Running head: EVALUATIVE PRIMING

Evaluative Priming in the Pronunciation Task: A Preregistered Replication and Extension

Karl Christoph Klauer and Manuel Becker
Albert-Ludwigs-Universität Freiburg

Adriaan Spruyt
Ghent University

Author Note

Karl Christoph Klauer, Institut für Psychologie; Manuel Becker, Institut für Psychologie; Adriaan Spruyt, Department of Experimental-Clinical and Health Psychology

Preparation of this article was supported by Methusalem Grant BOF09/01M00209 of Ghent University.

Correspondence concerning this article should be addressed to K. C. Klauer at the Institut für Psychologie, Albert-Ludwigs-Universität Freiburg, D-79085 Freiburg, Germany. Electronic mail may be sent to christoph.klauer@psychologie.uni-freiburg.de.

Address information for Christoph Klauer (corresponding author):

- E-mail: christoph.klauer@psychologie.uni-freiburg.de
- Phone: +49 761 2032469
- Postal address:
  - Institut für Psychologie
  - Albert-Ludwigs-Universität Freiburg
  - D-79085 Freiburg
  - Germany

WORD COUNT: 5,152

Abstract

We replicated and extended a study by Spruyt and Hermans (2008) in which picture primes engendered an evaluative-priming effect on the pronunciation of target words. As preliminary steps, we assessed data reproducibility of the original study, conducted Pilot Study I to identify highly semantically related prime-target pairs, reanalyzed the original data excluding such pairs, conducted Pilot Study II to demonstrate that we can replicate traditional associative priming effects in the pronunciation task, and conducted Pilot Study III to generate relatively unrelated sets of prime pictures and target words. The main study comprised three between-participants conditions: (a) A close replication of the original study, (b) the same condition excluding highly related prime-target pairs, and (c) a condition based on the relatively unrelated sets of prime pictures and target words developed in Pilot Study III. There was little evidence for an evaluative priming effect independent of semantic relatedness.

KEYWORDS: Evaluative priming, affective priming, pronunciation task, replicability

The evaluative priming paradigm is a sequential priming paradigm introduced by Fazio, Sanbonmatsu, Powell, and Kardes (1986). It allows one to gauge the impact of the valence of briefly shown prime stimuli on responses to subsequently presented target stimuli. An evaluative-priming effect is observed when responses to target stimuli (e.g., sunshine) occur faster and more accurately following evaluatively congruent prime stimuli (e.g., love) than following evaluatively incongruent prime stimuli (e.g., hate). The paradigm continues to attract a considerable amount of research interest, perhaps due to its role in studying the spontaneous activation of attitudes (e.g., Fazio et al., 1986), as one of the first implicit measures of attitudes (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009), and as a model of behavioral priming effects (Fiedler, Bluemke, & Unkelbach, 2011).

One line of research has looked at evaluative-priming effects in the pronunciation task. In that task, the targets are words and the participants' task is to read the target word out loud as fast as possible. The pronunciation task has gained some prominence because the occurrence of evaluative-priming effects in it would suggest relatively strongly (a) that prime valence affects the encoding of valenced targets (Wentura & Frings, 2008) and (b) that primes engender effects independently of a goal to evaluate (De Houwer, Hermans, & Spruyt, 2001). Both conclusions (a) and (b) would imply evaluative priming to be surprisingly general in scope, and they would strongly constrain theoretical accounts of evaluative priming.

Early attempts to find such effects yielded mixed results (see Klauer & Musch, 2003, for a summary). In particular, it seems difficult to replicate seminal findings by Bargh, Chaiken, Raymond, and Hymes (1996), who observed evaluative-priming effects in the traditional paradigm in which primes and targets are words (Klauer & Musch, 2001; Spruyt, Hermans, Pandelaere, De Houwer, & Eelen, 2004). But in recent years, evaluative-priming effects were repeatedly reported in the pronunciation task when picture primes and/or modified priming procedures were employed.

In particular, four studies have used picture primes and reported evaluative-priming effects in the pronunciation task (Duckworth, Bargh, Garcia, & Chaiken, 2002, Exp. 2; Giner-Sorolla, Garcia, & Bargh, 1999, Exp. 2; Spruyt & Hermans, 2008; Spruyt, Hermans, De Houwer, & Eelen, 2002, Exp. 3). A fifth such study (Everaert, Spruyt, & De Houwer, 2011) found a significant effect in one of two conditions. Other studies have used modified paradigms employing a variety of measures presumably affecting the semantic or evaluative processing of primes and/or targets (e.g., De Houwer, Hermans, & Spruyt, 2001; De Houwer & Randell, 2004; Pecchinenda, Ganteaume, & Banse, 2006; Everaert et al., 2011; Spruyt, De Houwer, & Hermans, 2009). A related line of research has used the picture-naming task, in which targets are pictures and the task is to name the object depicted in the picture (Spruyt et al., 2002; Spruyt, Hermans, De Houwer, Vandromme, & Eelen, 2007; Wentura & Frings, 2008). Collecting these different paradigms and tasks under the common categorization "pronunciation task/naming task", a recent meta-analysis (Herring et al., 2013) concluded that there was a significant evaluative-priming effect across these studies. The present paper is focused on effects reported for picture primes and target words in the non-modified evaluative-priming paradigm with the pronunciation task.[1]

