**Supplemental Materials for**

**How First- and Second-Language Emotion Words Influence
Emotion Perception in Swedish-English Bilinguals**

Marie-France Champoux-Larsson1, 2 & Erik C. Nook3

1 Department of Psychology and Social Work, Mid Sweden University

2 Department of Psychology, University of Chicago

3 Department of Psychology, Princeton University

**Table of Contents**

1. Analyses of Preliminary Preregistered Sample
2. Figure S1 – Reaction time and sensitivity results from preliminary preregistered sample
3. Figure S2 – L2 proficiency results from preliminary preregistered sample

## I. Analyses of Preliminary Preregistered Sample

## Sample description. As described in the main text, we set an initial target sample size of 123 participants. However, the COVID-19 pandemic halted our data collection at 115 participants. We preregistered our analysis plan at this point (<https://doi.org/10.17605/OSF.IO/YPHJ9>) and analyzed available data (presented below). Results were inconclusive, so we lodged an additional preregistration at a new target sample size of 140 usable participants once in-person data collection was possible again ([<https://doi.org/10.17605/OSF.IO/Z8BJK>](https://doi.org/10.17605/OSF.IO/Z8BJK)). That sample was collected and reported in the main text. Thus these supplemental analyses are performed on *N* = 105 usable participants (33 [31.43%] male, 70 [66.67%] female, and 2 [1.90%] other; age ranged 19-40 years, *Mage* = 28.09 years, *SDage* = 5.00 years, racial demographics were not assessed due to IRB restrictions on collecting these data in Sweden) from the initial sample of 115 (one excluded due to English being L1, three excluded due to age > 45, five excluded due to being an outlier > 3 standard deviation from sample mean on accuracy or reaction time measures in any condition, and one excluded due to not responding to 50% of trials).

**Overall analyses**. As in the main text (and the preliminary preregistration), we first report analyses of all conditions before following up main effects and potential interactions with focused tests of key research questions. A 2 [Congruence: congruent vs. incongruent] x 3 [Context: face vs. L1 word vs. L2 word] repeated-measures ANOVA showed a significant main effect of Congruence, *F*(1, 104) = 188.89, *p* < .001, ηp2 = .64, 90% CI = [.55, .71], a main effect of Context, *F*(2, 208) = 339.93, *p* < .001, ηp2 = .77, 90% CI = [.72, .80], and a non-significant interaction between Congruence and Context, *F*(2, 208) = 2.11, *p* = .123, ηp2 = .02, 90% CI = [0, .06] (**Figure S1a**). A one-way repeated-measures ANOVA performed on sensitivity scores also revealed a significant main effect of Context, *F*(2, 208) = 32.32, *p* < .001, ηp2 = .24, 90% CI = [.15, .31] (**Figure S1b**).

## Replicating Prior Work: Does Language Facilitate Emotion Perception?

## We next conducted the same analysis as in Nook et al. (2015) by ignoring L2 word trials and conducting a 2 [Congruence: congruent vs. incongruent] x 2 [Context: face vs. L1 word] ANOVA on RT. We found a main effect of Congruence, *F*(1, 104) = 150.43, *p* < .001, ηp2 = .59, 90% CI = [.49, .66], and a main effect of Context, *F*(1, 104) = 354.16, *p* < .001, ηp2 = .77, 90% CI = [.71, .81]. Finally, a trending interaction between Congruence and Context emerged, *F*(1, 104) = 3.42, *p* = .067, ηp2 = .03, 90% CI = [0, .10]. Although the interaction was not significant (a difference from the result in the main text), follow-up paired-samples *t*-tests showed that the impact of congruence on L1 word trials, *t*(104) = 11.50, *p* < .001, *d* = 0.76, 95% CI = [0.61, 0.90], was numerically larger than the impact of congruence on face trials *t*(104) = 7.61, *p* < .001, *d* = 0.71, 95% CI = [.51, .92]. A paired-sample *t*-test revealed a significant difference between sensitivity for face and L1 word trials, *t*(104) = 7.65, *p* < .001, *d* = 0.77, 95% CI = [0.54, 0.99]. These results replicated prior research that d’ is higher when matching facial expressions to first-language emotion words than when matching with other facial expressions. However, the interaction between Congruence and Context for RT was on the border of significance.

