components (response conflict and response facilitation). In contrast, size-related words produced no semantic effects: they did not differ from size-unrelated words or from each other, regardless of whether their meaning matched the presented size, and they differed from the size words of both relations. Semantic conflict, semantic facilitation, and task conflict were therefore not observed. These null results must be interpreted with caution. They were smaller than the minimum difference this experiment could detect with 80% power (about 29 ms), so it cannot be concluded that these components are absent with size; they were simply not detected with this sample and these data. To increase the potential semantic effects, the following experiment used synonyms and antonyms of the response words. In languages such as Spanish and English, finding synonyms or antonyms for color words is difficult, whereas it is easy for the words “large” and “small.” The following experiment evaluates this idea.

Experiment 2

Like Experiment 1, Experiment 2 has seven conditions, although the congruent and incongruent size-related word conditions were replaced with size synonym and size antonym conditions, respectively. The idea is that size-synonyms would produce a semantic facilitation effect while size-antonyms would produce semantic interference. Two Spanish words synonymous with “grande” (large) were chosen as antonyms of “pequeño” (small), and two words synonymous with “pequeño” were chosen as antonyms of “grande”. The words were either synonyms or antonyms with respect to the size presented, which participants had to identify. The other five conditions were the same as in the previous experiment.

Methods

Participants

The participants were 42 college students, 27 of whom were female (64.3%), 14 male (33.3%), and 1 non-binary (2.4%). The participants' ages ranged from 17 to 29 years (M = 18.9; SD = 2.0). Three participants were excluded for having less than 70% correct responses (see Procedure), so the analyses included 39 participants.

Design and Stimuli

The three control conditions and the congruent and incongruent size-word conditions of Experiment 1 were retained. The size-related words were replaced by synonyms of the two response words (ENORME and GIGANTE for GRANDE; MINÚSCULO and DIMINUTO for PEQUEÑO). Presented in the size they denote, these words were synonyms of the correct response (size synonyms); presented in the opposite size, they were antonyms of it (size antonyms; Table 1).

Results

With 39 participants, Experiment 2 had 80% power to detect differences of about 29 ms. Power was .16 for 15 ms, .71 for 25 ms, and .95 for 35 ms.

Of the 6,162 trials analyzed, 5,825 (94.5%) were correct, 39 (0.6%) were errors, 92 (1.5%) were other responses, and 206 (3.3%) were missing responses (Table 4). Condition affected the proportion of errors, $\chi^2(6) = 54.35$, $p < .001$, and of correct responses, $\chi^2(6) = 27.69$, $p < .001$. It did not affect missing responses, $\chi^2(6) = 7.80$, $p = .253$, or other responses, $\chi^2(6) = 4.66$, $p = .588$. Incongruent size words produced more errors (29) than size antonyms, symbol strings, pseudowords, size-unrelated words, and congruent size words (ORs = 8.1–33.5, ps ≤ .011). Size synonyms produced no errors, so their comparison could not be estimated. Correct responses were less frequent with incongruent size words than with symbol strings, pseudowords, size synonyms, and congruent size words (ORs = 0.37–0.53, ps ≤ .026).

The RT analysis included 5,819 correct responses from 39 participants; six correct responses were excluded because their onset could not be detected. Condition had a reliable effect on RTs, Wald $\chi^2(6) = 72.88$, $p < .001$. Figure 4 shows the estimated marginal means with their 95% confidence intervals, and Additional file 1: Table S3 shows the model parameters. From fastest to slowest, the conditions were congruent size words (M = 672.6 ms, 95% CI [638.0, 709.2]), size synonyms (M = 690.2, [654.7, 727.8]), size-unrelated words (M = 693.0, [657.3, 730.7]), symbol strings (M = 698.5, [662.5, 736.5]), pseudowords (M = 704.9, [668.6, 743.3]), size antonyms (M = 709.5, [672.9, 748.1]), and incongruent size words (M = 730.3, [692.6, 770.1]). As in Experiment 1, p values are

Holm-corrected across the 21 pairwise comparisons, and confidence intervals are unadjusted.

Standard Stroop interference was observed: incongruent size words were named more slowly than symbol strings (Δ = 31.8 ms, SE = 7.5, z = 4.25, p < .001, 95% CI [17.1, 46.4]). Interference was also significant against pseudowords (Δ = 25.3 ms, SE = 7.5, z = 3.38, p = .010, 95% CI [10.6, 40.1]) and size-unrelated words (Δ = 37.2 ms, SE = 7.5, z = 4.99, p < .001, 95% CI [22.6, 51.9]; Table 3). None of its components reached significance. In this experiment, size antonyms took the place of incongruent size-related words. Response conflict, the difference between incongruent size words and size antonyms, was not significant (Δ = 20.8 ms, SE = 7.5, z = 2.76, p = .070, 95% CI [6.0, 35.6]). Semantic conflict, the difference between size antonyms and size-unrelated words, was not significant either (Δ = 16.5 ms, SE = 7.3, z = 2.27, p = .210, 95% CI [2.2, 30.7]). Task conflict, the difference between size-unrelated words and symbol strings, was again negative (Δ = −5.5 ms, SE = 7.2, z = −0.76, p = 1.000, 95% CI [−19.5, 8.5]).

Standard Stroop facilitation was observed: congruent size words were named faster than size-unrelated words (Δ = 20.4 ms, SE = 7.0, z = 2.91, p = .047, 95% CI [6.7, 34.1]). Facilitation was also significant against pseudowords (Δ = 32.3 ms, SE = 7.1, z = 4.53, p < .001, 95% CI [18.3, 46.3]) and symbol strings (Δ = 25.9 ms, SE = 7.1, z = 3.66, p = .004, 95% CI [12.0, 39.7]). Neither of its components was significant. Response facilitation, the difference between size synonyms and congruent size words, was 17.6 ms (SE = 7.0, z = 2.52, p = .116, 95% CI [3.9, 31.3]). Semantic facilitation, the difference between size-unrelated words and size synonyms, was practically zero (Δ = 2.8 ms, SE = 7.1, z = 0.40, p = 1.000, 95% CI [−11.1, 16.7]).

Figure 5 shows the delta plots. The pattern was the same as in Experiment 1. Interference was absent in the fastest responses (−2.4 ms at the .1 quantile, 95% CI [−16.5, 11.7]) and increased steadily to 80.7 ms at the .9 quantile ([40.3, 121.1]). Facilitation was present from the fastest responses (22.3 ms, [8.7, 35.9]) and remained between 22 and 29 ms across the distribution, although its confidence interval included zero at the .9 quantile.

Size synonyms and antonyms shifted RTs in the expected directions, but by less than the minimum effect this experiment could detect. As reported above, neither differed significantly from size-unrelated words: size antonyms were 16.5 ms slower and size synonyms 2.8 ms faster. They did not differ from the other controls either: size antonyms did not differ from symbol strings (Δ = 11.0 ms, SE = 7.3, z = 1.51, p = .790, 95% CI [−3.3, 25.3]) or pseudowords (Δ = 4.6 ms, SE = 7.3, z = 0.62, p = 1.000, 95% CI [−9.8, 18.9]), and size synonyms did not differ from symbol strings (Δ = 8.3 ms, SE = 7.1, z = 1.16, p = 1.000, 95% CI [−5.7, 22.2]) or pseudowords (Δ = 14.7 ms, SE = 7.2, z = 2.05, p = .323, 95% CI [0.6, 28.8]). Size antonyms were slower than size synonyms, but the difference was not significant (Δ = 19.3 ms, SE = 7.2, z = 2.66, p = .085, 95% CI [5.1, 33.4]). Each condition differed from the standard condition of the opposite relation: size antonyms were slower than congruent size words (Δ = 36.9 ms, SE = 7.2, z = 5.12, p < .001, 95% CI [22.8, 50.9]), and size synonyms were faster than incongruent size words (Δ = 40.0 ms, SE = 7.4, z = 5.38, p < .001, 95% CI [25.5, 54.6]). In contrast, neither differed significantly from the standard condition with the same relation, as shown by the non-significant response conflict and response facilitation. Thus, synonyms and antonyms were closer to the size word with the same relation than to the size word with the opposite relation. This pattern suggests a

small semantic effect that drew each condition toward the size word it shared its relation with, although the effect could not be detected against the control conditions.

Experiment 2 replicated standard Stroop interference and facilitation with size words. Size synonyms and antonyms produced a pattern compatible with a small semantic effect: each differed from the size word of the opposite relation but not from the size word with the same relation, and antonyms were numerically slower than synonyms. However, neither condition differed from the control conditions, so semantic conflict and semantic facilitation were not detected, and no component of interference or facilitation reached significance. Response conflict (20.8 ms), semantic conflict (16.5 ms), and response facilitation (17.6 ms) were all below the minimum difference this experiment could detect with 80% power (about 29 ms). Therefore, these components cannot be considered absent; they were not detected with this sample and these data. As in Experiment 1, the difference between size-unrelated words and symbol strings went in the direction opposite to task conflict, although it was not significant.

Experiment 3

Some studies have demonstrated facilitation and interference effects due to the orthographic and phonological similarities between the stimulus and response sets of words (e.g., [58, 59]). In this study, we tested this effect using size words. As in Experiments 1 and Experiment 2, this experiment includes seven conditions. However, we replaced the size-synonyms and antonyms conditions from Experiment 2 with two conditions of words similar orthographically and phonologically to the size words, but are either congruent or incongruent with that size. We selected two Spanish words that are similar to GRANDE (big) and two that are similar to PEQUEÑO (small). The condition was considered

congruent when the words were similar to the one corresponding to the size to be named; and incongruent when the size was the opposite. The other five conditions from Experiment 1 and Experiment 2 were retained.

Methods

Participants

The participants were 40 college students, 25 female (62.5%) and 15 male (37.5%). Ages ranged from 17 to 25 years (M = 20.3; SD = 2.0). One participant was excluded for having less than 70% correct responses (see Procedure), so the analyses included 39 participants.

Design and Stimuli

The three control conditions and the two size-word conditions were retained. The two new conditions used words without size meaning that resemble one of the response words in spelling and sound: GRACIAS and GRANUJA resemble GRANDE, and PAQUETE and PECADO resemble PEQUEÑO. These words were presented either in the size named by the response word they resemble (congruent similar) or in the opposite size (incongruent similar; Table 1).

In this experiment, the stimuli for the size-unrelated words condition were changed to observe whether the behavior of this condition was dependent on the stimuli used in previous experiments. The words that comprised this condition in this experiment and the following one were: “NÚMERO”, “CALIDAD”, “ESTILO”, and “RELATO” (“number”, “quality”, “style”, and “story”).

Results

With 39 participants, Experiment 3 had 80% power to detect differences of about 30 ms. Power was .22 for 15 ms, .65 for 25 ms, and .96 for 35 ms.

Of the 6,162 trials analyzed, 5,906 (95.8%) were correct, 32 (0.5%) were errors, 85 (1.4%) were other responses, and 139 (2.3%) were missing responses (Table 5). Condition affected errors, $\chi^2(6) = 67.36$, $p < .001$; correct responses, $\chi^2(6) = 21.08$, $p = .002$; and missing responses, $\chi^2(6) = 13.22$, $p = .040$. It did not affect other responses, $\chi^2(6) = 8.27$, $p = .219$. Incongruent size words produced 25 errors, compared with one or two in each of the other conditions (ORs = 13.0–26.2, ps ≤ .028). Correct responses were less frequent with incongruent size words than with congruent size words (OR = 0.41, $p = .006$) and congruent similar words (OR = 0.45, $p = .018$). For missing responses, no pairwise comparison remained significant after Holm correction.