The just-mentioned meta-analysis also adduced evidence for publication bias in the evaluative-priming literature indicating that the published results may present a distorted picture. Given that the effect size estimated for the pronunciation/naming studies ($d = 0.29$) was substantially smaller than that for studies with the frequently used evaluative-decision task in which targets have to be classified as positive or negative ($d = 0.45$), it may be speculated that the pronunciation studies are especially likely to have produced null results and to have disappeared in file drawers. In fact, the first author has conducted multiple experiments using the pronunciation task producing null results and did not attempt or succeed to get them published. In the light of the relatively small number of published studies that have used the pronunciation task and considering the low power of methods for detecting publication bias, it is, however, difficult to assess whether publication bias is more or less pronounced for pronunciation studies than for studies using other tasks.

Given the evidence for publication bias, preregistered replications are desirable to enhance one's confidence in the reliability of the relevant literature despite publication bias. Another limitation of the literature is that priming effects in the pronunciation task using picture primes were reported by only two workgroups, making it desirable to assess the generalizability of effects beyond these laboratories (more than two workgroups were involved in the above-discussed set of studies sometimes categorized as pronunciation/naming-task studies). Furthermore, we engage in an adversarial collaboration; a widely recommended constructive method to resolve scientific conflict (e.g., Kahneman, 2003). For the present replication, we chose the study by Spruyt and Hermans (2008) on the basis of its methodological strengths. In particular, unlike almost all other studies, it avoided the use of target repetition, at least for a first block of trials. Klauer and Musch (2001) argued that target repetition jeopardizes the intended interpretation of priming effects as demonstrating facilitated target encoding.

In what follows, we first describe the study by Spruyt and Hermans (2008) and then assess its reproducibility, a necessary condition for replicability. Data reproducibility "means that Researcher B (e.g. the reviewer of a paper) obtains exactly the same results (e.g. statistics and parameter estimates) that were originally reported by Researcher A (e.g. the author of that paper) from A‘s data when following the same methodology" (Asendorpf et al., 2013, p. 109).

In the course of assessing reproducibility, a potential alternative hypothesis in terms of a semantic-relationship confound for Spruyt and Hermans' (2008) study was identified. This possibility was followed up in a reanalysis of Spruyt and Hermans' (2008) data based on ratings obtained in Pilot Study I. The reanalysis strongly motivates to include additional conditions over and above a replication condition, leading to an extended design. Pilot Study II asesses the suitability of our procedures and instruments for securing priming effects in the pronunciation task, another necessary condition for replicability. In Pilot Study III, new materials were generated for an additional condition of the main study.

Spruyt and Hermans (2008) Study

Procedure

Spruyt and Hermans (2008) used 20 positive and 20 negative words as targets along with 30 positive and 30 negative pictures as primes. Participants completed four blocks of 40 trials each. In each block, the 40 target words were randomly paired with 40 randomly selected pictures, the only restriction being that each trial type (positive-positive, positive-negative, negative-positive, negative-negative) occurred equally often. The experimental blocks were preceded by 10 practise trials using neutral prime pictures and neutral target words. Participants were to pronounce the target word as quickly as possible while ignoring the pictures. The procedures were further described as follows:

> Each trial started with a 500 ms presentation of a fixation cross in the centre of the screen. Five hundred milliseconds after the offset of the fixation cross, the primes were presented for 200 ms. Finally, after an inter stimulus interval of 50 ms, the target stimuli were presented until the participant gave a response or 2,000 ms elapsed. By pressing one of three keys of the computer keyboard, the experimenter coded whether the microphone was triggered accurately and whether the participant's response was correct. After the experimenter entered the code, the next trial was initiated after a time interval that varied randomly between 500 ms and 1,500 ms (Spruyt & Hermans, 2008, p. 239).

For the analyses, trials in which the voice key was not appropriately activated or an incorrect response was given were excluded. Responses past the 2,000 ms deadline were also excluded. Response latencies that deviated more than 2.5 standard deviations from a participant's conditional mean latency were also discarded, where "conditional" refers to the cells of the design spanned by the factors block and evaluative congruency. An analysis of variance with these factors revealed a significant main effect of evaluative congruency with estimated effect size $d = 0.44$. In addition, there was a significant evaluative priming effect for the first block of trials with estimated effect size $d = 0.35$.

