

Supplementary material: statistical models

Fixed effects

Our study involved two IE vowels. Vowel is a fixed effect, since it is expected to systematically affect formant frequencies (FFs). The experimental treatment is Context, i.e. whether the English target item was presented in an English (nonswitch) or Bengali (switch/mixed) context. Our first research question (RQ1—L2 vowels) asks whether L2 vowel quality of proficient bilinguals is affected by the Context of utterance. We hypothesize that the direction of such shift in the mixed condition, if any, is towards a corresponding L1 category. Therefore, we expect the effect of Context on formant frequencies, i.e. the direction of shift, to be moderated by the vowel category, leading to our second research question RQ2—Asymmetry). Thus, the main effect of interest is the interaction term Context*Vowel. This corresponds to the transfer pattern. Our third research question (RQ3—Paradigm) asks whether the language-mixing paradigm, or Task, affects the extent and direction of transfer. If this is the case, we expect Task to interact with the transfer term, giving a three way Task*Context*Vowel interaction. Since we measured FFs at five points in the vowel, Time is expected to affect the FF. However, the precise effect of formant dynamics is expected to depend on the vowel category. Thus, we expect the interaction of Vowel*Time to predict formant frequency. If transfer patterns change across time, then we should expect Time to interact with the transfer term, giving a Context*Vowel*Time interaction.

Random effects structure

Subject and Item were treated as random factors. Given our hypotheses, we included random intercepts for Subject and Item, by-subject random slopes for Vowel (formant targets for vowel categories differ across individuals) and Context (individuals differ in their re-

sponse to the experimental condition), and a by-item random slope for Context (the effect of experimental conditions differ across words).

We tested the significance of our hypothesized explanatory variables using t-values and p-values for individual variables in the models, and tested overall model fit by comparing the full model to a corresponding null model that lacks the variable of interest using ANOVAs. The following subsections summarize and discuss the sequence of models fitted for F1 and F2 respectively, to find the optimal models reported in the main text.

Models for F1: vowel height

First, to test our first and second research questions (L2 vowels and Asymmetry), we fitted a model with vowel*time, and vowel*context as fixed effects. The transfer term (vowel*context) is significant in this model (model coefficient/ unstandardized effect size (β)= -2.330e-01, standard error (SE)= 5.244e-02, t= -4.44, p=0.0002). Comparing this to a null model that lacks the transfer term shows that the full model is significantly better ($\chi^2(2) = 14.42, p = 0.0007$). To confirm whether the effect of context is moderated by vowel identity, we compared a second null model that contained Context as a fixed effect, but lacked the Context*Vowel interaction. The full model was better ($\chi^2(1) = 12.51, p = 0.0004$).

To examine the third research question, Paradigm, we fitted a model with vowel*time and task*vowel*context as fixed effects. This three-way interaction term is significant (β = -1.955e-01, SE= 5.453e-02, t= -3.59, p= 0.0003). We also find a significant effect of task on FFs (β =3.300e-01, SE=2.668e-02, t=12.371, p=2e-16). Comparing this model to a null model that lacks the fixed effect Task ($\chi^2(4) = 572.61, p = 2.2e - 16$) shows that the full model is better. Because Task has an independent effect on F1, and adding interactions increases model parameters, we fitted a second null model with Task as a separate fixed effect without the interaction term to test whether the interaction term is needed. However,

the full model was still a better fit for the data ($\chi^2(3) = 23.82, p = 2.722e - 05$). Next, we tested two other null models to verify that the full model with the three-way interaction term is indeed the optimal model for the data, by replacing the three-way interaction term with: (i) vowel*context + task*context (to test if task only interacts with context, rather than the transfer term vowel*context). The full model is better ($\chi^2(2) = 23.28, p = 8.78e - 06$); (ii) vowel*context + task*vowel (to test if task only interacts with vowel, rather than the transfer term vowel*context). The full model is better ($\chi^2(2) = 13.36, p = 0.001$). This shows that the three-way interaction is the best fit for the data. Because an independent effect of task on FFs is unexpected (c.f. model interpretation below), we wanted to examine if the effect might be local, i.e. decrease over time. We added a task*time interaction term to the full model, which is significant ($\beta = -1.216e-01$, SE= 1.347e-02, t= -9.028, p= 2e-16) and improves model fit ($\chi^2(1) = 81.22, p = 2.2e - 16$).

