**Inclusion criteria for participants.** All participants will be recruited from the Prolific participant pool. Participants will be screened before beginning the experiment, such that only participants meeting the following qualifications will be allowed to take part:

* Using a desktop computer, rather than tablet or mobile device
* Aged 18 or older
* Currently residing in the United States
* Approved on 90% or more of previously taken Prolific experiments
* At least 10 previous Prolific experiments submitted

We will disallow participants from prior experiments within this study, or who took part in any of the studies (reported and unreported) run using Prolific for Luthra, Mechtenberg, and Myers (2021), to take part in future experiments. Additionally, we will pre-screen participants based on their *self-reported* answers to the following questions:

* **Do you have any language related disorders?** (must answer “No”).
* **Are you an English-speaking monolingual, that is, are you fluent only in English? Or are you also fluent in any other language(s)?** (must answer “I only know English”).
* **What is your country of birth?** (must answer “United States”).

We will balance our sample across gender.

**Exclusion criteria for participants.** Among participants allowed to take part in the experiment, we will exclude from analyses those who fail the binaural Huggins-pitch headphone screening task twice. After failing once, participants will be given a second chance. Upon failing a second time, they will be redirected and will not complete the rest of the experiment.

Additionally, we will exclude participants who did not respond within the allotted 4 seconds to more than 10% of trials across the entire experiment. Participants who fail to correctly identify the talker with more than 80% accuracy during exposure will be excluded. Finally, we will exclude any participants for whom the main effect of continuum step is not significant.

We will recruit additional participants (following all inclusion criteria) to replace all excluded participants.

**Statistical analyses**. The primary dependent measures include talker identification accuracy (during exposure) and *asi* responses (at test). 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. 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. 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. We executed our power analyses using the simr package (Green & MacLeod, 2016) based on data from Luthra, Mechtenberg, and Myers (2021). This study is well-suited for this purpose because it shares the same paradigm (lexically guided learning, modified to include talker identification rather than the standard lexical decision task) and implementation (online crowdsourcing via the Prolific pool) with our proposed work.

Based on these analyses, our experiments will include 32 participants in each between-subjects condition. This sample size yields high power (85%) to detect effect sizes observed in Luthra, Mechtenberg, & Myers (2021) 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).