The RT analysis included 5,904 correct responses from 39 participants; two correct responses were excluded because their onset could not be detected. Condition had a reliable effect on RTs, Wald $\chi^2(6) = 102.48$, $p < .001$. Figure 6 shows the estimated marginal means with their 95% confidence intervals, and Additional file 1: Table S4 shows the model parameters. From fastest to slowest, the conditions were congruent size words (M = 669.5 ms, 95% CI [640.3, 700.1]), size-unrelated words (M = 698.4, [667.9, 730.3]), congruent similar words (M = 701.0, [670.3, 733.0]), pseudowords (M = 709.5, [678.5, 742.0]), incongruent similar words (M = 710.4, [679.3, 742.9]), symbol strings (M = 712.9, [681.7, 745.5]), and incongruent size words (M = 741.2, [708.8, 775.2]). As in the previous experiments, p values are Holm-corrected across the 21 pairwise comparisons, and confidence intervals are unadjusted.

Standard Stroop interference was observed: incongruent size words were named more slowly than symbol strings (Δ = 28.3 ms, SE = 7.6, z = 3.73, p = .002, 95% CI [13.5, 43.2]). Interference was also significant against pseudowords (Δ = 31.7 ms, SE = 7.6, z = 4.18, p < .001, 95% CI [16.9, 46.6]) and size-unrelated words (Δ = 42.8 ms, SE = 7.6, z = 5.67, p < .001, 95% CI [28.0, 57.6]; Table 3). Because this experiment did not include size-related words, semantic conflict and response conflict could not be estimated as defined by Augustinova et al. [32]. Task conflict, the difference between size-unrelated words and symbol strings, was again negative and not significant (Δ = −14.5 ms, SE = 7.3, z = −1.99, p = .468, 95% CI [−28.8, −0.2]).

Standard Stroop facilitation was observed: congruent size words were named faster than size-unrelated words (Δ = 28.9 ms, SE = 7.1, z = 4.08, p < .001, 95% CI [15.0, 42.7]). Facilitation was also significant against pseudowords (Δ = 40.0 ms, SE = 7.2, z = 5.59, p < .001, 95% CI [26.0, 54.0]) and symbol strings (Δ = 43.4 ms, SE = 7.2, z = 6.03, p < .001, 95% CI [29.3, 57.5]). For the same reason as above, semantic facilitation and response facilitation could not be estimated.

Figure 7 shows the delta plots. As in the previous experiments, interference was absent in the fastest responses (−3.9 ms at the .1 quantile, 95% CI [−20.7, 12.9]) and increased steadily to 73.7 ms at the .9 quantile ([35.3, 112.1]). Facilitation was small in the fastest responses (13.1 ms, [−1.5, 27.7]) and grew moderately across the distribution, reaching 44.8 ms at the .9 quantile ([13.1, 76.5]). In this experiment, therefore, facilitation also increased with response time, although less steeply than interference.

Words orthographically and phonologically similar to the response words produced neither conflict nor facilitation. Incongruent similar words did not differ from size-

unrelated words (Δ = 12.0 ms, SE = 7.3, z = 1.64, p = .913, 95% CI [−2.4, 26.3]), and congruent similar words were not faster than size-unrelated words (Δ = −2.6 ms, SE = 7.2, z = −0.36, p = 1.000, 95% CI [−16.7, 11.6]). Neither condition differed from the other controls: congruent similar words did not differ from symbol strings (Δ = 11.9 ms, SE = 7.3, z = 1.63, p = .913, 95% CI [−2.4, 26.2]) or pseudowords (Δ = 8.6 ms, SE = 7.3, z = 1.18, p = 1.000, 95% CI [−5.7, 22.8]), and incongruent similar words did not differ from symbol strings (Δ = 2.5 ms, SE = 7.4, z = 0.34, p = 1.000, 95% CI [−12.0, 17.0]) or pseudowords (Δ = 0.8 ms, SE = 7.4, z = 0.11, p = 1.000, 95% CI [−13.6, 15.3]). The two similar-word conditions did not differ from each other (Δ = 9.4 ms, SE = 7.3, z = 1.28, p = 1.000, 95% CI [−5.0, 23.8]). Both differed from the size words of either relation: incongruent size words were slower than incongruent similar words (Δ = 30.9 ms, SE = 7.6, z = 4.05, p < .001, 95% CI [15.9, 45.8]) and congruent similar words (Δ = 40.3 ms, SE = 7.6, z = 5.33, p < .001, 95% CI [25.5, 55.1]), and congruent size words were faster than congruent similar words (Δ = 31.5 ms, SE = 7.1, z = 4.44, p < .001, 95% CI [17.6, 45.4]) and incongruent similar words (Δ = 40.9 ms, SE = 7.2, z = 5.67, p < .001, 95% CI [26.7, 55.0]). In short, the two size-word conditions differed from all the other conditions, whereas none of the other five conditions differed from each other. Words that only resemble the response words behaved like the control conditions.

Experiment 3 again showed standard Stroop interference and facilitation, of similar magnitude to those of the previous experiments. Words orthographically and phonologically similar to GRANDE or PEQUEÑO, but unrelated to size, produced neither conflict nor facilitation: they did not differ from the control conditions or from each other, and they differed from the size words of both relations. However, the largest of these

differences (9 to 12 ms) were well below the minimum difference this experiment could detect with 80% power (about 30 ms), so an effect of orthographic and phonological similarity cannot be ruled out; it was not detected with this sample and these data. The control conditions showed the same order as in Experiment 1, with symbol strings slowest and size-unrelated words fastest. The difference between size-unrelated words and symbol strings again went in the direction opposite to task conflict, even though the size-unrelated words were different from those of Experiments 1 and 2. This suggests that this pattern does not depend on the specific words used.

Experiment 4

In Experiment 3, words that are similar orthographically and phonologically to the size words were used, although without any semantic relationship to size. The effect on RTs was very small. This experiment explores the combined effect of orthographic and phonological similarity with the semantic relationship of the words to the sizes. This combined effect yields four new related word conditions that are either similar or dissimilar in terms of orthography and phonology, and can be either congruent or incongruent with respect to stimulus size. Seven conditions were maintained for this experiment to avoid making the task longer than in the previous experiments. To achieve this, two of the control conditions were removed: symbol strings and pseudowords, retaining only size-unrelated words. This condition was kept instead of the symbol strings because the latter tended to produce longer RTs, much closer to the incongruent size words. The congruent and incongruent size word conditions were also retained.

Methods

Participants

Fifty-four college students participated, 37 female (68.5%) and 17 male (31.5%). Ages ranged from 18 to 33 years (M = 20.9; SD = 2.5). Four participants were excluded for having less than 70% correct responses (see Procedure), so the analyses included 50 participants.

Design and Stimuli

Size-unrelated words and congruent and incongruent size words were retained. The four new conditions used size-related words that resemble one of the response words in spelling and sound. In two pairs, meaning and form point to the same size: GRUESO and GRANJA are large and resemble GRANDE, and PICADURA and POQUITO are small and resemble PEQUEÑO. In the other two pairs, meaning and form point to opposite sizes: GRAMO and GRANADA are small but resemble GRANDE, and PAQUIDERMO and PAQUETÓN are large but resemble PEQUEÑO. Each trial was classified by two features: whether the word resembled the correct response (similar) or the opposite response (dissimilar), and whether its size meaning matched the presented size (congruent) or not (incongruent). Each word therefore appeared in two conditions, depending on the size in which it was presented (Table 1).

Results

With 50 participants, Experiment 4 had 80% power to detect differences of about 34 ms. Power was .21 for 15 ms, .46 for 25 ms, and .85 for 35 ms.

Of the 7,900 trials analyzed, 7,483 (94.7%) were correct, 82 (1.0%) were errors, 70 (0.9%) were other responses, and 265 (3.4%) were missing responses (Table 6). Condition

affected errors, $\chi^2(6) = 74.74$, $p < .001$, and correct responses, $\chi^2(6) = 49.15$, $p < .001$. It did not affect missing responses, $\chi^2(6) = 5.99$, $p = .424$, or other responses, $\chi^2(6) = 3.97$, $p = .680$. Incongruent size words produced more errors (50) than every other condition (ORs = 6.5–29.7, all ps < .001), and fewer correct responses (ORs = 0.31–0.55, ps ≤ .036).

The RT analysis included 7,482 correct responses from 50 participants; one correct response was excluded because its onset could not be detected. Condition had a reliable effect on RTs, Wald $\chi^2(6) = 149.21$, $p < .001$. Figure 8 shows the estimated marginal means with their 95% confidence intervals, and Additional file 1: Table S5 shows the model parameters. From fastest to slowest, the conditions were congruent size words (M = 662.3 ms, 95% CI [630.3, 695.9]), similar congruent words (M = 683.1, [650.8, 717.0]), similar incongruent words (M = 685.7, [653.3, 719.7]), size-unrelated words (M = 698.4, [665.4, 733.1]), dissimilar incongruent words (M = 711.3, [677.7, 746.7]), dissimilar congruent words (M = 712.5, [678.8, 747.8]), and incongruent size words (M = 735.4, [699.8, 772.8]). As in the previous experiments, p values are Holm-corrected across the 21 pairwise comparisons, and confidence intervals are unadjusted.

Because this experiment did not include symbol strings or pseudowords, standard interference was estimated against size-unrelated words, the only control condition. Incongruent size words were named more slowly than size-unrelated words (Δ = 37.0 ms, SE = 9.7, z = 3.83, p = .002, 95% CI [18.0, 55.9]; Table 3). Dissimilar incongruent words, whose meaning and form both point to the opposite response, took the place of incongruent size-related words. Response conflict, the difference between incongruent size words and dissimilar incongruent words, was not significant (Δ = 24.0 ms, SE = 9.7, z = 2.47, p = .122, 95% CI [5.0, 43.1]). Semantic conflict, the difference between dissimilar incongruent

words and size-unrelated words, was not significant either (Δ = 12.9 ms, SE = 8.6, z = 1.50, p = .526, 95% CI [−3.9, 29.8]). Task conflict could not be estimated.

Standard Stroop facilitation was observed: congruent size words were named faster than size-unrelated words (Δ = 36.1 ms, SE = 9.0, z = 4.00, p < .001, 95% CI [18.4, 53.8]). Similar congruent words, whose meaning and form both point to the correct response, took the place of congruent size-related words. Response facilitation, the difference between similar congruent words and congruent size words, was not significant (Δ = 20.7 ms, SE = 8.9, z = 2.33, p = .147, 95% CI [3.3, 38.2]). Semantic facilitation, the difference between size-unrelated words and similar congruent words, was not significant either (Δ = 15.3 ms, SE = 8.4, z = 1.82, p = .409, 95% CI [−1.1, 31.8]).

Figure 9 shows the delta plots, with interference computed against size-unrelated words. Interference was again absent in the fastest responses (3.6 ms at the .1 quantile, 95% CI [−18.5, 25.7]) and increased to 75.0 ms at the .9 quantile ([44.9, 105.1]). Facilitation was present from the fastest responses (27.3 ms, [11.1, 43.5]) and increased to 54.5 ms at the .9 quantile ([26.9, 82.1]). As in Experiment 3, facilitation grew with response time, although interference was the effect most dependent on response speed.

