**Participants.**

* Recruited via prolific
* Balance gender
* Age range? (probably unbounded, as we check for adequate sound-sensitivity in the headphone screener?)
* Make sure they haven’t taken our previous hits (or, minimally, weren’t in Tzeng et al. 2021)
  + Can we exclude participants who took part in LMM21?

**Exclusion criteria.**

LMM21:

* Failed headphone screener twice (we’ll probably keep this)
* did not respond to more than 10% of trials during the exposure or phonetic categorization phase (N/A to us… unless we should add a time limit per trial?)
* were less than 70% accurate in classifying the unambiguous endpoints during the phonetic categorization task (We’re using no unambiguous, just the 7 middle points of a 9-step continuum, so this should maybe stay but be altered? I like the Bushong & Jaeger method, provided it’s done right?)
* Random additional exclusions to have equal number of participants in each counterbalance (let’s just fill up our lists instead?)

Additional:

* Talker identification, somewhere around 80% correct? (we can also track how/whether this tracks with extent of adaptation)

**Statistical analyses**. The primary dependent measures include accuracy (during exposure) and *asi* responses (at test). Listeners must achieve ≥ 80% accuracy *for each item type* presented during exposure (e.g., listeners must meet this criterion for /s/ words, /ʃ/ words, filler words, and nonwords) to be included. Data will be analyzed using generalized linear mixed effects models with the binomial response family as implemented using lme4 (Bates et al., 2015) in R. We have extensive experience with this analysis approach (e.g., Stilp & Theodore, 2019; Theodore & Monto, 2019; Tzeng et al., 2020; Giovannone & Theodore, 2020; Giovannone & Theodore, 2021; Theodore et al., 2020). Models will include the maximal random effect structure licensed by the experimental design (Barr et al., 2013). Convergence issues, rare in our past work and expected to be similarly rare here, will be resolved by iteratively removing random slopes for higher-order interactions until convergence is reached. Models will operate on trial-level data unless otherwise noted. Continuous variables will be entered into models scaled/centered around the mean; categorical variables will be orthogonally coded as appropriate (e.g., sum contrasts will be used for bias; in some cases, sliding contrasts may be used to compare performance across a set of experiments or conditions). Interactions will be tested using simple slopes or by conducting paired comparisons, adjusting alpha to account for multiple comparisons as implemented in the emmeans package (Lenth, 2019). In all cases, we will follow best practices for model selection (e.g., Bates et al., 2015), contrast coding (e.g., Schad et al., 2020), and model reporting (e.g., Meteyard & Davies, 2020).

**Power analyses and sample size**. The primary analyses consist of mixed effects models for trial-level, binary responses. We followed emerging best practices for conducting *a priori* power analyses for mixed effects models (e.g., Green & MacLeod, 2016; Kumle et al., 2021), which entails (1) measuring effect sizes and the covariance structure from an existing model, (2) using those parameters to simulate new data sets with different numbers of participants, trials, and/or effect sizes, (3) analyzing each simulated data set to test for statistical significance of the fixed effect(s) of interest, and (4) calculating power based on the proportion of statistically significant effects relative to all simulations. Evaluating our hypotheses requires power to detect an effect of bias and, in some cases, interactions with bias. We executed our power analyses using the simr package (Green & MacLeod, 2016) based on data from our previous work (Tzeng et al., 2021). This study is well-suited for this purpose because it provides data from a standard lexically guided learning task (experiment 1) and two input manipulations that also used the standard lexically guided learning task (experiments 2 and 3). The results showed a monotonic decrease in the magnitude of the bias effect across experiments (as described previously). Thus, using these data allowed us to estimate effects sizes for the main effect of bias and for bias-by-input interactions. Moreover, each experiment was conducted twice, once with each of two stimulus sets, allowing us to assess convergence of our power analyses across effect size estimates. Based on these analyses, most experiments will include 40 participants in each between-subjects condition. This sample size yields high power (≥ 87%) to detect effect sizes observed in Tzeng et al. (2021) for the main effect of bias, experiment by bias interactions, and a conservative estimate of *half* of the effect size for the main effect of bias. The results of our power analyses converge with others who have used a similar approach with a different data set as the starting point for simulations (Liu & Jaeger, 2019). Experiments that examine individual differences (in Aim 3) will include 80 participants in each between-subjects condition, which yields high power (90%) to detect a moderate effect size (*r* = 0.35) in the proposed correlation analyses. We acknowledge that interaction effects are more difficult to predict *a priori*. We address this challenge by conducting multiple replications in the proposed activities. The sample size for each replication will be determined by *a priori* power analyses following the simulation procedure described here, except that we will use data for the study-to-be-replicated as input for the simulations. This strategy allows us to refine our sample size to be consistent with observed effects sizes, mitigating the concern that our initial estimates may not perfectly capture effect sizes of interest.