Next, to examine the dynamics of transfer, we added a three-way interaction between the transfer term (vowel*context) and time. This term has a marginally significant effect ($\beta = -4.604e-02$, SE= 2.692e-02, t= -1.71, p= 0.08), but does not significantly improve model fit compared to a null model lacking the vowel*context*time term ($\chi^2(2) = 3.41, p = 0.18$).

Thus, the optimal model was: $f1 \sim \text{time*vowel} + \text{vowel*context*task} + \text{time*task} + (1 + \text{context}|\text{word}) + (1 + \text{context} + \text{vowel}|\text{subject})$.

Models for F2: vowel backness

First, to test RQ1 (L2 transfer) and RQ2(Asymmetry), we fitted a model with vowel*time, and vowel*context as fixed effects. The transfer term (vowel*context) is significant in this model (standardized effect size (β)= 1.234e-01, standardized error (SE)= 2.055e-02, t= 6.00, p=5.46e-06). Comparing this to a null model that lacks the transfer term shows that the full model is significantly better ($\chi^2(2) = 16.20, p = 0.0003$). To confirm whether the effect

of context is moderated by vowel identity, we compared a second null model that contained Context as a fixed effect, but lacked the Context*Vowel interaction. The full model was better ($\chi^2(1) = 16.02, p = 6.245e - 05$).

To test RQ3 (Paradigm), we fitted a model with vowel*time and task*vowel*context as fixed effects. While task has an independent effect on F2 ($\beta = -8.758e-02$, $SE = 1.756e-02$, $t = -4.98$, $p = 6.18e-07$), the three-way interaction term is not significant ($\beta = 3.349e-02$, $SE = 3.565e-02$, $t = 0.94$, $p = 0.34$). However, there are significant two-way interactions between task*context ($\beta = -1.099e-01$, $SE = 2.483e-02$, $t = -4.427$, $p = 9.66e-06$) and task*vowel ($\beta = -8.983e-02$, $SE = 2.535e-02$, $t = -3.544$, $p = 0.000396$). Comparing this to a null model that lacks the fixed effect Task confirms that the full model is still a better fit for the data ($\chi^2(4) = 572.61, p = 2.2e - 16$), indicating that an optimal model should contain the term Task. To confirm whether Task interacts with any of the other terms, we fitted a model with Task as a separate fixed effect without the interaction term. Comparing this to the full (interaction) model shows that the latter is still a better fit ($\chi^2(3) = 44.8, p = 1.02e - 09$). Next, we tested two other null models to verify that the three-way interaction term is indeed optimal for the data, by replacing the three-way interaction term with: (i) vowel*context + task*context (to test if task only interacts with context, rather than the transfer term vowel*context). The full model is better ($\chi^2(2) = 17.28, p = 0.0001$); (ii) vowel*context + task*vowel (to test if task only interacts with vowel, rather than the transfer term vowel*context). The full model is better ($\chi^2(2) = 28.45, p = 6.633e - 07$). This shows that the three-way interaction is the best fit for the data.

Because an independent effect of task on FFs is unexpected (c.f. model interpretation below), we wanted to see if the effect might be local, i.e. decrease over time. We added a task*time interaction term to the full model, which is significant ($\beta = 6.237e-02$, $SE = 8.876e-03$, $t = 7.027$, $p = 2.23e-12$) and improves model fit ($\chi^2(1) = 49.271, p = 2.23e - 12$).

Next, to examine the dynamics of transfer, we added a three-way interaction between

the transfer term (vowel*context) and time. This term is not significant ($\beta = -6.274\text{e-}03$, $\text{SE} = 1.774\text{e-}02$, $t = -0.35$, $p = 0.72$), and the addition does not significantly improve model fit compared to a null model lacking the vowel*context*time term ($\chi^2(2) = 2.2443$, $p = 0.32$).

Thus, the optimal model is: $f2 \sim \text{vowel*time} + \text{vowel*context*task} + \text{task*time} + (1+\text{context}|\text{word}) + (1+\text{context+vowel}|\text{subject})$.