None of the four new conditions differed from size-unrelated words. Besides the semantic conflict and facilitation reported above, similar incongruent words did not differ from size-unrelated words (Δ = 12.7 ms, SE = 8.4, z = 1.51, p = .526, 95% CI [−3.8, 29.2]), and neither did dissimilar congruent words (Δ = 14.0 ms, SE = 8.6, z = 1.64, p = .510, 95% CI [−2.8, 30.9]). Among the four new conditions, all comparisons between dissimilar and similar words were significant: dissimilar incongruent vs. similar congruent (Δ = 28.3 ms, SE = 6.8, z = 4.18, p < .001, 95% CI [15.0, 41.5]), dissimilar congruent vs. similar

incongruent (Δ = 26.7 ms, SE = 6.7, z = 3.97, p = .001, 95% CI [13.5, 40.0]), dissimilar congruent vs. similar congruent (Δ = 29.4 ms, SE = 8.5, z = 3.45, p = .007, 95% CI [12.7, 46.1]), and dissimilar incongruent vs. similar incongruent (Δ = 25.6 ms, SE = 8.5, z = 3.00, p = .030, 95% CI [8.9, 42.4]). In contrast, congruent and incongruent words did not differ when they had the same form: similar congruent vs. similar incongruent (Δ = 2.6 ms, SE = 8.3, z = 0.32, p = 1.000, 95% CI [−13.7, 19.0]) and dissimilar congruent vs. dissimilar incongruent (Δ = 1.1 ms, SE = 8.7, z = 0.13, p = 1.000, 95% CI [−15.9, 18.1]). Similar words differed from incongruent size words (similar congruent: Δ = 52.3 ms, SE = 9.6, z = 5.44, p < .001, 95% CI [33.5, 71.2]; similar incongruent: Δ = 49.7 ms, SE = 9.6, z = 5.16, p < .001, 95% CI [30.8, 68.5]) but not from congruent size words (similar incongruent: Δ = 23.4 ms, SE = 8.9, z = 2.62, p = .087, 95% CI [5.9, 40.9]; for similar congruent, see response facilitation). Conversely, dissimilar words differed from congruent size words (dissimilar congruent: Δ = 50.1 ms, SE = 9.1, z = 5.49, p < .001, 95% CI [32.2, 68.0]; dissimilar incongruent: Δ = 49.0 ms, SE = 9.1, z = 5.36, p < .001, 95% CI [31.1, 66.9]) but not from incongruent size words (dissimilar congruent: Δ = 22.9 ms, SE = 9.7, z = 2.36, p = .147, 95% CI [3.9, 42.0]; for dissimilar incongruent, see response conflict). Thus, RTs for these words depended on whether their form resembled the correct or the opposite response, by about 26 to 29 ms, and not on whether their size meaning matched the presented size: words that resembled the correct response grouped with congruent size words, and words that resembled the opposite response grouped with incongruent size words.

Although congruence produced no significant differences, the congruent conditions occupied the extremes of the four new conditions: similar congruent words were the fastest

and dissimilar congruent words the slowest. Consequently, the effect of form was slightly larger for congruent words (29.4 ms) than for incongruent words (25.6 ms), but this difference was far from significant (interaction contrast, unadjusted: $\Delta = 3.8$ ms, SE = 14.1, $z = 0.27$, $p = .790$, 95% CI [−23.9, 31.4]). If confirmed, it would suggest that a size meaning that matches the presented size does not act as an independent source of facilitation but slightly amplifies the effect of form, whichever response the form points to.

Experiment 4 replicated standard Stroop interference and facilitation, both measured against size-unrelated words. When size meaning and orthographic and phonological similarity were combined, form prevailed over meaning. Words that resembled the correct response were named about 26 to 29 ms faster than words that resembled the opposite response, regardless of their size meaning, and they did not differ from congruent size words, whereas words that resembled the opposite response did not differ from incongruent size words. Size meaning made no detectable difference once form was held constant; congruent words were only slightly more extreme than incongruent words, a difference that was not significant. This contrasts with Experiment 3, in which similarity alone, without size meaning, produced no reliable effect. Response conflict (24.0 ms), response facilitation (20.7 ms), semantic conflict (12.9 ms), and semantic facilitation (15.3 ms) were not significant. All were below the minimum difference this experiment could detect with 80% power (about 34 ms), so they cannot be considered absent; they were not detected with this sample and these data.

Discussion

Summary of Findings

Across four experiments, participants named the physical size of written words whose meaning was congruent, incongruent, or unrelated to that size. Table 7 summarizes the standard effects and their components. Three results replicated in every experiment: standard Stroop interference, standard Stroop facilitation, and a concentration of naming errors on incongruent size words. Other results were specific to one experiment. Response conflict and response facilitation were significant only in Experiment 1, and in Experiment 4 RTs depended on whether the form of a size-related word resembled the correct or the opposite response. In Experiment 2, size synonyms and antonyms differed from the size word of the opposite relation but not from the size word with the same relation, a pattern compatible with a small semantic effect that could not be detected against the control conditions. Finally, several components were not detected: task conflict, semantic conflict, semantic facilitation, and the effects of orthographic and phonological similarity alone. However, their estimates were below the minimum differences the experiments could detect (about 29 to 34 ms), so these results are inconclusive rather than evidence that the components are absent.

Standard Stroop Effects with Size as the Target Dimension

The first objective was to determine whether standard Stroop interference and facilitation occur when the target dimension is the size of written words. They did. Presenting GRANDE and PEQUEÑO in the size opposite to their meaning slowed naming by 24 to 37 ms, and presenting them in the size they denote speeded it by 20 to 36 ms (Table 7). In Experiment 4, interference was measured against size-unrelated words, the

only control condition. Both effects were present against every available baseline (Table 3), and naming errors were concentrated on incongruent size words in all four experiments. These results show that the standard Stroop effect is also obtained when the target dimension is the physical size of the word, at least with Spanish size words.

In the color task, interference is usually two to three times larger than facilitation [5, 16, 32, 34, 60]. In our experiments, the two effects were of similar magnitude. This pattern is descriptive, because it was not tested formally, and it depends on the baseline. Against the standard baselines, facilitation was larger than interference in Experiment 1 and smaller in Experiment 2. Against a common baseline, size-unrelated words, interference (37 to 43 ms) exceeded facilitation (20 to 36 ms) in all experiments, but at most by a factor of about 1.8 (Experiment 2). Two features of the task may contribute to this pattern. First, the response set had only two words, both of which were also the congruent and incongruent distractors, so every congruent trial presented the exact correct response. Second, size is a relational property that participants judged against only two learned values, which may have made the congruent word a particularly effective cue.

The delta plots revealed a second difference between the two effects. In all four experiments, interference was absent in the fastest responses and grew across the RT distribution, reaching 58 to 81 ms at the .9 quantile. This increasing profile is typical of Stroop interference [61]. Facilitation, in contrast, was already present in the fastest responses (13 to 27 ms at the .1 quantile). It remained roughly constant in Experiments 1 and 2 and increased more moderately than interference in Experiments 3 and 4. Following Smith and Ulrich [57], who recommend examining the time course of interference and facilitation separately, this dissociation suggests that the benefit of a congruent size word

arises early in processing, whereas the cost of an incongruent size word builds up in slower responses.

Components of Interference and Facilitation

The second objective was to assess whether the components of interference and facilitation described in the color task are also present with size. Response conflict and response facilitation were significant in Experiment 1 (40.5 and 20.8 ms). In Experiments 2 and 4, both pointed in the same direction and ranged from 18 to 24 ms, but they were not significant and were below the minimum detectable difference. In Experiment 3, which did not include size-related words, they could not be estimated. In Experiment 4, they were estimated with the new conditions whose meaning and form pointed to the same response, but task conflict could not be estimated because symbol strings were not included. Response-level components therefore appear to contribute to the Stroop effect with size, but our data establish this firmly only for Experiment 1. Because no integrated model across experiments was fitted, we do not compare the magnitude of these components between experiments. In addition, response conflict exceeded standard interference in Experiment 1 only because symbol strings were named more slowly than size-unrelated words. This is an arithmetic consequence of the baseline, not evidence about the underlying processes.

No semantic conflict or semantic facilitation was detected with the criteria of Augustinova et al. [32]. The estimates ranged from −1.5 to 16.5 ms, and their confidence intervals are compatible with small effects in either direction that the present samples could not detect reliably. However, Experiment 2 provided indirect evidence of a small semantic effect. Size synonyms and antonyms differed significantly from the standard condition of

the opposite relation (synonyms from incongruent size words and antonyms from congruent size words), but not from the standard condition with which they shared their relation (synonyms from congruent size words and antonyms from incongruent size words), and antonyms were slower than synonyms, although not significantly. Each condition was thus drawn toward the size word with the same relation, as expected if its meaning partially activated the corresponding response. In Experiment 1, by contrast, size-related words did not differ from each other and differed from the size words of both relations, which suggests that words that merely name large or small objects activate the size concept too weakly to affect naming.

Orthographic and phonological similarity to a response word, in words without size meaning (Experiment 3), did not affect size naming. This contrasts with color studies that found such effects (e.g., [34, 58, 59]), although the differences we observed (9 to 12 ms) were well below our minimum detectable difference. Our design differed from that of Qiu and van Heuven [34] in several respects. They used Chinese homophones, which share the full pronunciation of a color word, whereas our similar words shared only part of their spelling and sound with the response words. They also estimated phonological conflict and facilitation against color-associated words of the same congruency (incongruent homophones minus incongruent associated words, and congruent associated words minus congruent homophones), whereas we compared similar words with size-unrelated words, a baseline without size meaning. In addition, in their task the word and the color were presented separately, and participants chose among three colors rather than two sizes. These differences may have contributed to the absence of detectable similarity effects in our Experiment 3.

In Experiment 4, similarity was combined with size meaning. Words that resembled the correct response were named 26 to 29 ms faster than words that resembled the opposite response, regardless of whether their meaning matched the presented size, and meaning made no detectable difference once form was held constant. This exploratory result suggests that, when a size-related distractor shares its form with a response word, the activation of that response through form outweighs the activation produced by its meaning. Testing the same conditions with color words would show whether formal overlap between distractor and response acts at the level of verbal output regardless of the target dimension. Because these conditions were tested in a single exploratory experiment, this interpretation requires replication.

Task Conflict and Lexicality

Task conflict refers to the competition between two task sets, reading the word and naming the target dimension. It arises whenever the stimulus is readable, independently of the informational conflict between the meaning of the word and the target ([62–65]; for a review, see [66]). In color naming, task conflict is usually indexed by slower responses to neutral words than to letter or symbol strings, and by a lexicality gradient in which words are slower than pseudowords and pseudowords slower than nonreadable strings [65, 67, 68]. The lexicality gradient is therefore a way of measuring task conflict, not the mechanism itself. Moreover, task conflict does not always appear in RTs. Pupillometric studies have detected it even when RTs to neutral words and nonreadable stimuli did not differ [69–71].

In Experiments 1 to 3, size-unrelated words were named numerically faster than symbol strings (by 5.5 to 15.4 ms), but none of these differences was significant after

correction, and pseudowords did not differ from either condition. Our data thus provide no evidence of task conflict when the target is size, but they do not show a reliable advantage for readable stimuli either. What they suggest is that, when participants name the size of a stimulus, readable stimuli do not produce the cost they typically produce when participants name its color. This does not mean that reading was not activated. A null or slightly negative difference in RTs is compatible with task conflict being present but offset by another factor that favors readable stimuli, as the pupillometric findings illustrate. Accordingly, the composition of standard interference described for color, as the sum of task, semantic, and response conflict, may differ when the lexical baseline is not slower than the symbol baseline, as tended to occur in Experiments 1 to 3.

Three explanations are compatible with this pattern, and future studies could separate them. The first concerns task control. Task conflict is larger when proactive control of the reading task is weak, for instance when most trials present nonreadable stimuli, and smaller when it is strong [63, 64]. In Experiments 1 to 3, five of the seven conditions presented words, a proportion that may have kept proactive control high and reduced task conflict. The second concerns the baseline. Symbol strings may not be neutral for size naming, because their shapes and density differ from those of letters and could make their size harder to judge. The third is perceptual and speculative. Words are recognized from their letters [72], and familiar letter strings might allow a faster judgment of size than unfamiliar symbols, an advantage that would speed size naming for readable stimuli and could mask the task conflict they generate. A preregistered replication with a larger sample, the same controls in every experiment, normed stimuli, a manipulation of the proportion of word trials, and pupillometric measures could distinguish among these explanations.