## Comparing Emotion Words in First vs. Second Language.

## We then compared L1 and L2 contexts by ignoring face trials and analyzing RT using a 2 [Congruence: Congruent vs. Incongruent] x 2 [Context: L1 word vs. L2 word] ANOVA. We again found a main effect of Congruence, *F*(1, 104) = 169.68, *p* < .001, ηp2 = .62, 90% CI = [.52, .69]; however, we did not observe a main effect of Context, *F*(1, 104) = 0.81, *p* = .369, ηp2 = .008, 90% CI = [0, .06], nor an interaction between Congruence and Context, *F*(1, 104) = 3.22, *p* = .076, ηp2 = .03, 90% CI = [0, .10]. A follow-up paired-samples *t*-test revealed that the impact of congruence on L2 word trials, *t*(104) = 8.68, *p* < .001, *d* = 0.64, 95% CI = [0.48, 0.80], was not significantly different than for L1 word trials (*d* = 0.76, see above). A paired-sample t-test also showed that d’ for L1 and L2 word trials did not significantly differ, *t*(104) = 0.63, *p* = .530, *d* = 0.06, 95% CI = [-0.12, 0.24]. Thus, native Swedish speakers did not significantly differ in their ability to match facial expressions of emotion with L1 vs. L2 emotion words, and there was not a significant difference in the extent to which congruent facial expressions primed L1 vs. L2 emotion words.

## Influence of Proficiency

First, a 2 [Congruence] x 3 [Context] repeated-measures x self-reported English proficiency ANCOVA was used to analyze RT (**Figure S2a**). As in prior analyses, we observed main effects of both Congruence, *F*(2, 103) = 187.87, *p* < .001, ηp2 = .65, 90% CI = [.55, .71], and Context, *F*(2, 206) = 361.91, *p* < .001, ηp2 = .78, 90% CI = [.74, .81]. We also observed a significant interaction between L2 proficiency and Context, *F*(2, 206) = 7.73, *p* < .001, ηp2 = .07, 90% CI = [.02, .13]. No other effects or interactions were significant, *p*s > .12. We unpacked this interaction by computing correlations between L2 proficiency and average RT for each Context. L2 proficiency was not significantly related to RT in any condition, but the relationship was positive for face trials, *r*(103) = .11, *p* = .251, and negative for both L1, *r*(103) = -.17, *p* = .080, and L2 word trials, *r*(103) = -.13, *p* = .183. This differed from results in the main text, which showed that self-reported L2 proficiency was significantly correlated with faster RTs for both L1 and L2 word trials but not face trials.

Results for the ANCOVA including LexTALE scores (**Figure S2b**) showed again main effects of Congruence, *F*(1, 103) = 196.10, *p* < .001, ηp2 = .66, 90% CI = [.57, .72], and Context, *F*(2, 206) = 343.96, *p* < .001, ηp2 = .77, 90% CI = [.73, .80]. A significant interaction also emerged between LexTALE and Congruence, *F*(1, 103) = 4.97, *p* = .028, ηp2 = .05, 90% CI = [.003, .13]. No other main effects or interactions were significant, *p*s > .08. To unpack this interaction, tested how LexTALE scores correlated with “priming scores” (i.e., the difference in RT between congruent and incongruent trials for each Context). LexTALE scores correlated negatively with priming scores on Face trials, *r*(103) = -.28, *p* = .004, but not L1 word trials, *r*(103) = -.06, *p* = .568, or L2 word trials, *r*(103) = -.07, *p* = .449. As such, just as in the main text, greater objectively measured L2 proficiency was associated with less congruent cue priming on face trials.

Turning to d’, we first conducted a 3 [Context] x self-reported L2 proficiency ANCOVA (**Figure S3c**). As in the main text, we observed main effects of Context, *F*(1, 206) = 32.42, *p* < .001, ηp2 = .24, 90% CI = [.16, .31], and L2 proficiency, *F*(1, 103) = 6.60, *p* = .012, ηp2 = .06, 90% CI = [.01, .15]. The L2 proficiency x Context interaction was not significant, *p* = .275. The LexTALE analyses were the same (**Figure S3d**), yielding significant main effects of Context, *F*(1, 206) = 32.14, *p* < .001, ηp2 = .24, 90% CI = [.15, .31], and LexTALE, *F*(1, 103) = 4.33, *p* = .040, ηp2 = .04, 90% CI = [.001, .12], but no interaction emerged, *p* = .657. Thus, as in the main text, participants in this initial sample who had higher L2 proficiency (both self-reported and objectively assessed) showed higher performance on this task.

**Equivalence tests**. Two key tests of our hypotheses yielded p values between .05 and .10 (i.e., the interaction between Context and Congruence for the 2x2 ANOVA analyses of RT both when L2 and face trials were ignored). These results are central to interpreting two of our key research questions, namely (i) whether facial expressions prime congruent L1 words more than congruent faces in native Swedish speakers and (ii) whether facial expressions prime congruent L1 words more than congruent L2 words. As such, we used equivalence tests (Lakens et al., 2020) to assess whether our data allowed us to *affirm* the null hypothesis that these effects were 0 (or trivially small). To simplify this analysis into a t-test framework, we computed “priming scores” for each participant in each condition, which represented the average difference in RT between congruent and incongruent trials in each condition. We then tested whether (i) priming scores for face trials could be considered equivalent to priming scores for L1 word trials and (ii) whether priming scores for L1 word trials could be considered equivalent to priming scores for L2 word trials. We set a Cohen’s *d* of 0.2 as our minimally interesting effect size, borrowing from the conventional definition of a small effect. Equivalence tests for paired-samples t-tests were conducted using the TOSTER package in R (Lakens, 2017).

Results revealed non-significant equivalence tests for both the difference in priming scores between face and L1 word trials, *t*(104) = -0.20, *p* = .420, and for the difference in priming scores between L1 word trials and L2 word trials, *t*(104) = -0.26, p = .399. Thus, given that we did not have power to either affirm or reject the null hypothesis for two key research questions (i.e., whether priming was greater for L1 words than both L2 words and faces), we set a new target sample size of 140 usable participants (i.e., 35 additional usable participants). We then collected, analyzed, and reported that sample in the main text.

**Summary of supplemental results.** Supplemental analyses of the initial sample are almost identical in significance to those presented in the main text. One difference is that the trending interaction between Context and Congruence in predicting RT when comparing only face trials and L1 trials in the initial sample became significant in the final sample. This provided evidence of replication of the Nook et al. (2015) results in a Swedish sample. Although the difference in priming between L1 and L2 trials was unclear in this initial sample (null-hypothesis test: *p* = .076; equivalence test *p* = .399), this result remained non-significant in the larger final sample (*p* = .121). As such, if there is a significant difference in priming in this task for L1 vs. L2 words, it is likely very small in size, and we consequently interpret this as a null effect. Finally, the significance of results for proficiency analyses are largely the same between the initial sample and final sample, except for the first analysis testing whether self-reported L2 proficiency affects RTs. Nonetheless, the key interaction between L2 proficiency and Context remains, and the overall conclusions are unchanged.

**II. Figure S1**

*Reaction Time and Sensitivity in Initial Dataset*

****

*Note.* Panel A displays reaction times (RT) for each condition in the initial sample. Results are largely identical to those in the main text. Main effects indicated that participants were faster for congruent than incongruent trials and for both L1 and L2 word trials than face trials. However, the interaction between Congruence and Context was not significant for the ANOVA comparing RT for face trials and L1 word trials. Similarly, the impact of congruence on L1 and L2 words was not significantly different. Panel B displays sensitivity (d’), a signal detection measure of accuracy, for each condition. Participants were less able to correctly identify cue-target matches in the face condition than both the L1 and L2 word conditions. Error bars represent 95% confidence intervals, adjusted for within-subjects comparisons.

**III. Figure S2**

*Relations with Self-Reported L2 Proficiency and LexTALE Scores in Initial Dataset*



*Note.* Panel A presents relations between reaction time (RT) and self-reported L2 proficiency for each condition in the interim sample (not centered). Results are largely similar to those presented in the main text. L2 proficiency moderated the impact of Context on RT such that participants with higher L2 proficiency were faster to respond to L1 and L2 word trials and slower to respond to face trials (although correlations were not significant). Panel B represents relations between RT and objectively assessed L2 proficiency using the LexTALE test. Results show that increasing LexTALE scores correlated with reductions in the difference between Face-Incongruent (light green) and Face-Congruent (dark green) trials. As such, individuals with lower L2 proficiency showed greater tendency for facial expressions to prime (i.e., speed responding) of congruent facial expressions. Panels C and D present relations between sensitivity (d’), a signal-detection measure of accuracy, and both self-reported and LexTALE L2 proficiency. Results for both analyses show that increased L2 proficiency is related to higher overall sensitivity. No interaction with context conditions emerged. Grey shaded regions indicate 95% confidence intervals.