Reproducibility

Based on the files with the raw data, we reproduced all reported statistics precisely with the exception of the number of excluded outliers which amounted to 2.04% of all trials rather than 1.70% as stated in the article. In total, 5.30% of trials were excluded for different reasons.[2]

The raw data suggested a potentially more important discrepancy between the article and the data, however. The set of targets listed in the appendix of Spruyt and Hermans (2008) differs from the set of targets written into the raw data files. Discussion with Adriaan Spruyt revealed that the set reported in the article is not the one that was actually used in the experiment although there is considerable overlap between the two. This discrepancy led to the publication of an erratum, "Correction to Spruyt and Hermans (2008)" (2014), in which the correct target words are reported. The discrepancy does not affect any of the statistical analyses or conclusions drawn from the original paper.

A Potential Confound, Pilot Study I, and Reanalysis

Inspection of the target words and of the prime pictures reveals a potential confound: The target words are relatively frequently strongly descriptive of the prime pictures. For example, the target "tumor" can be seen as a description of one of the prime pictures depicting a huge tumor; the target 'haat' [hate] is exemplified by several prime pictures depicting threatening, angry men; the target 'romantiek' [romance] is allegorically depicted by pictures showing happy couples, a bride, a rose, and so forth. Such close semantic relationships raise the possibility that part of the evaluative-priming effects were in fact due to uncontrolled strong associative-semantic links between primes and targets (Neely, 1991).

Pilot Study I

To assess this possibility, we had six raters rate all 2400 possible prime-target pairs on a four-point rating scale with respect to how well they fit each other. The instructions explained that a picture and a word fit each other if, for example, the word describes the picture well, the picture is an allegorical depiction of the word, the picture gives an example of the word, or if the picture makes one think spontaneously of the word. The four points of the rating scale were labeled from 1 to 4, in order, as "do not fit at all", "fit rather not", "fit somewhat", "fit very well". One rater only used the cautious middle categories with few exceptions, and we did not use her data further.

Not surprisingly, evaluatively congruent stimuli ($M = 2.75$, $SD = 0.72$) fit each other better than evaluatively incongruent stimuli ($M = 1.21$, $SD = 0.24$); in particular, none of the raters gave the highest rating "4" for any evaluatively incongruent pair.

Reanalysis

For a reanalysis of Spruyt and Hermans' (2008) data, we excluded all pairs for which a majority of the five raters had selected the highest rating, thereby excluding 233 of the 2400 pairs. This reduced the mean fit rating for evaluatively congruent stimuli slightly to $M = 2.52$, $SD = 0.60$.

The total percentage of excluded trials was thereby elevated to 14.19% (from 5.30%). An analysis of variance with within-participants factors block and evaluative congruency showed that the main effect of block reported by Spruyt and Hermans (2008) was as strong as in the original analysis, $F(2.39, 107.67) = 8.99$, $p < .01$, previously $F(2.50, 112.44) = 8.42$, whereas the main effect of congruency was erased: $F(1, 45) = 1.49$, $p = .23$, from $F(1, 45) = 8.83$, $p < .005$ ($F < 1$ for the interaction of both factors). The mean priming effect was 1.8 ms ($SD = 9.7$), down from $M = 4.0$ ms ($SD = 9.2$) in the original study. Neither was the evaluative-priming effect significant in the first block of trials, $t(45) = 1.17$, $p = .25$, $M = 3.6$ ms, $SD = 21$, previously $t(45) = 2.40$, $p < .05$, $M = 6.9$ ms, $SD = 20$, nor in an analysis of variance with first block of trials eliminated (all Fs involving congruency smaller than 1).

This suggests that the evaluative-priming effects reported by Spruyt and Hermans (2008) may partly or completely have been based on a few strongly semantically related prime-target pairs. Of course, this evidence is only suggestive, because (a) eliminating trials reduces the statistical test power for detecting effects (note, however, that the block effect remained, if anything, as strong as in the original analysis), and because (b) the fit ratings were based on target words translated into German and on German rather than Belgian participants so that there may be cultural differences in the degree of fit as perceived by Spruyt and Hermans's (2008) participants and the present raters. Nevertheless, the results of the pilot study and reanalysis strongly motivate to extend the replication by additional conditions in which at least the most strongly semantically related prime-target pairs are excluded a priori.

Pilot Study II

The purpose of Pilot Study II was to demonstrate that the procedures and instruments (such as our voicekey) are capable of documenting priming effects if they exist. Like data replicability, this is a necessary condition for replicability, because sloppy and noisy data collection would decrease effect sizes and can thereby prevent a true effect from emerging reliably.

In Pilot Study II, we implemented a sequential priming paradigm using the pronunciation task and exactly the same procedures for the timing of primes and targets and the recording of responses as in Spruyt and Hermans (2008). Primes and targets were, however, both words that were either strongly associated (e.g., mother-father) or not related. The related prime-target pairs were the same 50 pairs already used by Klauer and Musch (2001). The 50 unrelated prime-target pairs were formed by randomly repairing primes and targets. The 100 prime-target pairs were presented in a random order that was determined anew for each participant with the restriction that the same target word did not appear in two consecutive trials. Associative priming in the pronunciation task is a well documented effect (Neely, 1991) and therefore an appropriate benchmark for testing our procedures.

The 20 participants were mostly students from the University of Freiburg with different majors (mean age 23.5 years, $SD = 3.4$, 13 female). They received course credit or a monetary gratification of Euro 2.00 for participating. Primes and targets were presented in the center of a 58 cm LCD monitor with a resolution of 1920 by 1080 pixels and a seating distance of approximately 60 cm. Words were presented in Times New Roman font and subtended 54 pixels vertically.

Data were preprocessed exactly as in Spruyt and Hermans (2008; see above), leading to the exclusion of 4.10% of the trials. An associative priming effect of 10 ms ($SD = 12$) and estimated effect size of $d = .88$ emerged that was significantly different from zero: $t(19) = 3.93$, $p < .01$.

Main Study

The main study comprised three conditions implemented as between-participants factor. A replication condition was a close replication of the study by Spruyt and Hermans (2008). In a second condition, all stimuli from the original study were retained, but the 233 strongly related prime-target combinations identified in Pilot Study I were never be presented. A third condition used different sets of prime pictures and target words for which highly related prime-target pairs were a priori unlikely to occur.

A sample size of $N = 58$ is required for a power of .95 to detect an effect of evaluative congruency of the size observed in the original study ($d = 0.44$) by means of a one-sided t test, and we planned to collect $N = 60$ participants per condition. An anonymous reviewer suggested prior to data collection to use the above-mentioned effect size of $d = 0.29$ estimated for pronunciation/naming studies and to obtain enough participants such that the two relatively unconfounded replication attempts (second and third condition) taken together achieve a test power of .95 in a one-sided t test. This requires $N = 131$ participants, and following the reviewer's suggestion, we sampled $N = 66$ participants for each of the three conditions.

Pilot Study III

For the third condition, we selected pictures of ugly and beautiful landscapes as negative and positive primes, respectively, all of them of a size of 512 by 384 pixels, like the pictures used by Spruyt and Hermans (2008). Target words denoted positive and negative traits excluding those that might be readily applied to landscapes (such as beautiful, attractive, etc.). We obtained valence ratings on a 7-point scale for 60 such pictures and 60 such words from $N = 10$ raters. We selected two sets of 40 pictures and 40 words for use in the experiment on the basis of these valence ratings and so that positive and negative words were matched in word frequency and in word length. Table 1 shows descriptive statistics for the final picture and word sets. The target words are listed in the appendix, the prime pictures can be obtained from the authors upon request.

Method

The procedures closely followed those of Spruyt and Hermans (2008) unless where otherwise mentioned. A replication condition provided a close replication of the original study, a condition with reduced prime-target relatedness sampled from the same primes and targets as the original study, but excluded the 233 highly related prime-target pairs identified in Pilot Study I from being sampled. A condition with unrelated stimuli was based on the landscape primes and trait words just described. The stimuli for the 10 practice trials were sampled from the sets of primes and targets not in use for the participant's condition (excluding highly-related prime-target pairs). Departing from the proposal submitted prior to data collection, we used, however, the neutral stimuli (with neutral words translated into German) employed in the original experiment for the practice trials of the replication condition. This change was cleared with the action editor prior to data collection (Christian Frings, personal communication, October, 23rd, 2014).

Participants. Participants were mostly University-of-Freiburg students with different majors. They received either partial course credit or 2 Euros for participating. A total of 204 participants were sampled.[3] Five of these were excluded as extreme outliers according to Tukey's criterion (more than three times the interquartile range above the upper or below the lower quantile) in the sample's distribution of the number of correctly pronounced words ($M = 157$, $SD = 4.67$) with less than 141 correct responses. One was excluded as extreme outlier in the sample's distribution of correct response latencies ($M = 496$ ms, $SD = 58$) with an average response latency of 728 ms. After these exclusions, there were $N = 66$ persons per condition for the analyses.

Of these 198 participants, demographic data were lost for one participant due to computer problems. Of the remaining 197 participants, 75 were male, 121 female, and one person chose not to respond to the gender question. Participants' mean age was 22.62 years ($SD = 3.93$).

Results

Preregistered analyses. The main analyses closely followed Spruyt and Hermans (2008) in terms of outlier criteria and statistical tests. Like in Spruyt and Hermans (2008), a few responses coded as correct occurred past the response deadline of 2 s, with response latency set to 2 s by the program, and in the present data, a few such responses occurred with unrealistically short latencies (such as smaller than 50 ms). Comparing the codings of the Spruyt and Hermans (2008) data and the present data, coders were generally more liberal for the Spruyt and Hermans (2008) data in accepting too late responses and coders of the present data in accepting responses with short latencies. This suggests that there is some leeway for subjective coder criteria in the coding of such responses. This in turn suggests that it is probably wise to have coders blind to the experimental hypotheses in future studies.

Responses past the response deadline were excluded as in Spruyt and Hermans (2008), and the three authors also agreed to take out responses that occurred 150 ms or sooner after target onset for the present analyses. Note, however, that the results pattern of the preregistered analysis is robust against not taking out such responses or using different cut-off values such as 50 ms or 100 ms. Response latencies that deviated more than 2.5 standard deviations from a participant's conditional mean latency were excluded as in Spruyt and Hermans (2008).

This led to an exclusion of 1.75% of the trials due to voicekey failures, wrong or missing responses and to an additional exclusion of 1.75% of the remaining correct-response trials as outliers. For the analyses of variance reported below, degrees of freedom are Greenhouse-Geisser corrected.

Table 2 presents means and standard deviations of the correct-response latencies as a function of the between-participants factor condition, and within-participants factors block and evaluative congruency. An analysis of variance with these factors revealed a significant main effect of block, $F(1.87, 364.81) = 80.09$, $p < .01$, and a significant interaction of block and condition, $F(3.74, 364.81) = 4.28$, $p < .01$. There was no main effect of evaluative congruency, $F(1, 195) = 1.24$, $p = .27$, nor an interaction of block and evaluative congruency, $F(2.87, 560.03) = 1.60$, $p = .19$; all other $F \leq 1.01$. The average priming effect was 0.9 ms ($SD = 11.4$, $d = 0.08$).

A second preregistered analysis of variance with condition and evaluative congruency was conducted just for the first block of trials in which there is no stimulus repetition. Again, the main effect of evaluative congruency was not significant, $F(1, 195) = 1.47$, $p = .23$; all other $F < 1$.

Exploratory Analyses

Both the first two authors and the third author, independently from each other, explored the effects of various analytic choices on the results. These analytic choices include the exclusion/exclusion of the block factor, the excluding/inclusion of different subsets of trials or participants on various grounds, the use of a logarithmic transformation of the raw data, and combinations of these settings. The first two authors also looked at mixed models that include random intercepts and slopes for participants and/or items. Both the first two authors and the third author found that some (combinations of) analytic choices resulted in significant results for the evaluative-priming effect, whereas others did not. For the sake of brevity, we report only a small subset of these analyses.

In particular, a joint analysis (total $N = 285$) of the present data, the original data ($N = 46$) collected by Spruyt and Hermans (2008), and the data of a recent study ($N = 41$), run at Ghent university, in which the second condition of the present experiment was replicated, was also performed. This second replication study revealed no reliable evidence for an evaluative priming effect, neither across blocks of trials, $F < 1$, nor in the first block of trials, $F(1, 40) = 1.10$, $p = .30$.[4]

Across all studies, however, a small but reliable evaluative priming effect emerged.[5] As can be seen in Table 3, this evaluative priming effect was not significantly moderated by the semantical relatedness of prime-target pairs, the language/lab (location) in which the experiment was run, nor the precise study. This data pattern also replicated irrespective of whether (a) the block variable was included in the design or (b) the analyses were restricted the data of the first block only. Still, in each of these analyses, the overall evaluative priming effect was numerically very small, as were the respective effect sizes (see Table 3). Moreover, when the trials with the most strongly related prime-target pairs identified in Pilot Study I were removed, just like we did for the reanalysis of Spruyt and Hermans (2008), these effects dropped to non-significance, all $F < 2.21$, all $p > .14$.

Discussion

There was little evidence for an evaluative-priming effect in pronunciation latencies in the preregistered analyses. This was in particular true for the data from the first block, in which there was no stimulus repetition and in which an evaluative-priming effect of the size reported by Spruyt and Hermans (2008) would therefore have been especially conclusive in its theoretical implications had it occurred. The evidence for small evaluative-priming effects in the exploratory analyses of the joint data appears to hinge on the inclusion of highly semantically related prime-target pairs in the analyses, that is on the semantic-relatedness confound already discussed.

It is possible that the present small 0.9 ms priming effect ($d = 0.08$) would turn out significant if an even larger sample of participants were obtained. Note, however, that such a small priming effect would be difficult to interpret theoretically, given that residual confounds of evaluative congruency with associative and semantic relatedness cannot be ruled out even when the most highly semantically related prime-target pairs are taken out of the analyses. For example, ratings of semantic relatedness were still substantially higher for evaluatively-congruent prime-target pairs than for incongruent pairs even after the most highly related pairs were taken out (see Pilot Study I). Future work might make an effort to contrast evaluatively congruent and incongruent stimulus pairs equated on ratings of semantic relatedness or might select pairs for which evaluative congruency and semantic relatedness vary orthogonally.

Priming results may differ as a function of the language in which the study is run. For example, one criticism of the null findings reported by Klauer and Musch (2001) and in particular of their failure to replicate the original Bargh et al. (1996) priming effects in the pronunciation task was that the German language may be orthographically less deep than the English language. Orthographic depth increases with the complexity of the print-to-speech correspondences, and languages differ systematically in orthographic depth. The impact of lexical and semantic variables on pronunciation latencies may be reduced in shallow orthographies, because a simple direct route from print to speech is available (e.g., Frost, Katz, & Bentin, 1987; but see Pagliuca, Arduino, Barca, & Burani, 2008). To address this criticism, Klauer and

Musch (2001) also ran a study with English-speaking participants and English stimuli, which did not change results. Similarly, Spruyt et al. (2004) also failed to replicate the result using native English speakers. One reviewer of the present replication proposal analogously raised the issue that the German language may be less deep orthographically than the Dutch language in which Spruyt and Hermans' (2008) study was couched. According to the reviewer, this might make it more difficult to find an effect.

A number of objective quantifications of orthographic depth have been proposed (Schmalz, Marinus, Coltheart, & Castles, 2015). According to a quantification based on the dual route cascaded model of reading (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001), "English is a 'deep' orthography, in that it has many rules, and a particularly high percentage of irregular words, while Dutch and German are 'shallow', in that they have few rules and a small proportion of irregular words." (Schmalz et al., 2015, no page number available yet). In this quantification, German and Dutch do not differ substantially in orthographic depth. According to an alternative quantification proposed by Borgwaldt, Hellwig, and Groot (2005), the German language is even substantially deeper orthographically than the Dutch language. This is consistent with the fact that our participants were slower by an average amount of about 50 ms than Spruyt and Herman's (2008) in naming the targets (see also Footnote 5). Note finally that the failed replication presented in the section on exploratory analyses was run in Dutch.

In closing, let us emphasize once more that the present failures to replicate do not question the evaluative-priming literature on the pronunciation task and the picture-naming task as a whole. Although they raise a concern for the many studies that did not control for semantic relatedness of primes and targets, the present results directly speak only to the paradigm and study by Spruyt and Hermans (2008). Each of the other paradigms reviewed in the introduction needs to be scrutinized separately and a few attempts to do so are underway (e.g., Becker, Klauer, & Spruyt, 2015).

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119. doi: 10.1002/per.1919

Bargh, J. A., Chaiken, S., Raymond, P., & Hymes, C. (1996). The automatic evaluation effect: Unconditional automatic attitude activation with a pronunciation task. *Journal of Experimental Social Psychology*, *32*, 104-128. doi: 10.1006/jesp.1996.0005

Becker, M., Klauer, K. C., & Spruyt, A. (2015). *Is attention enough? A re-examination of the impact of feature-specific attention allocation on semantic priming effects in the pronunciation task.* (Preregistered report accepted by Attention, Perception, & Psychophysics.)

Borgwaldt, S. R., Hellwig, F. M., & Groot, A. M. B. D. (2005). Onset entropy matters: Letter-to-phoneme mappings in seven languages. *Reading and Writing*, *18*, 211–229. doi: 10.1007/s11145-005-3001-9

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental Psychology*, *58*, 412–424. doi: 10.1027/1618-3169/a000123

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–256. doi: 10.1037/0033-295X.108.1.204

"Correction to Spruyt and Hermans (2008)". (2014). Correction to Spruyt and Hermans (2008). *Canadian Journal of Experimental Psychology*, *68*, 132–132. doi: 10.1037/cep0000022

De Houwer, J., Hermans, D., & Spruyt, A. (2001). Affective priming of pronunciation responses: Effects of target degradation. *Journal of Experimental Social Psychology*, *37*, 85-91. doi: 10.1006/jesp.2000.1437

De Houwer, J., & Randell, T. (2004). Robust affective priming effects in a conditional pronunciation task: Evidence for the semantic representation of evaluative information. *Cognition & Emotion*, *18*, 251–264. doi: 10.1080/02699930341000022

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, *135*, 347-368. doi: 10.1037/a0014211

Duckworth, K. L., Bargh, J. A., Garcia, M., & Chaiken, S. (2002). The automatic evaluation of novel stimuli. *Psychological Science*, *13*, 513-519. doi: 10.1111/1467-9280.00490

Everaert, T., Spruyt, A., & De Houwer, J. (2011). On the (un)conditionality of automatic attitude activation: The valence proportion effect. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *65*, 125–132. doi: 10.1037/a0022316

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*, 229-238. doi: 10.1037/0022-3514.50.2.229

Fiedler, K., Bluemke, M., & Unkelbach, C. (2011). On the adaptive flexibility of evaluative priming. *Memory & Cognition*, *39*, 557–572. doi: 10.3758/s13421-010-0056-x

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multinlingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 104-115. doi: 10.1037/0096-1523.13.1.104

Giner-Sorolla, R., Garcia, M. T., & Bargh, J. A. (1999). The automatic evaluation of pictures. *Social Cognition*, *17*, 76-96. doi: 10.1521/soco.1999.17.1.76

Herring, D. R., White, K. R., Jabeen, L. N., Hinojos, M., Terrazas, G., Reyes, S. M., . . . Crites, S. L. J. (2013). On the automatic activation of attitudes: A quarter century of evaluative priming research. *Psychological Bulletin*, *139*, 1062–1089. doi: 10.1037/a0031309

Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist*, *58*, 723–730. doi: 10.1037/0003-066X.58.9.723

Klauer, K. C., & Musch, J. (2001). Does sunshine prime loyal? Affective priming in the naming task. *Quarterly Journal of Experimental Psychology*, *54*, 727-751. doi: 10.1080/713755986

Klauer, K. C., & Musch, J. (2003). Affective priming: Findings and theories. In J. Musch & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (p. 7-49). Mahwah, NJ: Lawrence Erlbaum.

Kupietz, M., Belica, C., Keibel, H., & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari et al. (Eds.), *Proceedings of the seventh international conference on language resources and evaluation.* Valletta, Malta: European Language Resources Association (ELRA).

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (p. 264-336). Hillsdale, NJ: Erlbaum.

Pagliuca, G., Arduino, L. S., Barca, L., & Burani, C. (2008). Fully transparent orthography, yet lexical reading aloud: The lexicality effect in Italian. *Language and Cognitive Processes*, *23*, 422–433. doi: 10.1080/01690960701626036

Pecchinenda, A., Ganteaume, C., & Banse, R. (2006). Investigating the mechanisms underlying affective priming effects using a conditional pronunciation task. *Experimental Psychology*, *53*, 268–274. doi: 10.1027/1618-3169.53.4.268

Schmalz, X., Marinus, E., Coltheart, M., & Castles, A. (2015). Getting to the bottom of orthographic depth. *Psychonomic Bulletin & Review*, 1–16. doi: 10.3758/s13423-015-0835-2

Spruyt, A., De Houwer, J., & Hermans, D. (2009). Modulation of automatic semantic priming by feature-specific attention allocation. *Journal of Memory & Language*, *61*, 37-54. doi: 10.1016/j.jml.2009.03.004

Spruyt, A., & Hermans, D. (2008). Affective priming of naming responses does not depend on stimulus repetition. *Canadian Journal of Experimental Psychology*, *62*, 237–241. doi: 10.1037/1196-1961.62.4.237

Spruyt, A., Hermans, D., De Houwer, J., Vandromme, H., & Eelen, P. (2007). On the nature of the affective priming effect: Effects of stimulus onset asynchrony and congruency proportion in naming and evaluative categorization. *Memory and Cognition*, *35*, 95-106. doi: 10.3758/BF03195946

Spruyt, A., Hermans, D., Houwer, J. D., & Eelen, P. (2002). On the nature of the affective priming effect: Affective priming of naming responses. *Social Cognition*, *20*, 227–256. doi: 10.1521/soco.20.3.227.21106

Spruyt, A., Hermans, D., Pandelaere, M., De Houwer, J., & Eelen, P. (2004). On the replicability of the affective priming effect in the pronunciation task. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie)*, *51*, 109–115. doi: 10.1027/1618-3169.51.2.109

Wentura, D., & Frings, C. (2008). Response-bound primes diminish affective priming in the naming task. *Cognition & Emotion*, *22*, 374–384. doi: 10.1080/02699930701446064

Appendix

Target Words in the Condition with Unrelated Stimuli

Negative Words

Müde [tired], langsam [slow], eitel [vain], hämisch [gleeful], bissig [snappy], naiv [naive], berechnend [scheming], gelangweilt [bored], dumm [stupid], launisch [fickle], eifersüchtig [jealous], eingebildet [conceited], unfähig [unable], bessesen [obsessed], bestechlich [bribable], süchtig [addicted], neidisch [envious], arrogant [arrogant], gierig [greedy], betrügerisch [fraudulous]

Positive Words

Altruistisch [altruistic], diskret [discrete], raffiniert [refined], sensibel [sensitive], behutsam [gentle], ausdauernd [enduring], lässig [cool], pünktlich [punctual], barmherzig [compassionate], diplomatisch [diplomatic], zuvorkommend [courteous], ausgelassen [exuberant], geschickt [skillful], gütig [benign], belesen [well-read], fleißig [assiduous], aktiv [active], selbstsicher [self-assured], humorvoll [humorous], witzig [funny]

Footnotes

[1] The first two authors believe that the above-mentioned different paradigms are vulnerable to different alternative explanations and confounds and that it is therefore misleading to lump them together. We intend to spell out and examine these alternative explanations in the course of a larger project of which the present manuscript constitutes the first step targeting specifically and only evaluative priming in the pronunciation task using picture primes.

[2] In the original article, degrees of freedom are Greenhouse-Geisser corrected for the F tests involving the factor "block", but this is not the case in the computation of the reported *MSE* values as some readers might perhaps have expected. Finally, in Footnote 3, the degrees of freedom reported for the main effect of block are wrongly specified, the correct values are 1.75 and 78.52.

[3] A number of additional participants were excluded a priori because technical problems occurred during the experimental session (2 participants), because they did not complete the experiment (4 participants), because they had already participated once in the same experiment (2 participants), or because German was not their first language (2 participants). One of the last participants to take part in the experiment was excluded because we had already sampled the planned N of 66 per condition for the participant's condition.

[4] Nevertheless, the overall evaluative priming effect did correlate significantly ($r = .36$, $p < .05$) with the extent to which participants deemed the experiment to have important implications (i.e., as captured by a post-experimental questionnaire that was included for exploratory purposes). This observation suggests that the evaluative priming effect in the pronunciation task may be dependent upon person-specific beliefs and expectations, which might be an interesting avenue for further research, not only for researchers working on the evaluative priming effect but also for replication attempts in general.

[5] The joint analysis also revealed that the number of coded voice key failures was

significantly smaller in the present study as compared to the original study by Spruyt and Hermans (2008) and the recent replication attempt run at Ghent University, $F(1, 283) = 116.95$, $p < .001$, $\Delta = 2.84\%$. This observation suggests that the coding accuracy, coding criteria, and/or voice key sensitivity settings were different in both (sets of) studies.

Table 1

Descriptive Statistics for the Materials in the Condition with Unrelated Stimuli (SDs in parentheses)

		Valence rating		Word frequency		
		Mean	Polarization	Written	Spoken	Word length
Pictures	Negative	1.52 (0.23)	2.49 (0.23)			
	Positive	6.51 (0.16)	2.51 (0.16)			
Words	Negative	2.42 (0.54)	1.59 (0.54)	2.02 (0.82)	14.80 (1.89)	7.80 (0.59)
	Positive	5.69 (0.48)	1.69 (0.48)	1.94 (0.68)	14.25 (2.07)	8.75 (0.53)

Note. The norms for spoken language stem from Brysbaert, Buchmeier, Conrad, Jacobs, Bölte, and Böhl (2011); for written language from the German reference corpus (Kupietz, Belica, Keibel, & Witt, 2010). One negative word was not part of the database for spoken words. For both pictures and words, valence ratings depart significantly from the midpoint for both positive and negative stimuli (all four $t > 13.0$, $p < .001$) and positive and negative stimuli do not differ in this degree of polarization (both $t < 1$).

Table 2

Means and SDs of Correct-Response Latencies and Priming Effects in ms

		Condition 1			Condition 2			Condition 3		
Block	Statistic	Inc.	Con.	PE	Inc.	Con.	PE	Inc.	Con.	PE
1	*M*	514.6	513.7	0.9	504.3	503.6	0.8	521.9	516.8	5.1
	SD	75.3	80.1	23.2	58.2	54.2	24.0	71.1	64.2	30.6
2	*M*	492.8	498.3	-5.5	483.6	483.2	0.5	486.8	487.4	-0.6
	SD	64.9	66.1	22.2	46.5	48.0	16.6	50.6	51.8	21.8
3	*M*	495.5	492.6	2.9	485.2	485.0	0.2	486.4	484.6	1.8
	SD	61.4	62.2	19.7	50.1	49.6	19.1	57.1	53.14	19.8
4	*M*	499.5	495.7	3.8	486.3	486.0	0.3	483.4	482.7	0.7
	SD	65.3	64.8	19.0	50.2	53.3	19.5	48.5	47.9	16.5

Note. Con.= evaluatively congruent; Inc = evaluatively incongruent; PE = priming effect. Condition 1 is the replication condition; in Condition 2, highly related prime-target pairs are never presented; in Condition 3, prime-target pairs are less related a priori.

Table 3

Overview of the Effects Observed in a Joint Analysis of the Present Study, the Original Study Reported by Spruyt and Hermans (2008), and a Recent Replication Attempt by the Third Author (Total $N = 285$)

	Evaluative congruence			Interaction effects		
Analysis	F	PE	η_p^2	Study	Location	Relatedness
Block factor included	5.18^{*}	1.44 ms	.018	$F < 1.60$	$F < 1.65$	$F < 1$
Block factor excluded	7.48^{**}	1.68 ms	.026	$F < 1$	$F < 1$	$F < 1$
First block only	4.80^{*}	3.16 ms	.017	$F < 1$	$F < 1$	$F < 1$

Note. PE = priming effect.

$^{*}p < .05$, $^{**}p < .01$.