Theoretical Accounts

The multistage account of the Stroop effect [16, 20, 32] predicts that standard interference is the sum of task, semantic, and response conflict, and that standard facilitation is the sum of semantic and response facilitation. Our results are consistent with this account in one respect. With size as the target, part of the standard effects occurred at the response level: response conflict and response facilitation were significant in Experiment 1, and pointed in the same direction in Experiments 2 and 4. The semantic components were not detected, and task conflict produced no cost, so the remaining predictions could not be confirmed. The decomposition of Stroop effects into components has already been applied outside the color domain, for example in picture–word interference [35], spatial Stroop tasks [36], and Stroop-matching tasks [37]. The present study extends this approach to the physical size of written words.

An alternative account is the diffusion model for conflict tasks (DMC) [73]. In this model, the irrelevant dimension produces an automatic activation that rises and then decays over time and is superimposed on the controlled processing of the relevant dimension. DMC explains interference and facilitation with a single mechanism and predicts the shape of delta functions: they increase with RT when the automatic activation peaks late or decays slowly, as is typically assumed for the Stroop task. The increasing interference delta plots observed in all four experiments fit this pattern. The facilitation delta plots, which were already positive in the fastest responses and flatter than those of interference, are less easily explained by a single automatic activation with the same time course for congruent and incongruent trials. They may require additional assumptions, such as a faster or stronger contribution of congruent size words, which coincide exactly with the correct response (see

902 [57]). Fitting DMC to these data was beyond the scope of this exploratory study, but it
903 would allow testing whether the Stroop effect with size words is driven by the same
904 automatic process as the Stroop effect with color.

905 **Limitations**

906 Several limitations qualify these conclusions. First, sample sizes were not
907 determined a priori. The sensitivity analysis showed that each experiment could detect, with
908 80% power, differences of about 29 to 34 ms, so the components we did not detect (task,
909 semantic, and phonological effects of 3 to 16 ms) remain inconclusive. Second, the study
910 was exploratory and not preregistered; each experiment was designed after the results of the
911 previous one, and all pairwise comparisons were tested with a correction for multiplicity
912 rather than a small set of confirmatory contrasts. Third, participants came from a single
913 private university in Bogotá, mostly psychology students. Fourth, the stimuli were not
914 selected from normed databases. We controlled their orthographic similarity to the response
915 words (Table 1), but not their frequency, familiarity, length, or number of syllables, which
916 may have contributed to differences between conditions. Fifth, each condition included
917 only 24 trials per participant, which limits the precision of individual estimates and of the
918 distributional analyses. Sixth, the control conditions changed across experiments: the size-
919 unrelated words were replaced in Experiment 3, and Experiment 4 had no symbol strings or
920 pseudowords. Because no integrated model across experiments was fitted, comparisons of
921 effect magnitudes between experiments are descriptive.

922 A further limitation concerns the response set. Size is a relational property:
923 something is large or small relative to something else, and participants had to discriminate
924 only two sizes, which they could learn within the first trials. Unlike color words, which

name relatively discrete categories, "large" and "small" also apply to continuous dimensions such as volume, speed, or duration. A larger response set, with intermediate sizes and terms such as "medium," would make the task more similar to color naming but would also increase decision difficulty, RTs, and errors. In color naming, response set size affects overall RTs but not the presence of Stroop interference [27, 30, 74, 75]. Whether this holds for size, and whether the two-word response set contributed to the absence of task conflict or to the similar magnitudes of interference and facilitation, remains to be tested. Finally, response onsets were detected automatically from the recordings and transcriptions were reviewed manually. These procedures were identical across conditions, but they may have added measurement noise.

Conclusions

Naming the physical size of written words produces a Stroop effect. Across four experiments, Spanish size words presented in the size opposite to their meaning slowed size naming, and size words presented in the size they denote speeded it, against every available baseline. Errors were concentrated on incongruent size words. Interference was absent in the fastest responses and grew with RT, whereas facilitation was present from the fastest responses. Response-level components contributed to these effects, at least in Experiment 1. Semantic components, task conflict, and effects of orthographic and phonological similarity alone were not detected, but the experiments were not sensitive enough to rule out effects smaller than about 30 ms. Readable stimuli did not produce the cost that characterizes task conflict in color naming, and in Experiment 4 the form of size-related words outweighed their meaning. These two exploratory findings suggest that the

composition of Stroop effects may depend on the target dimension, and they should be tested in preregistered studies with larger samples.

Abbreviations

AIC: Akaike information criterion; ANOVA: analysis of variance; BIC: Bayesian information criterion; CI: confidence interval; DMC: diffusion model for conflict tasks; GLMM: generalized linear mixed model; OR: odds ratio; OSF: Open Science Framework; RT: reaction time; SD: standard deviation; SE: standard error.

Declarations

Ethics approval and consent to participate

The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Fundación Universitaria Konrad Lorenz (Minutes No. 31, October 26, 2023). All participants gave written informed consent before taking part. For participants under 18 years of age, written consent was also obtained from a parent or legal guardian.

Consent for publication

Not applicable.

Availability of data and materials

The trial-level data and the analysis scripts for the four experiments are available in the Open Science Framework repository, https://doi.org/10.17605/OSF.IO/J4ZUV (CC-BY 4.0 license). The audio recordings are not publicly available because they contain participants' voices. None of the experiments was preregistered.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by Fundación Universitaria Konrad Lorenz (Bogotá, Colombia) through the research project “Study of the Symbolic Properties of Words Through the Stroop Effect” (9INV2241), awarded to EAP and JCFV. The funder had no role in the design of the study, the analysis of the data, or the writing of the manuscript.

Authors’ contributions

EAP and JCFV designed the study. JCFV designed the data collection program. EAP supervised data collection. JAVM performed the orthographic similarity measurement between stimuli. JCFV and JAVM performed the statistical data analysis. EAP and JCFV wrote the manuscript. JAVM proofread the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Diego Andrés Sichaca for assistance with data collection.

Supplementary Information

Additional file 1 (.docx). Supplementary tables. Table S1: comparison of random-effects structures. Tables S2–S5: estimated fixed parameters of the models of Experiments 1–4. Table S6: residual diagnostics of the final models. Table S7: repeated-measures ANOVA on participant mean RTs. Table S8: participant mean RTs by condition.

References

1. MacLeod CM. Half a century of research on the Stroop effect: An integrative review. Psychol Bull. 1991;109(2):163–203. https://doi.org/10.1037/0033-2909.109.2.163

2. MacLeod CM. The Stroop task in cognitive research. In: Wenzel A, Rubin DC, editors. Cognitive methods and their application to clinical research. Washington, DC: American Psychological Association; 2005. p. 17–40. https://doi.org/10.1037/10870-002

3. Stroop JR. Studies of interference in serial verbal reactions. J Exp Psychol. 1935;18(6):643–62. https://doi.org/10.1037/h0054651

4. Stroop JR. Factors affecting speed in serial verbal reactions. Psychol Monogr. 1938;50(5):38–48. https://doi.org/10.1037/h0093516

5. Dalrymple-Alford EC. Associative facilitation and interference in the Stroop color-word task. Percept Psychophys. 1972;11(4):274–6. https://doi.org/10.3758/BF03210377

6. Hilbert S, Nakagawa TT, Bindl M, Bühner M. The spatial Stroop effect: A comparison of color-word and position-word interference. Psychon Bull Rev. 2014;21(6):1509–15. https://doi.org/10.3758/s13423-014-0631-4

7. Palef SR. Judging pictorial and linguistic aspects of space. Mem Cognit. 1978;6(1):70–5. https://doi.org/10.3758/BF03197430

8. Warren RE, Lasher MD. Interference in a typeface variant of the Stroop test. Percept Psychophys. 1974;15(1):128–30. https://doi.org/10.3758/BF03205840

9. Stirling N, Coltheart M. Stroop interference in a letter naming task. Bull Psychon Soc. 1977;10(1):31–4. https://doi.org/10.3758/BF03333538

10. Green EJ, Barber PJ. Interference effects in an auditory stroop task: Congruence and correspondence. Acta Psychol (Amst). 1983;53(3):183–94. https://doi.org/10.1016/0001-6918(83)90001-X

11. Hamers JF, Lambert WE. Bilingual interdependencies in auditory perception. J Verbal Learn Verbal Behav. 1972;11(3):303–10. https://doi.org/10.1016/S0022-5371(72)80091-4

12. Dupuis A, Berent I. Signs are symbols: evidence from the Stroop task. Lang Cogn Neurosci. 2015;30(10):1339–44. https://doi.org/10.1080/23273798.2015.1069363

13. Grégoire L, Poulin-Charronnat B, Perruchet P. Stroop interference depends also on the level of automaticity of the to-be-interfered process. Acta Psychol (Amst). 2019;197:143–52. https://doi.org/10.1016/j.actpsy.2019.05.013

14. Zakay D, Glicksohn J. Stimulus congruity and S-R compatibility as determinants of interference in a Stroop-like task. Can J Psychol. 1985;39(3):414–23. https://doi.org/10.1037/h0080069

15. Klein GS. Semantic power measured through the interference of words with color-naming. Am J Psychol. 1964;77(4):576–88. https://doi.org/10.2307/1420768

16. Augustinova M, Silvert L, Spatola N, Ferrand L. Further investigation of distinct components of Stroop interference and of their reduction by short response-stimulus intervals. Acta Psychol (Amst). 2018;189:54–62. https://doi.org/10.1016/j.actpsy.2017.03.009

17. Glaser MO, Glaser WR. Time course analysis of the Stroop phenomenon. J Exp Psychol Hum Percept Perform. 1982;8(6):875–94. https://doi.org/10.1037/0096-1523.8.6.875

18. Hasshim N, Downes M, Bate S, Parris BA. Response Time Distribution Analysis of Semantic and Response Interference in a Manual Response Stroop Task. Exp Psychol. 2019;66(3):231–8. https://doi.org/10.1027/1618-3169/a000445

19. Müller VI, Cieslik EC, Ficco L, Tyralla S, Sepehry AA, Aziz-Safaie T, Feng C, Eickhoff SB, Langner R. Not all Stroop-type tasks are alike: Assessing the impact of stimulus material, task design, and cognitive demand via meta-analyses across neuroimaging studies. Neuropsychol Rev. 2025;35(3):449–82. https://doi.org/10.1007/s11065-024-09647-1

20. Parris BA, Hasshim N, Wadsley M, Augustinova M, Ferrand L. The loci of Stroop effects: a critical review of methods and evidence for levels of processing contributing to color-word Stroop effects and the implications for the loci of

attentional selection. Psychol Res. 2022;86(4):1029–53. https://doi.org/10.1007/s00426-021-01554-x

21. Dimoska-Di Marco A, McDonald S, Kelly M, Tate R, Johnstone S. A meta-analysis of response inhibition and Stroop interference control deficits in adults with traumatic brain injury (TBI). J Clin Exp Neuropsychol. 2011;33(4):471–85. https://doi.org/10.1080/13803395.2010.533158

22. Ferrand L, Ducrot S, Chausse P, Maïonchi-Pino N, O'Connor RJ, Parris BA, Perret P, Riggs KJ, Augustinova M. Stroop interference is a composite phenomenon: Evidence from distinct developmental trajectories of its components. Dev Sci. 2020;23(2):e12899. https://doi.org/10.1111/desc.12899

23. Mills L, Kinoshita S, Norris D. No negative priming effect in the manual Stroop task. Front Psychol. 2019;10:1764. https://doi.org/10.3389/fpsyg.2019.01764

24. Scarpina F, Tagini S. The Stroop Color and Word Test. Front Psychol. 2017;8:557. https://doi.org/10.3389/fpsyg.2017.00557

25. Algom D, Chajut E. Reclaiming the Stroop Effect Back From Control to Input-Driven Attention and Perception. Front Psychol. 2019;10:1683. https://doi.org/10.3389/fpsyg.2019.01683

26. MacLeod CM. John Ridley Stroop: Creator of a landmark cognitive task. Can Psychol. 1991;32(3):521–4. https://doi.org/10.1037/h0079012

1065 27. Sabri M, Melara RD, Algom D. A confluence of contexts: Asymmetric versus global
1066 failures of selective attention to Stroop dimensions. J Exp Psychol Hum Percept
1067 Perform. 2001;27(3):515–37. https://doi.org/10.1037/0096-1523.27.3.515

1068 28. Treisman A, Fearnley S. The Stroop Test: Selective Attention to Colours and Words.
1069 Nature. 1969;222(5192):437–9. https://doi.org/10.1038/222437a0

1070 29. Blais C, Besner D. When the visual format of the color carrier word does and does
1071 not modulate the Stroop effect. Mem Cognit. 2005;33(8):1337–44.
1072 https://doi.org/10.3758/BF03193366

1073 30. Kanne SM, Balota DA, Spieler DH, Faust ME. Explorations of Cohen, Dunbar, and
1074 McClelland's (1990) connectionist model of Stroop performance. Psychol Rev.
1075 1998;105(1):174–87. https://doi.org/10.1037/0033-295X.105.1.174

1076 31. Kinoshita S, Mills L, Norris D. The semantic Stroop effect is controlled by
1077 endogenous attention. J Exp Psychol Learn Mem Cogn. 2018;44(11):1730–42.
1078 https://doi.org/10.1037/xlm0000552

1079 32. Augustinova M, Parris BA, Ferrand L. The Loci of Stroop Interference and
1080 Facilitation Effects With Manual and Vocal Responses. Front Psychol.
1081 2019;10:1786. https://doi.org/10.3389/fpsyg.2019.01786

1082 33. Martinon LM, Ferrand L, Burca M, Hasshim N, Lakhzoum D, Parris BA, Silvert L,
1083 Augustinova M. Distributional analyses reveal the polymorphic nature of the
1084 Stroop interference effect: It's about (response) time. Mem Cognit.
1085 2024;52(6):1229–45. https://doi.org/10.3758/s13421-024-01538-3

34. Qiu Y, van Heuven WJ. Distinct components of Stroop interference and facilitation: The role of phonology and response modality. Q J Exp Psychol (Hove). 2025;78(5):997–1015. https://doi.org/10.1177/17470218241302490

35. Starreveld PA, La Heij W. Picture-word interference is a Stroop effect: A theoretical analysis and new empirical findings. Psychon Bull Rev. 2017;24(3):721–33. https://doi.org/10.3758/s13423-016-1167-6

36. Spinelli G, Lupker SJ. A spatial version of the Stroop task for examining proactive and reactive control independently from non-conflict processes. Atten Percept Psychophys. 2024;86(4):1259–86. https://doi.org/10.3758/s13414-024-02892-9

37. Machado-Pinheiro W, dos Santos Afonso A Jr, Caldas AL, de Almeida Portugal AC, Mocaiber I, David IA, Carreiro LRR. Disentangling three different conflicts at non-response/semantic and response levels in a new Stroop-matching task. Learn Motiv. 2024;85:101955. https://doi.org/10.1016/j.lmot.2023.101955

38. Brown TL. The relationship between stroop interference and facilitation effects: Statistical artifacts, baselines, and a reassessment. J Exp Psychol Hum Percept Perform. 2011;37(1):85–99. https://doi.org/10.1037/a0019252

39. Smith P, Ulrich R. The neutral condition in conflict tasks: On the violation of the midpoint assumption in reaction time trends. Q J Exp Psychol (Hove). 2024;77(5):1023–43. https://doi.org/10.1177/17470218231201476

40. Burca M, Beaucousin V, Chausse P, Ferrand L, Parris BA, Augustinova M. Is There Semantic Conflict in the Stroop Task? Exp Psychol. 2021;68(5):274–83. https://doi.org/10.1027/1618-3169/a000530

41. Kumle L, Võ MLH, Draschkow D. Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. Behav Res Methods. 2021;53(6):2528–43. https://doi.org/10.3758/s13428-021-01546-0

42. Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK. PsychoPy2: Experiments in behavior made easy. Behav Res Methods. 2019;51(1):195–203. https://doi.org/10.3758/s13428-018-01193-y

43. Galer S, Schmitz R, Leproult R, De Tiège X, Van Bogaert P, Peigneux P. Response-stimulus interval duration modulates interference effects in the Stroop task. Psychol Belg. 2014;54(1):97–110. https://doi.org/10.5334/pb.ad

44. Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, Skaug HJ, Mächler M, Bolker BM. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. R J. 2017;9(2):378–400. https://doi.org/10.32614/RJ-2017-066

45. Jadoul Y, Thompson B, de Boer B. Introducing Parselmouth: A Python interface to Praat. J Phon. 2018;71:1–15. https://doi.org/10.1016/j.wocn.2018.07.001

46. Bain M, Huh J, Han T, Zisserman A. WhisperX: time-accurate speech transcription of long-form audio. In: Proceedings of Interspeech 2023. 2023. p. 4489–93. https://doi.org/10.21437/Interspeech.2023-78

47. Lo S, Andrews S. To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. Front Psychol. 2015;6:1171. https://doi.org/10.3389/fpsyg.2015.01171

48. R Core Team. R: a language and environment for statistical computing. Version 4.5.0. Vienna: R Foundation for Statistical Computing; 2025. https://www.R-project.org/. Accessed 11 Oct 2026.

49. Fox J, Weisberg S. An R companion to applied regression. 3rd ed. Thousand Oaks: Sage; 2019.

50. Barr DJ, Levy R, Scheepers C, Tily HJ. Random effects structure for confirmatory hypothesis testing: Keep it maximal. J Mem Lang. 2013;68(3):255–78. https://doi.org/10.1016/j.jml.2012.11.001

51. Matuschek H, Kliegl R, Vasishth S, Baayen H, Bates D. Balancing Type I error and power in linear mixed models. J Mem Lang. 2017;94:305–15. https://doi.org/10.1016/j.jml.2017.01.001

52. Hartig F. DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. R package version 0.4.7. 2024. https://github.com/florianhartig/DHARMa. Accessed 11 Oct 2026.

53. Lenth RV, Piaskowski J. emmeans: estimated marginal means, aka least-squares means. R package. 2017. https://doi.org/10.32614/CRAN.package.emmeans

54. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65–70.

55. De Jong R, Liang CC, Lauber E. Conditional and unconditional automaticity: A dual-process model of effects of spatial stimulus–response correspondence. J Exp Psychol Hum Percept Perform. 1994;20(4):731–50. https://doi.org/10.1037/0096-1523.20.4.731

56. Ratcliff R. Group reaction time distributions and an analysis of distribution statistics. Psychol Bull. 1979;86(3):446–61. https://doi.org/10.1037/0033-2909.86.3.446

57. Smith P, Ulrich R. Decomposing delta plots: Exploring the time course of the congruency effect using inhibition and facilitation curves. Psychol Res. 2025;89(1):52. https://doi.org/10.1007/s00426-024-02075-z

58. Dalrymple-Alford EC. Sound similarity and color-word interference in the Stroop task. Psychon Sci. 1972;28(4):209–10. https://doi.org/10.3758/BF03328712

59. Marmurek HHC, Proctor C, Javor A. Stroop-Like Serial Position Effects in Color Naming of Words and Nonwords. Exp Psychol. 2006;53(2):105–10. https://doi.org/10.1027/1618-3169.53.2.105

60. Quétard B, Spatola N, Parris BA, Ferrand L, Augustinova M. A mouse-tracking study of the composite nature of the Stroop effect at the level of response execution. PLoS One. 2023;18(1):e0279036. https://doi.org/10.1371/journal.pone.0279036

61. Pratte MS, Rouder JN, Morey RD, Feng C. Exploring the differences in distributional properties between Stroop and Simon effects using delta plots. Atten Percept Psychophys. 2010;72(7):2013–25. https://doi.org/10.3758/APP.72.7.2013

62. Entel O, Tzelgov J, Bereby-Meyer Y, Shahar N. Exploring relations between task conflict and informational conflict in the Stroop task. Psychol Res. 2015;79(6):913–27. https://doi.org/10.1007/s00426-014-0630-0

63. Goldfarb L, Henik A. Evidence for task conflict in the Stroop effect. J Exp Psychol Hum Percept Perform. 2007;33(5):1170–6. https://doi.org/10.1037/0096-1523.33.5.1170

64. Kalanthroff E, Davelaar EJ, Henik A, Goldfarb L, Usher M. Task conflict and proactive control: A computational theory of the Stroop task. Psychol Rev. 2018;125(1):59–82. https://doi.org/10.1037/rev0000083

65. Monsell S, Taylor TJ, Murphy K. Naming the color of a word: Is it responses or task sets that compete? Mem Cognit. 2001;29(1):137–51. https://doi.org/10.3758/BF03195748

66. Littman R, Keha E, Kalanthroff E. Task conflict and task control: A mini-review. Front Psychol. 2019;10:1598. https://doi.org/10.3389/fpsyg.2019.01598

67. Kinoshita S, De Wit B, Norris D. The magic of words reconsidered: Investigating the automaticity of reading color-neutral words in the Stroop task. J Exp Psychol Learn Mem Cogn. 2017;43(3):369–84. https://doi.org/10.1037/xlm0000311

68. Levin Y, Tzelgov J. What Klein's "Semantic Gradient" Does and Does Not Really Show: Decomposing Stroop Interference into Task and Informational Conflict Components. Front Psychol. 2016;7:249. https://doi.org/10.3389/fpsyg.2016.00249

69. Hershman R, Henik A. Dissociation between reaction time and pupil dilation in the Stroop task. J Exp Psychol Learn Mem Cogn. 2019;45(10):1899–909. https://doi.org/10.1037/xlm0000690

70. Hershman R, Henik A. Pupillometric contributions to deciphering Stroop conflicts. Mem Cognit. 2020;48(2):325–33. https://doi.org/10.3758/s13421-019-00971-z

71. Hershman R, Levin Y, Tzelgov J, Henik A. The contribution of meaning to the detection of task conflict. Q J Exp Psychol (Hove). 2021;74(9):1553–61. https://doi.org/10.1177/17470218211001331

72. Pelli DG, Farell B, Moore DC. The remarkable inefficiency of word recognition. Nature. 2003;423(6941):752–6. https://doi.org/10.1038/nature01516

73. Ulrich R, Schröter H, Leuthold H, Birngruber T. Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions. Cogn Psychol. 2015;78:148–74. https://doi.org/10.1016/j.cogpsych.2015.02.005

74. Lemercier C, Simoës-Perlant A, Schmidt JR, Boujon C. Stroop interference and development: Influence of expectation on color-naming response times. Eur Rev Appl Psychol. 2017;67(1):43–50. https://doi.org/10.1016/j.erap.2016.09.001

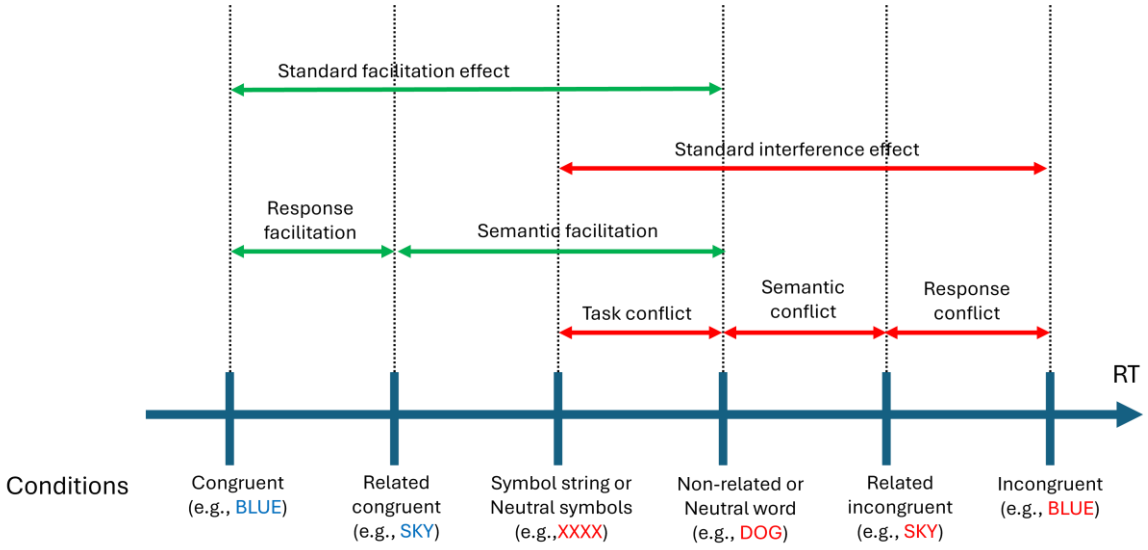
75. McClain L. Effects of Response Type and Set Size on Stroop Color-Word Performance. Percept Mot Skills. 1983;56(3):735–43. https://doi.org/10.2466/pms.1983.56.3.735

76. Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. J Am Stat Assoc. 1989;84(406):414–20. https://doi.org/10.1080/01621459.1989.10478785

77. Winkler WE. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In: Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association; 1990. p. 354–9.

Figure 1

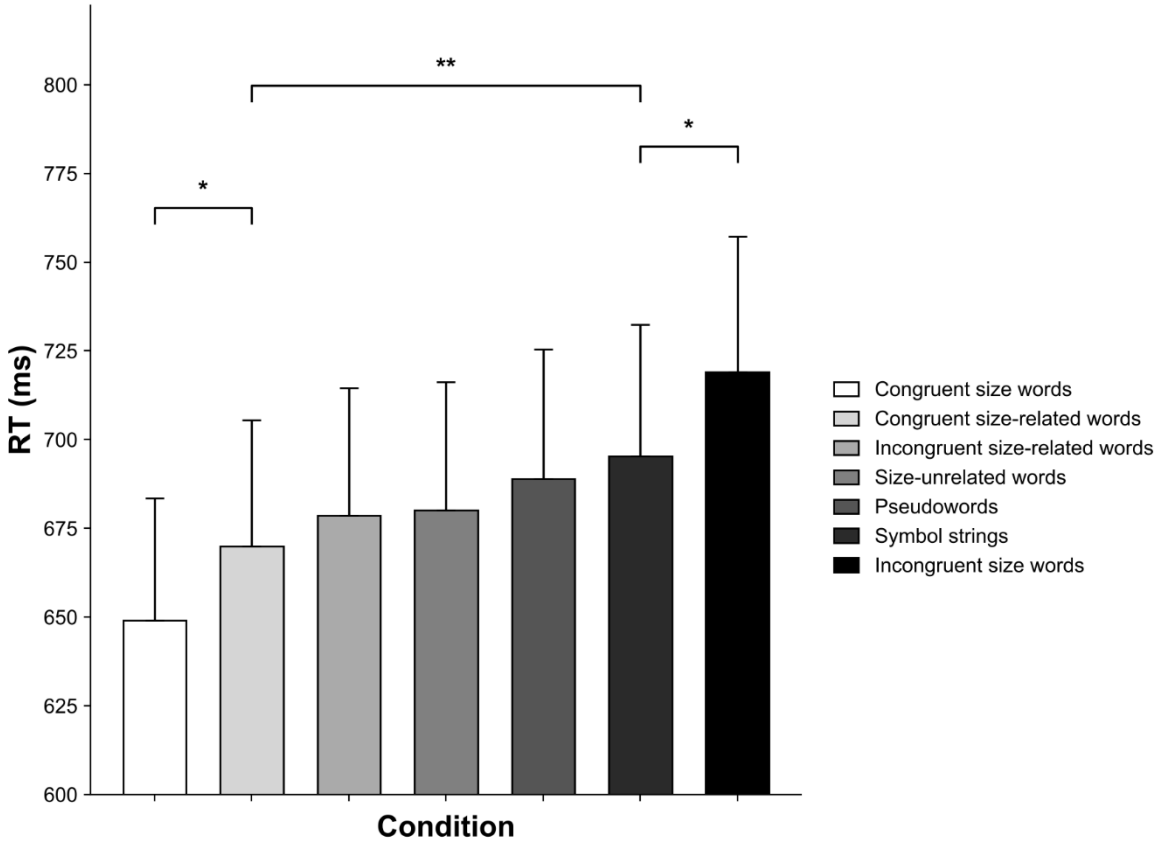
Components of interference and facilitation Stroop effects



Note: This figure is adapted from Augustinova et al. [16], Parris et al. [20], and Qiu and van Heuven [34]. Standard Stroop interference and its components are in red, while facilitation and its components are in green. The conditions are ordered according to their typical relative RT in color Stroop task research.

Figure 2

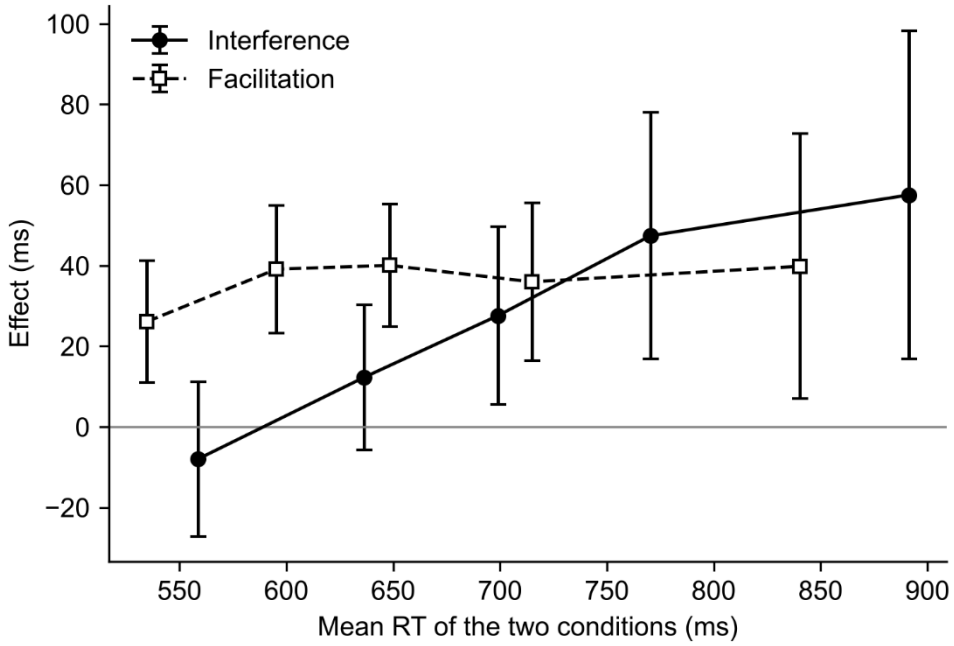
Mean RT by condition in Experiment 1



Note. Bars show estimated marginal means, ordered from fastest to slowest, and error bars the upper limit of their 95% confidence intervals. Each bracket joins two conditions that differ significantly; each of these conditions also differs from every condition beyond the other end of the bracket. Together, the brackets show all the significant pairwise differences (Holm-corrected across 21 comparisons). Asterisks give the significance of the bracketed pair: * $p < .05$, ** $p < .01$, *** $p < .001$.

Figure 3

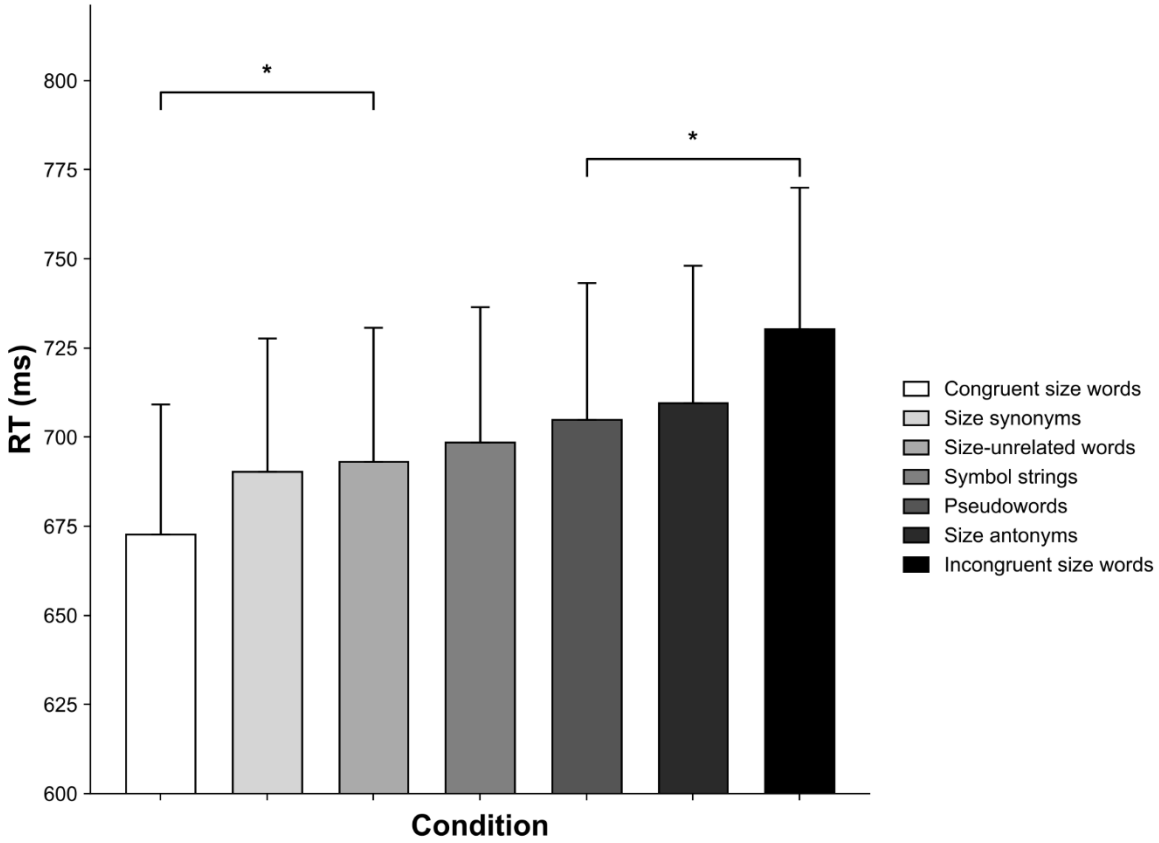
Delta Plots of Standard Interference and Facilitation in Experiment 1



Note. Effects at the .1, .3, .5, .7, and .9 quantiles of the RT distribution, plotted against the mean RT of the two conditions at each quantile. Interference = incongruent size words minus symbol strings; facilitation = size-unrelated words minus congruent size words. Error bars are 95% confidence intervals across participants.

Figure 4

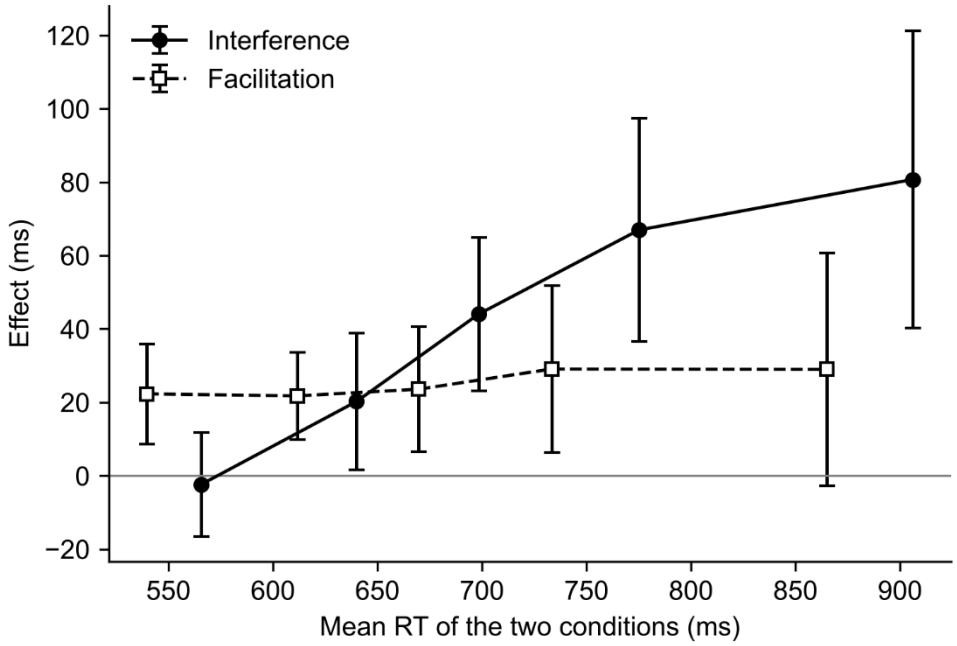
Mean RT by condition in Experiment 2



Note. Bars show estimated marginal means, ordered from fastest to slowest, and error bars the upper limit of their 95% confidence intervals. Each bracket joins two conditions that differ significantly; each of these conditions also differs from every condition beyond the other end of the bracket. Together, the brackets show all the significant pairwise differences (Holm-corrected across 21 comparisons). Asterisks give the significance of the bracketed pair: * $p < .05$, ** $p < .01$, *** $p < .001$.

Figure 5

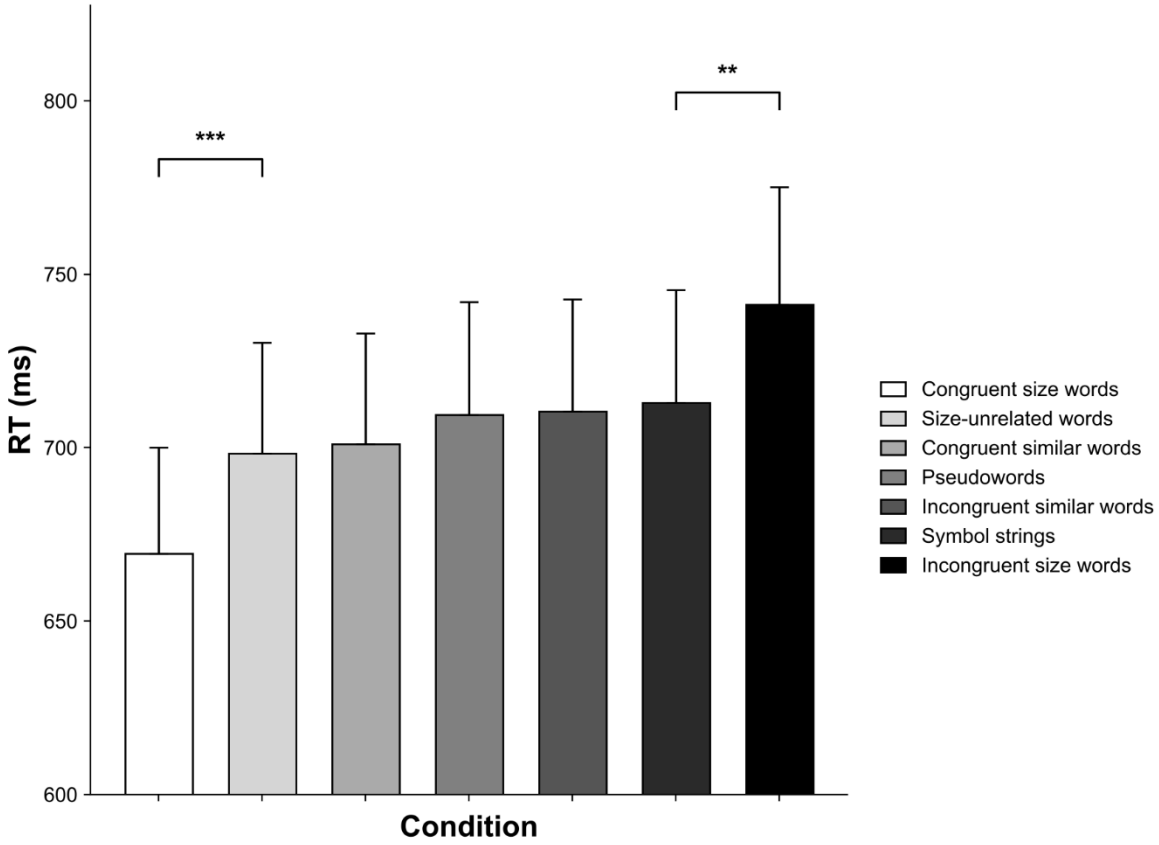
Delta Plots of Standard Interference and Facilitation in Experiment 2



Note. Effects at the .1, .3, .5, .7, and .9 quantiles of the RT distribution, plotted against the mean RT of the two conditions at each quantile. Interference = incongruent size words minus symbol strings; facilitation = size-unrelated words minus congruent size words. Error bars are 95% confidence intervals across participants.

Figure 6

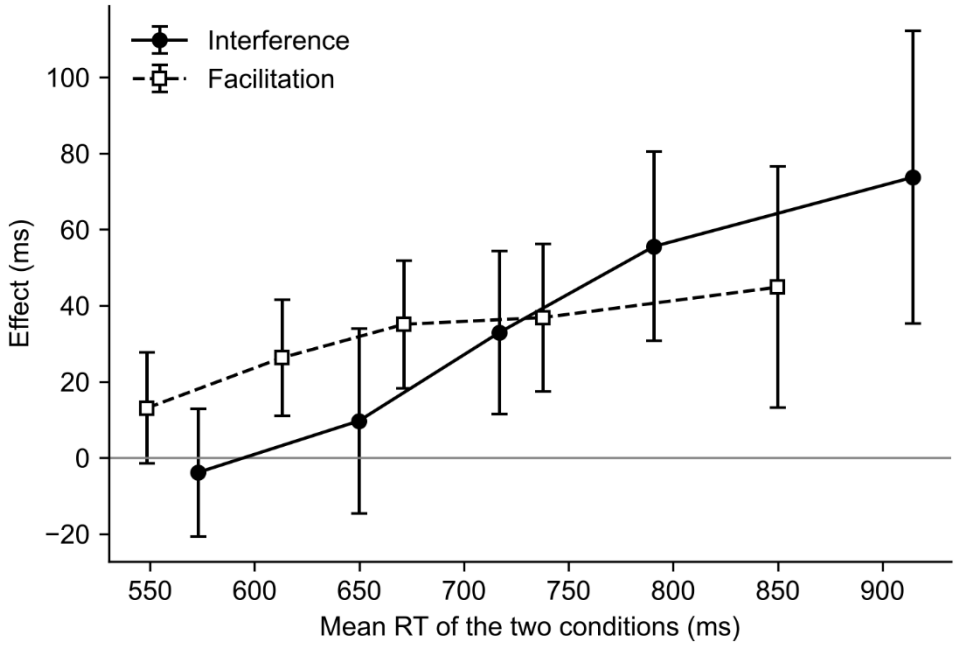
Mean RT by condition in Experiment 3



Note. Bars show estimated marginal means, ordered from fastest to slowest, and error bars the upper limit of their 95% confidence intervals. Each bracket joins two conditions that differ significantly; each of these conditions also differs from every condition beyond the other end of the bracket. Together, the brackets show all the significant pairwise differences (Holm-corrected across 21 comparisons). Asterisks give the significance of the bracketed pair: * $p < .05$, ** $p < .01$, *** $p < .001$.

Figure 7

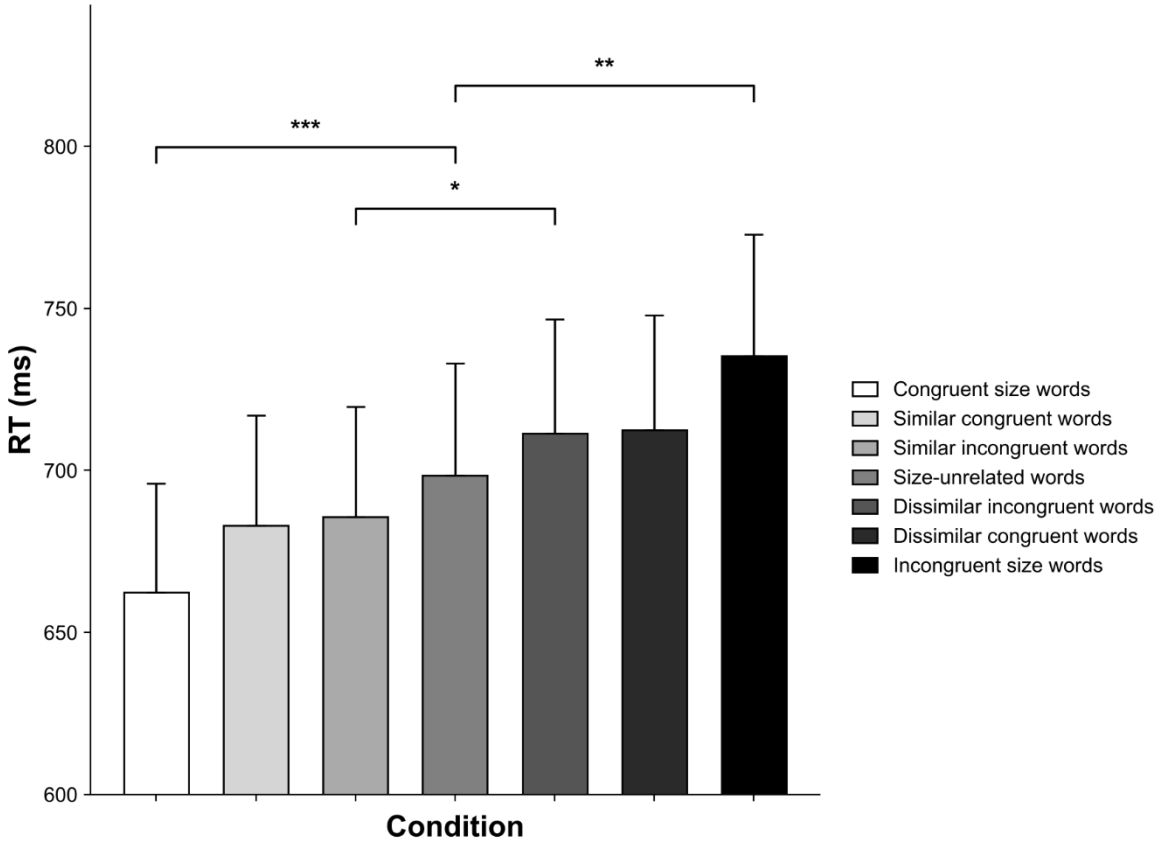
Delta Plots of Standard Interference and Facilitation in Experiment 3



Note. Effects at the .1, .3, .5, .7, and .9 quantiles of the RT distribution, plotted against the mean RT of the two conditions at each quantile. Interference = incongruent size words minus symbol strings; facilitation = size-unrelated words minus congruent size words. Error bars are 95% confidence intervals across participants.

Figure 8

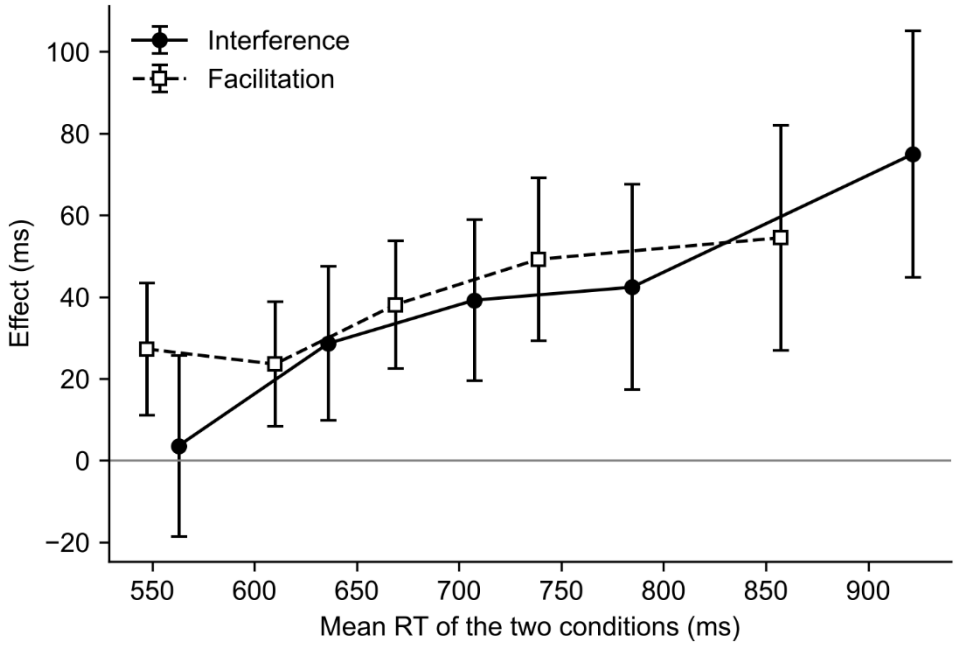
Mean RT by condition in Experiment 4



Note. Bars show estimated marginal means, ordered from fastest to slowest, and error bars the upper limit of their 95% confidence intervals. Each bracket joins two conditions that differ significantly; each of these conditions also differs from every condition beyond the other end of the bracket. Together, the brackets show all the significant pairwise differences (Holm-corrected across 21 comparisons). Asterisks give the significance of the bracketed pair: * $p < .05$, ** $p < .01$, *** $p < .001$.

Figure 9

Delta Plots of Standard Interference and Facilitation in Experiment 4



Note. Effects at the .1, .3, .5, .7, and .9 quantiles of the RT distribution, plotted against the mean RT of the two conditions at each quantile. Interference = incongruent size words minus size-unrelated words; facilitation = size-unrelated words minus congruent size words. Error bars are 95% confidence intervals across participants.

Table 1

Stimuli by Experiment and Condition

Condition	Experiments	Presented large (100 px)	Presented small (20 px)	Presentations
Congruent size words	1–4	GRANDE (large) [.00, .56]	PEQUEÑO (small) [.56, .00]	12
Incongruent size words	1–4	PEQUEÑO [.56, .00]	GRANDE [.00, .56]	12

Condition	Experiments	Presented large (100 px)	Presented small (20 px)	Presentations
Symbol strings	1–3	#####, &&&&&, ?????, €€€€€ [1.00, 1.00 each]	#####, &&&&&, ?????, €€€€€ [1.00, 1.00 each]	3 per size
Pseudowords	1–3	SOLIMU [1.00, .56], LURATEMA [.38, .49], VESLARIO [.57, .49], FEBAMINO [.47, .49]	SOLIMU, LURATEMA, VESLARIO, FEBAMINO	3 per size
Size-unrelated words	1, 2	CONOCER (to know), FORMAL (formal), IMAGINAR (to imagine), ÁMBITO (scope)	CONOCER [.46, .57], FORMAL [.44, 1.00], IMAGINAR [.49, 1.00], ÁMBITO [1.00, .56]	3 per size
Size-unrelated words	3, 4	NÚMERO (number), CALIDAD (quality), ESTILO (style), RELATO (story)	NÚMERO [.56, .46], CALIDAD [.46, 1.00], ESTILO [1.00, .46], RELATO [.44, .46]	3 per size
Congruent size-related words	1	MONTAÑA (mountain) [.63, .57], BALLENA (whale) [.47, .57]	HORMIGA (ant) [.56, 1.00], SEMILLA (seed) [1.00, .57]	6
Incongruent size-related words	1	HORMIGA, SEMILLA	MONTAÑA, BALLENA	6
Size synonyms	2	ENORME (enormous) [.44, .46], GIGANTE (giant) [.23, .57]	MINÚSCULO (minuscule) [.57, .50], DIMINUTO (tiny) [.57, .49]	6
Size antonyms	2	MINÚSCULO, DIMINUTO	ENORME, GIGANTE	6
Congruent similar words	3	GRACIAS (thank you) [.25, 1.00], GRANUJA (rascal) [.15, .57]	PAQUETE (package) [.46, .26], PECADO (sin) [.44, .29]	6
Incongruent similar words	3	PAQUETE, PECADO	GRACIAS, GRANUJA	6
Similar congruent words	4	GRUESO (thick) [.27, .47], GRANJA (farm) [.13, 1.00]	PICADURA (sting) [.47, .44], POQUITO (a little) [1.00, .26]	6
Similar incongruent words	4	GRAMO (gram) [.21, .55], GRANADA (pomegranate) [.09, 1.00]	PAQUIDERMO (pachyderm) [.40, .24], PAQUETÓN (big package) [.47, .39]	6

Condition	Experiments	Presented large (100 px)	Presented small (20 px)	Presentations
Dissimilar congruent words	4	PAQUIDERMO, PAQUETÓN	GRAMO, GRANADA	6
Dissimilar incongruent words	4	PICADURA, POQUITO	GRUESO, GRANJA	6

Note. Stimuli were presented in black uppercase Arial on a white background. Each condition comprised 24 trials per participant (eight per block, three blocks). English translations are given in parentheses at first mention. In Experiment 3, similar words were orthographically and phonologically similar to GRANDE or PEQUEÑO but unrelated to size; they were congruent when similar to the correct response and incongruent when similar to the opposite response. In Experiment 4, similar and dissimilar refer to orthographic and phonological similarity to the correct response, and congruent and incongruent refer to whether the size meaning of the word matched the presented size. Values in brackets are the Jaro-Winkler distances between each stimulus and the two response words, GRANDE and PEQUEÑO, respectively. The Jaro-Winkler index [76, 77] measures orthographic similarity from the number of letters that two strings share in approximately the same positions and the number of transpositions among those letters, and it gives additional weight to strings that begin with the same letters (up to four initial letters, prefix scale p = .10). Distance is 1 minus similarity: 0 indicates identical strings and 1 indicates strings with no letters in common.

Table 2

Frequency of response types by condition in Experiment 1

	Type of response			
Conditions	No-response	Correct	Incorrect	Other

Symbol strings	9	864	4	6
Pseudowords	13	862	5	3
Size-unrelated words	18	850	5	5
Congruent size-related words	17	848	7	7
Incongruent size-related words	23	838	10	6
Congruent size words	10	862	2	5
Incongruent size words	18	816	42	7

Table 3

Standard Interference and Facilitation Estimated Against Each Baseline

	Interference vs. symbol strings	Interference vs. pseudowords	Interference vs. unrelated words	Facilitation vs. unrelated words	Facilitation vs. pseudowords	Facilitation vs. symbol strings
Experiment 1	23.6* [8.3, 38.9]	30.1* [14.8, 45.3]	39.0* [23.7, 54.2]	31.0* [16.7, 45.3]	39.9* [25.5, 54.3]	46.4* [31.9, 60.9]
Experiment 2	31.8* [17.1, 46.4]	25.3* [10.6, 40.1]	37.2* [22.6, 51.9]	20.4* [6.7, 34.1]	32.3* [18.3, 46.3]	25.9* [12.0, 39.7]
Experiment 3	28.3* [13.5, 43.2]	31.7* [16.9, 46.6]	42.8* [28.0, 57.6]	28.9* [15.0, 42.7]	40.0* [26.0, 54.0]	43.4* [29.3, 57.5]
Experiment 4	n.a.	n.a.	37.0* [18.0, 55.9]	36.1* [18.4, 53.8]	n.a.	n.a.

Note. Values are differences in estimated marginal means (ms) from the final GLMM of each experiment, with 95% confidence intervals in brackets. Interference = incongruent size words minus baseline; facilitation = baseline minus congruent size words. n.a. = condition not included in the experiment. * $p < .05$, Holm-corrected across the 21 pairwise comparisons of each experiment.

Table 4

Frequency of response types by condition in Experiment 2

	Type of response			
Conditions	No-response	Correct	Incorrect	Other
Symbol strings	24	835	2	16
Pseudowords	29	831	2	14
Size-unrelated words	37	840	1	12
Size-synonyms words	24	848	0	16
Size-antonyms words	35	818	4	11
Congruent size words	23	850	1	8
Incongruent size words	34	803	29	15

Table 5

Frequency of response types by condition in Experiment 3

	Type of response			
Conditions	No-response	Correct	Incorrect	Other
Symbol strings	25	847	1	8
Pseudowords	16	853	1	16
Size-unrelated words	20	851	1	12
Congruent similar words	15	850	1	12

Incongruent similar words	26	829	1	17
Congruent size words	11	857	2	13
Incongruent size words	26	819	25	7

Table 6

Frequency of response types by condition in Experiment 4

	Type of response			
Conditions	No-response	Correct	Incorrect	Other
Size-unrelated words	35	1080	4	11
Similar congruent words	37	1079	4	11
Similar incongruent words	32	1087	4	7
Dissimilar congruent words	44	1079	9	11
Dissimilar incongruent words	43	1057	9	11
Congruent size words	31	1089	2	6
Incongruent size words	43	1012	50	13

Table 7

Standard Stroop Effects and Their Components in the Four Experiments

Effect (contrast)	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Standard interference (incongruent size words − baseline)	23.6* [8.3, 38.9]	31.8*** [17.1, 46.4]	28.3** [13.5, 43.2]	37.0**[a] [18.0, 55.9]
Task conflict (unrelated words − symbol strings)	−15.4 [−30.1, −0.7]	−5.5 [−19.5, 8.5]	−14.5 [−28.8, −0.2]	n.e.
Semantic conflict (incongruent related − unrelated)	−1.5 [−16.1, 13.1]	16.5 [2.2, 30.7]	n.e. (12.0 [−2.4, 26.3])[b]	12.9 [−3.9, 29.8]
Response conflict (incongruent size words − incongruent related)	40.5*** [25.2, 55.8]	20.8 [6.0, 35.6]	n.e.	24.0 [5.0, 43.1]
Standard facilitation (unrelated words − congruent size words)	31.0*** [16.7, 45.3]	20.4* [6.7, 34.1]	28.9*** [15.0, 42.7]	36.1*** [18.4, 53.8]
Semantic facilitation (unrelated − congruent related)	10.1 [−4.3, 24.6]	2.8 [−11.1, 16.7]	n.e. (−2.6 [−16.7, 11.6])[b]	15.3 [−1.1, 31.8]
Response facilitation (congruent related − congruent size words)	20.8* [6.7, 35.0]	17.6 [3.9, 31.3]	n.e.	20.7 [3.3, 38.2]
Minimum detectable difference (80% power)	≈ 29	≈ 29	≈ 30	≈ 34

Note. Differences in estimated marginal means (ms) with unadjusted 95% confidence intervals. Related words were size-related words in Experiment 1, size antonyms and synonyms in Experiment 2, and dissimilar incongruent and similar congruent words in Experiment 4. n.e. = not estimable with the conditions of that experiment. [a] Against size-unrelated words, the only control condition in Experiment 4. [b] Experiment 3 had no size-related words; the value in parentheses is the corresponding contrast with incongruent or congruent similar words. * $p < .05$, ** $p < .01$, *** $p < .001$, Holm-corrected across the 21 pairwise comparisons of each experiment.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx