Supplementary Information for Bushong & Jaeger (JEP:LMC)

1 Analyses confirming that participant exclusions do not affect the reported results

The analyses reported in the main text did not exclude participants based on their use of VOT. As we discussed in the main text, theories of subcategorical information maintenance predict that individual participants should use *both* VOT and context in their categorization decisions. Thus, excluding participants on the basis of one of these effects biases the interpretation of results.

However, there are some reasons why excluding participants without a significant VOT effect may be justified. First, participants may not have had sufficiently good audio equipment, thus reducing the signal. Second, participants may not be native speakers of English (despite their responses to our post-experiment survey) and thus not interpret VOT in a native-like way. Third, participants may have simply not paid attention to the task and answered randomly. For these reasons and for the sake of comparability with previous work in our lab, we decided to additionally run the models of the test phase from the main text with subject exclusions based on VOT effects.

1.1 Exclusion procedure

In our past work (Bushong & Jaeger, 2017, 2019), we fit a simple logistic regression model to each individual participant's data predicting responses from VOT. We then excluded participant for whom the effect of VOT was not positive (increasing /t/-responses for increasing VOT) and significant. There are, however, two problems with this approach in the context of the current study. First, in the current work we manipulated the correlation between VOT and context in the stimuli during the exposure phase, such that the extreme /t/ endpoint was never paired with /d/-biasing contexts, and vice versa for for the /d/ endpoint. This introduces a bias towards finding a VOT effect in individual participants by introducing a contrast in responses between the two extreme endpoints.[1] Second, subjects with very extreme VOT effects may be erroneously excluded due to non-convergence of models fit to individual data.

Given these issues, we developed a new approach to participant exclusion that reduces the severity of these issues. Instead of fitting models to individual participant data from the entire experiment, we took the full model fitted to the test phase data for all participants (specifically, the GLMMs presented in the main text of the paper). We then estimated each participant's VOT coefficient and 95% confidence intervals from the random effects of the model. Participants whose

[1] We are grateful to Kaidi Chen and Rachel Theodore for making us aware of this issues.

95% confidence intervals did not include 0 were included. This solves both of the aforementioned problems: (i) the test phase stimuli contain no correlation between VOT and context, thus not biasing inclusion towards participants with strong effects of context and none of VOT; and (ii) the assumption of normally distributed individual differences shrinks the magnitude of participant-specific VOT slopes, reducing issues with overfitting and convergence.

This exclusion procedure resulted in quite stringent exclusions, particularly in Experiment 3 (perhaps not surprising given the inattentiveness of subjects in that experiment). The procedure resulted in the exclusion of 45 (35%) of participants from Experiment 1, 44 (34%) from Experiment 2, and 182 (77%) from Participant 3. We only report results from the test phase since those results test the primary hypotheses presented in the main text.

1.2 Experiment 1

We found significant main effects of VOT ($\hat{\beta} = 9, z = 8.6, p < .001$) and subsequent context ($\hat{\beta} = 2.11, z = 4.54, p < .001$). The interaction between right context and exposure group was marginally significant ($\hat{\beta} = -.68, z = -1.75, p = .08$). Trial also interacted significantly with both context ($\hat{\beta} = -.27, z = -2.67, p < .01$) and VOT ($\hat{\beta} = .91, z = 4.71, p < .001$). There was also a significant interaction between VOT and exposure group ($\hat{\beta} = 6.24, z = 4.88, p < .001$). These results are qualitatively very similar to the model in the main text, though the context by group interaction coefficient is marginal (compared to significant in the main text no-exclusions analysis), and the group by VOT interaction is significant (compared to marginal in the main text analysis).

1.3 Experiment 2

We again found a significant main effect of VOT ($\hat{\beta} = 7.88, z = 7.24, p < .001$). The effect of subsequent context was only marginal ($\hat{\beta} = 1.04, z = 1.73, p = .08$). The interaction between right context and exposure group was marginally significant ($\hat{\beta} = -.62, z = -1.78, p = .08$). Trial interacted significantly with exposure group ($\hat{\beta} = -.28, z = -2.15, p = .03$). There were no other significant main effects or interactions. These results largely replicate the main text analysis of Experiment 2, excepting that the context effect is marginal in the exclusions analysis, and there is a significant trial by group interaction.

1.4 Experiment 3

We found significant main effects of VOT ($\hat{\beta} = 4.25, z = 3.74, p < .001$) and subsequent context ($\hat{\beta} = 2.09, z = 2.87, p < .01$). The interaction between right context and exposure group was marginally significant ($\hat{\beta} = -.66, z = -1.64, p = .1$). Trial also interacted significantly with VOT ($\hat{\beta} = .69, z = 2.85, p < .01$). The main results are qualitatively similar to the model in the main text, though the context by group interaction coefficient is marginal (compared to significant in the main text no-exclusions analysis). Some of the other effects were discordant—in particular, there was a significant VOT by trial effect not present in the no-exclusions analysis; and no main effect of group or interaction between group and trial, as opposed to the no-exclusions analysis. This is likely due to the small number of participants in Experiment 3 displaying significant VOT effects under the revised–very stringent—inclusion criterion.

As also discussed in footnote 3 in the main text, shrinkage of by-participant random slope estimates that is inherent to GLMMs can make the approach we committed to here (prior to conducting Experiment 3) overly conservative. Section §5 in the SI presents an alternative, and arguably more informative, approach to participant exclusions for experiments like Experiment 3, in which a large proportion of participants (though *not* as many as 77%) were inattentive or non-cooperative.

2 GAMM analyses of non-linear changes across the experiment

We conducted additional analyses that analyze potential changes in the effects of VOT and subsequent context through the exposure and test phases. These analyses use generalized additive mixed-effects models (GAMMs) with a binomial (logit) link function. We originally considered these analyses as our primary approach but came to realize that properties of our design—and the amount of available data—made it difficult to reliably estimate all relevant effects of interest. This meant that we had to limit the predictors included in the GAMM analysis. Combined with technical limitations of existing implementations of GAMMs (specifically, the difficulty of assessing higher-order interactions between smooths and multiple categorical predictors), this motivated our decision to present GLMM analyses in the main text. Here we present the additional GAMM analyses, as they detect a number of additional patterns that—if confirmed by future work—would further speak to how recent experience comes to affect listeners' maintenance of subcategorical information.

We used the `bam` function of the `mgcv` library (Wood, 2011) in `R` (R Core Team, 2016) to fit the GAMMs. Except where explicitly mentioned otherwise, this analysis follows the same approach as in Bushong and Jaeger (2019), reducing researchers' degrees of freedom. We analyzed the log-odds of /t/-responses as a linear function of context (sum-coded: 0.5 = tent-biasing, -0.5 = dent-biasing), exposure group (sum-coded: 0.5 = low-informativity exposure, -0.5 = high-informativity exposure), and their interaction.

We included several non-linear terms in the model, including a main effect of VOT and VOT's interaction with context and group (including the three-way interaction). To model changes across trials, we included smooth (non-parametric) terms to capture (1) the two-way interaction between VOT and trial, (2) the two-way interaction between context and trial, and (3) the three-way interaction between trial ($\log_2$-transformed), context, and exposure group. Since available GAMM implementations do not accommodate a three-way smooth interaction with multiple categorical predictors, we created a dummy-coded context-exposure group variable following our approach in Bushong and Jaeger (2019). The three-way interaction was thus represented by a reference smooth (tent-biasing, high-informativity group) and three difference smooths representing the three other context-group combinations. Finally, the analysis included by-item random intercepts and slopes for all effects and by-participant random intercepts and random slopes for all effects except those including group (which was manipulated between participants).[2]

For the analysis of responses during the test phase, our primary prediction corresponding to the results in the main text is a (negative) interaction between group and context, indicating a reduced

[2]As of yet, `mgcv::bam()` does not support correlations between random effects. This means that the random intercepts and slopes are estimated under the assumption of 0 correlation between random effects.

effect of context in the low-informativity group. The additional time-course analyses also allow us to interrogate the nature of this interaction: do the high- and low-informativity groups differ in their context effect from the beginning of the experiment, as would be expected if the exposure phase is the driver of the reduction of the context effect in the low-informativity group? Secondarily, we can also ask how the context effect changes for the low-informativity group. If informativity of context causes listeners to modulate their maintenance of subcategorical information, then we would expect the low-informativity group's context effect to potentially increase across test as they learn that context is again informative to word categorization. Any effects of this nature are likely to be relatively subtle, due to the small number of trials in test compared to exposure, and the low correlation between cues in the test phase.

In addition to the critical analysis of responses during the test phase, we fit GAMMs to the exposure phase data. Since the low-informativity group did not contain a manipulation of context, we fit separate GAMMs to the low- and high-informativity groups. For both analyses, we removed group and its interactions from the GAMM. For the low-informativity group, we additionally removed context and its interactions from the GAMM (since there only was one context condition for the low-informativity group).

The GAMM analyses of the exposure phase of the high-informativity group assesses our secondary research question, whether the high-informativity group showed a context effect from the beginning of the exposure phase—as would be expected if subcategorical information maintenance is typical in spoken word recognition, rather than a response to the specific task participants are asked to complete in this type of behavioral experiment. The same analysis also sheds light on whether the effect of context remained constant across the exposure phase, as would be expected if our design was successful in avoiding cue conflict (Bushong & Jaeger, 2019).

The GAMM analyses of the exposure phase of the low-informativity group was purely included for visual comparison (see below).

2.1 Experiment 1

Figure S1 shows the GAMM-predicted /t/-responses across trials in the exposure and test phases. Figure S2 visualizes the GAMM-predicted context effect across trials.

2.1.1 Exposure phase

In line with Figure S1 (top left panel) and the GLMM results reported in the main text, the GAMM analysis of the high-informativity group found effects of VOT ($edf = 2.7, \chi^2 = 75.3, p < .001$) and subsequent context ($\hat{\beta} = 2.59, z = 6.14, p < .001$). As evident in Figure S1 (top left), subsequent context affected responses already on the very first trial of the exposure phase (estimate = 3.47, CIs = $[2.93, 4.02]$). This effect then declined numerically over trials, but not significantly; neither trial smooth by context was significant ($ps > .1$). There were no other significant effects.

2.1.2 Test phase

In line with Figures S1 (right panels) and S2, and the GLMM results reported in the main text, the GAMM analysis found effects of VOT ($edf = 3.86, \chi^2 = 273.45, p < .001$), subsequent

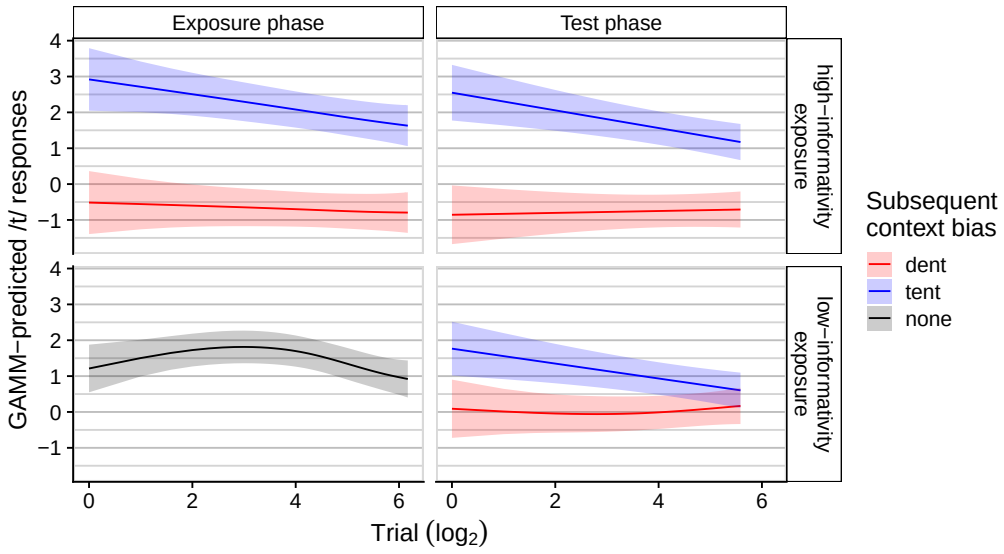

Figure S1: GAMM-predicted log-odds of /t/ responses over the exposure and test phase of Experiment 1, assuming a VOT value in the center of our tested continuum. Bands are 95% confidence intervals. Note that separate GAMMs were fit to the exposure and test phase, so predicted context effects are not expected to precisely line up between phases.

context ($\hat{\beta} = 1.57, z = 7.13, p < .001$), and the interaction between context and exposure group ($\hat{\beta} = -1.48, z = -3.56, p < .001$).

The context effect changed over the course of the test phase, and the pattern of these changes differed between the two exposure groups. For the high-informativity group, the reference smooth representing tent-biasing contexts was significant, with /t/-responses declining over the course of the test phase ($edf = 1, \chi^2 = 6.91, p = .009$); the difference smooth comparing dent- to the tent-biasing contexts was significant, suggesting that /t/-responses in the dent-biasing condition did not change ($edf = 1, \chi^2 = 6.88, p = .009$). Together this suggests that the context effect—the difference between the red and blue lines in Figure S1 (top right)—decreased across trials (see also green line in Figure S2, right). This decrease in the effect of right context across test trials replicates the pattern observed during exposure as well as previous findings in the presence of cue conflict (Bushong & Jaeger, 2019).

For the low-informativity group, the context effect also changed significantly across trials but did so in a non-linear fashion. Like the high-informativity group, /t/-responses in the tent-biasing condition decreased linearly over trials ($edf = 1, \chi^2 = 6.09, p = .01$). The /t/-responses in the dent-biasing condition were significantly different ($edf = 2.19, \chi^2 = 2.71, p = .02$), staying flat for the majority of the test phase, then decreasing toward the end. Qualitatively, however, the results are largely concordant; in both groups, the effect of context decreases across test.

At the very beginning of the test phase, the two exposure groups exhibit different context effects: whereas participants in the high-informativity group exhibited significant effects of right context even on the first trial of the test phase (estimate of context effect at Trial 1 = 4.17, CIs = $[2.88, 5.45]$), the context effect for participants in the low-informativity group was *not* significantly different from zero at the first trial (estimate = $.93$, CIs = $[-0.37, 2.23]$).

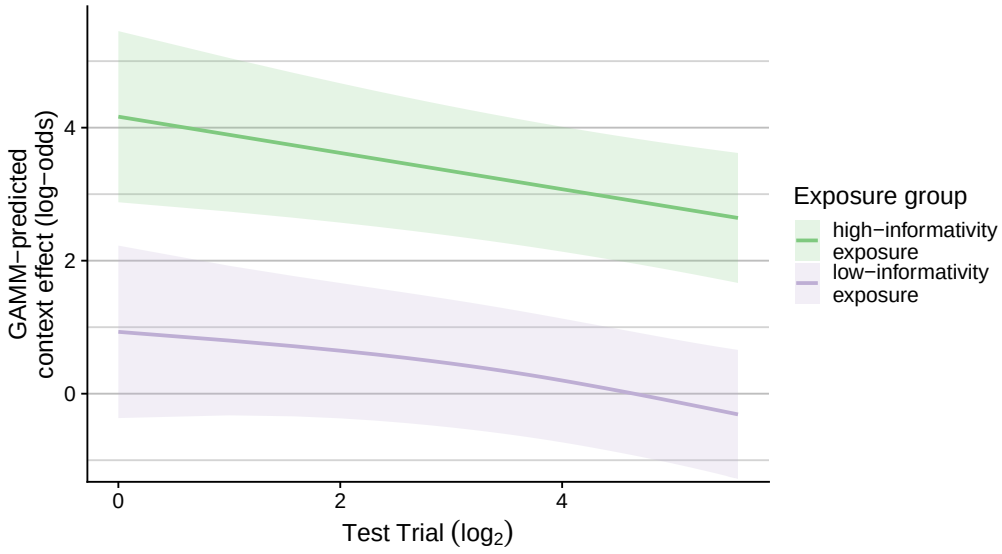

Figure S2: GAMM-predicted differences in context effects between exposure groups over trials in Experiment 1 (difference between the red and blue lines from Figure S1). Shaded regions indicate 95% confidence intervals.

2.2 Experiment 2

Figure S3 shows the GAMM-predicted /t/-responses over the exposure phase (for the high-informativity group) and test phase. Figure S4 visualizes the GAMM-predicted context effect across trials.

2.2.1 Exposure phase

The results replicated Experiment 1: in line with Figure S3 (left), we found effects of VOT ($edf = 2.2, \chi^2 = 109.18, p < .001$) and subsequent context ($\hat{\beta} = 2.14, z = 4.14, p < .001$).

Like in Experiment 1, the context effect was already present on the very first trial of the exposure phase (estimate = 5.08, CIs = $[3.71, 6.45]$; see Figure S3). Like Experiment 1, the context effect did not change significantly over the course of the exposure phase, with neither context condition smooth reaching significance ($ps > 0.19$); though, like Experiment 1, there appeared to be a numerical trend downward in the size of the context effect.

2.2.2 Test phase

The results of the test phase also largely replicated Experiment 1. We again found significant main effects of VOT ($edf = 3.68, \chi^2 = 121.85, p < .001$) and context ($\hat{\beta} = 2.2, z = 10.82, p < .001$). Echoing the GAMM results from the main text, the interaction between context and exposure group was marginally significant ($\hat{\beta} = -0.7, z = -1.82, p = .068$).

Next we turn to the changes of the context effect across the test phase, visualized in Figure S4.

The reference smooth for the trial by group by context interaction was marginally significant ($edf = 2.32, \chi^2 = 2.69, p = .053$), such that /t/-responses increased over time in the high-informativity group. The difference smooth for the dent-biasing condition in the high-informativity group was not significant ($edf = 1, \chi^2 = 0.96, p = .33$), suggesting overall that the context effect

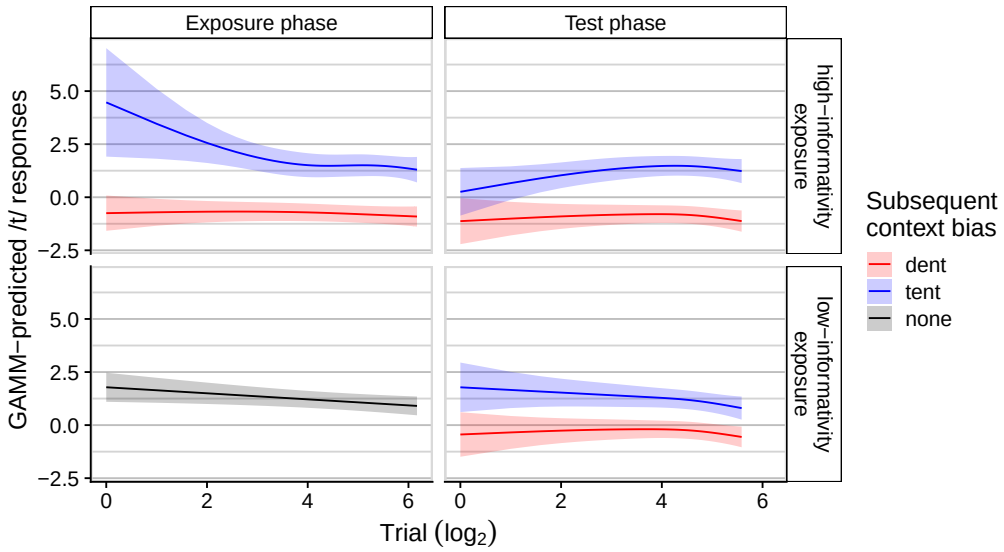

Figure S3: GAMM-predicted log-odds of /t/ responses over the exposure and test phase by context condition for Experiment 2, assuming a VOT value in the center of our tested continuum. Bands are 95% confidence intervals.

did not change significantly over time (as shown in Figure S4, top right). In the low-informativity group, the reference smooth was marginally significant ($edf = 1, \chi^2 = 2.89, p = .09$), such that /t/-responses slightly increased over trials. The difference smooth for the dent-biasing condition was not significant ($p = .39$). Together, this suggests that there was very little change in the context effect over trials in the low-informativity group.

The timecourse results of Experiment 2 stand in apparent contrast to Experiment 1: the context effect did not clearly differ between the two groups at the onset of the test phase, and there were no significant changes in the context effect across trials in either experiment. Although the context effect was marginally smaller in the low-informativity group, it was not indistinguishable from zero at the onset of testing (low-informativity group, estimate of context effect = 1.86, CIs = $[0.22, 3.49]$; high-informativity group estimate = 1.69, CIs = $[0.04, 3.34]$).

2.3 Experiment 3

2.3.1 Exposure phase

The results replicated Experiments 1 and 2 (and the GLMM analyses of Experiment 3 reported in the main text): in line with Figure S5 (left), we found effects of VOT ($edf = 1, \chi^2 = 21.86, p < .001$) and subsequent context ($\hat{\beta} = 1.43, z = 6.05, p < .001$).

Like in Experiments 1 and 2, the context effect was already present on the very first trial of the exposure phase (estimated effect at trial 1 = 2.6, CIs = $[1.94, 3.264]$). Also like Experiments 1-2, the context effect did not change significantly over the course of the exposure phase, with neither context condition smooth reaching significance ($ps > 0.57$).

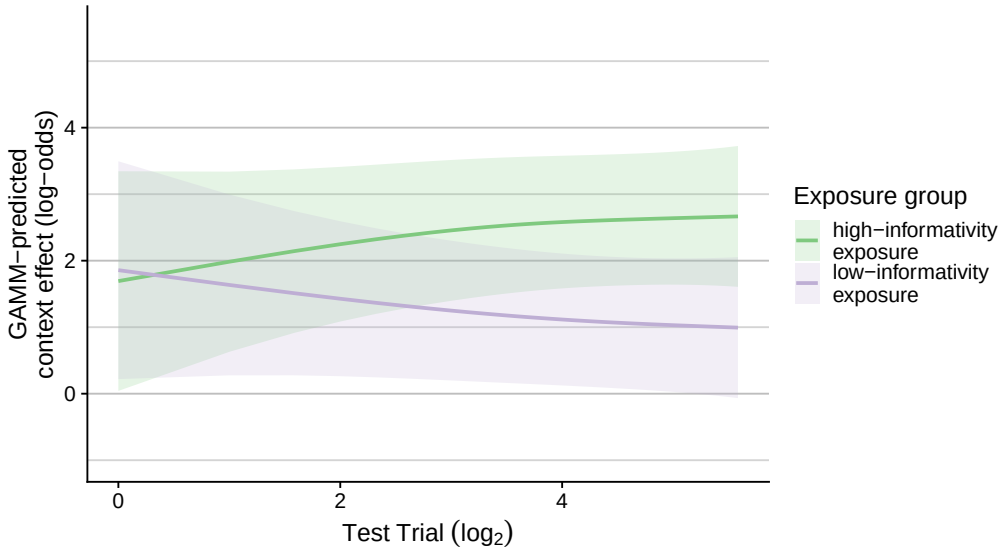

Figure S4: GAMM-predicted differences in context effects between exposure groups over trials in Experiment 2 (difference between the red and blue lines from Figure S3). Shaded regions indicate 95% confidence intervals.

2.3.2 Test phase

The results of the test phase also largely replicated Experiment 1. We again found significant main effects of VOT ($edf = 3.34, \chi^2 = 40.25, p < .01$) and context ($\hat{\beta} = .89, z = 6.67, p < .001$). Critically, as in Experiment 1, the interaction between context and exposure group was significant ($\hat{\beta} = -0.47, z = -2.13, p = .033$).

Finally, we turn to the changes of the context effect across the test phase, visualized in Figure S6. The reference smooth for the trial by group by right context interaction was not significant ($edf = 1, \chi^2 = .19, p = .66$), such that /t/-responses in the tent-biasing condition in the high-informativity group changed little over time. The difference smooth for the dent-biasing condition in the high-informativity group was not significant ($edf = 1, \chi^2 = .8, p = .37$), suggesting overall that the context effect changed little over time (as shown in Figure S6, top right). In the low-informativity group, the reference smooth was not significant ($edf = 1, \chi^2 = .45, p = .5$). The difference smooth for the dent-biasing condition was marginally significant ($edf = 1, \chi^2 = 3.04, p = .08$), such that while /t/-responses in the tent-biasing condition changed little, responses in the dent-biasing condition slightly decreased over trials. Together, this pattern suggests that the context effect grew marginally larger throughout the test phase for the low-informativity group.

Like Experiment 1 (and in contrast to Experiment 2), the difference in context effect between exposure groups was apparent from the beginning of the test phase. On the first trial, the high-informativity group's context estimate (1.68, CIs = $[0.52, 2.84]$) was significantly higher than the low-informativity group ($-.156$, CIs = $[-1.28, 1]$), which was indistinguishable from zero.

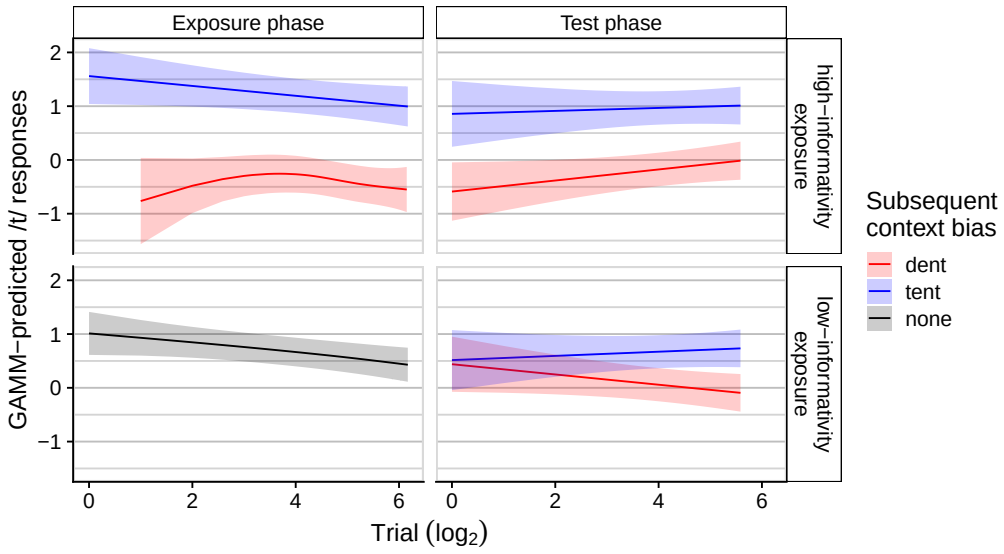

Figure S5: GAMM-predicted log-odds of /t/ responses over the exposure and test phase by context condition for Experiment 3, assuming a VOT value in the center of our tested continuum. Bands are 95% confidence intervals. Due to fully randomizing lists, no list in Experiment 3 had a dent-biasing sentence on the first trial of exposure; thus, data is missing for trial 1 of exposure in that condition.

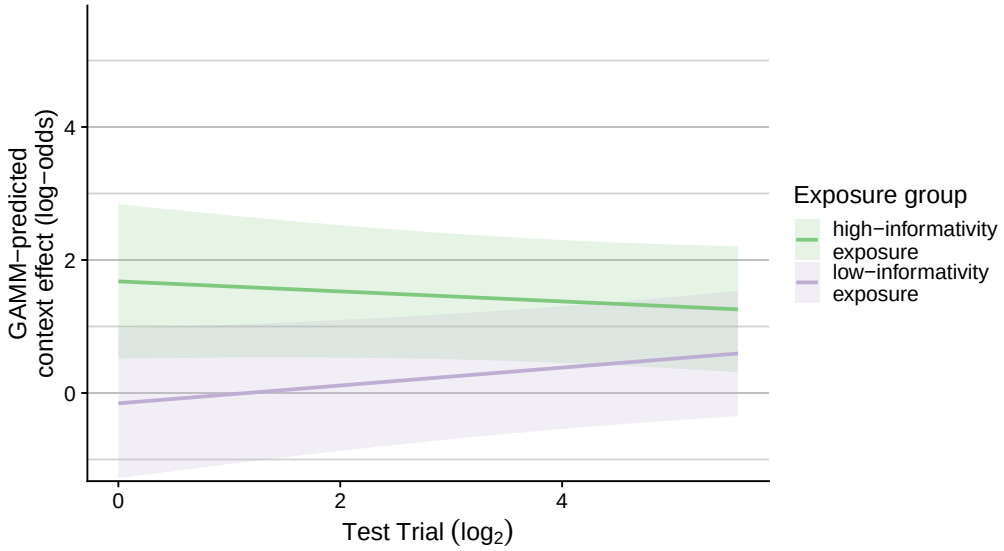

Figure S6: GAMM-predicted differences in context effects between exposure groups over trials in Experiment 3 (difference between the red and blue lines from Figure S5). Shaded regions indicate 95% confidence intervals.

2.4 Discussion of GAMM analyses

The results of the GAMM analyses are qualitatively similar to the results presented in the main text with respect to our major predictions, finding the same pattern of effects of context, VOT, and

context by group interaction.

One motivation for conducting the GAMM analysis is that several features of our design may induce non-linear trial effects. First, if listeners continue to monitor the informativity of subsequent context not only when first exposed to the new input but throughout the entire experiment, we would expect to see that the two groups become more similar over the course of the test phase (since both groups receive the same input during test). Specifically, participants in the low-informativity group might begin to show larger context effects as the test phase progresses, as subsequent context is informative during the test phase (unlike during low-informativity exposure). Complicating the predicted changes across the test phase, however, is the presence of trials with cue conflict during the test phase (absent during exposure). If previous findings about the effect of cue conflict replicate (Bushong & Jaeger, 2019), this is expected to lead to decreasing effects of subsequent context for *both* groups as the test phase progresses. Taken together, these considerations predict that we might see nonlinear changes in the context effect across the test phase, including monotonically decreasing effects for the high-informative group and potentially non-monotonic effects for the low-informativity group.

The GAMM results are inconsistent on this point. The context effect in the low-informativity group decreases (primarily at the end of the test phase) in Experiment 1, stays the same in Experiment 2, and marginally increases in Experiment 3. We likely need more trials and participants in order to fully understand the timecourse of effects in this paradigm, if any exist.

3 Comparison of Experiments 1 & 2

While we found significant effects of context, VOT, and the context by group interaction in both Experiments 1 and 2, there were some numeric differences in the magnitude of these effects with potential theoretical implications. In order to test whether these effects were in fact significantly different between experiments, we decided to conduct a post-hoc combined analysis pooling data from both experiments. This analysis is somewhat pre-empted by the fact that we ran a third experiment that fixed design flaws in Experiment 2 and increased the number of subjects. However, we include the combined analysis in its original form in these materials.

In order to test the relationship between the group and context effects across the experiments, we fit a mixed-effects logistic regression model to data combined from both experiments, allowing experiment (sum coded: 0.5 = Experiment 2, -0.5 = Experiment 1) to interact with context, group, VOT, and trial (coded identically to the analyses in the main text). We included the maximal random effects structure allowing convergence, which was a random intercept and slopes for context and VOT by subject, and a random intercept and slopes by context and experiment by item.

3.1 Results

Figure S7 shows the results of the combined analysis of Experiments 1 and 2. We found main effects of VOT ($\hat{\beta} = 4.94, z = 11.69, p < .001$) and subsequent context ($\hat{\beta} = 2.28, z = 7.35, p < .001$). There was also a significant interaction between group and context ($\hat{\beta} = -1.65, z = -6.99, p < .001$) such that the context effect was larger for the high-informativity group than the low-informativity group. There was an interaction between group and VOT ($\hat{\beta} = 0.97, z = 2.834, p = .005$), such that the VOT effect was larger in the low-informativity group. Context and

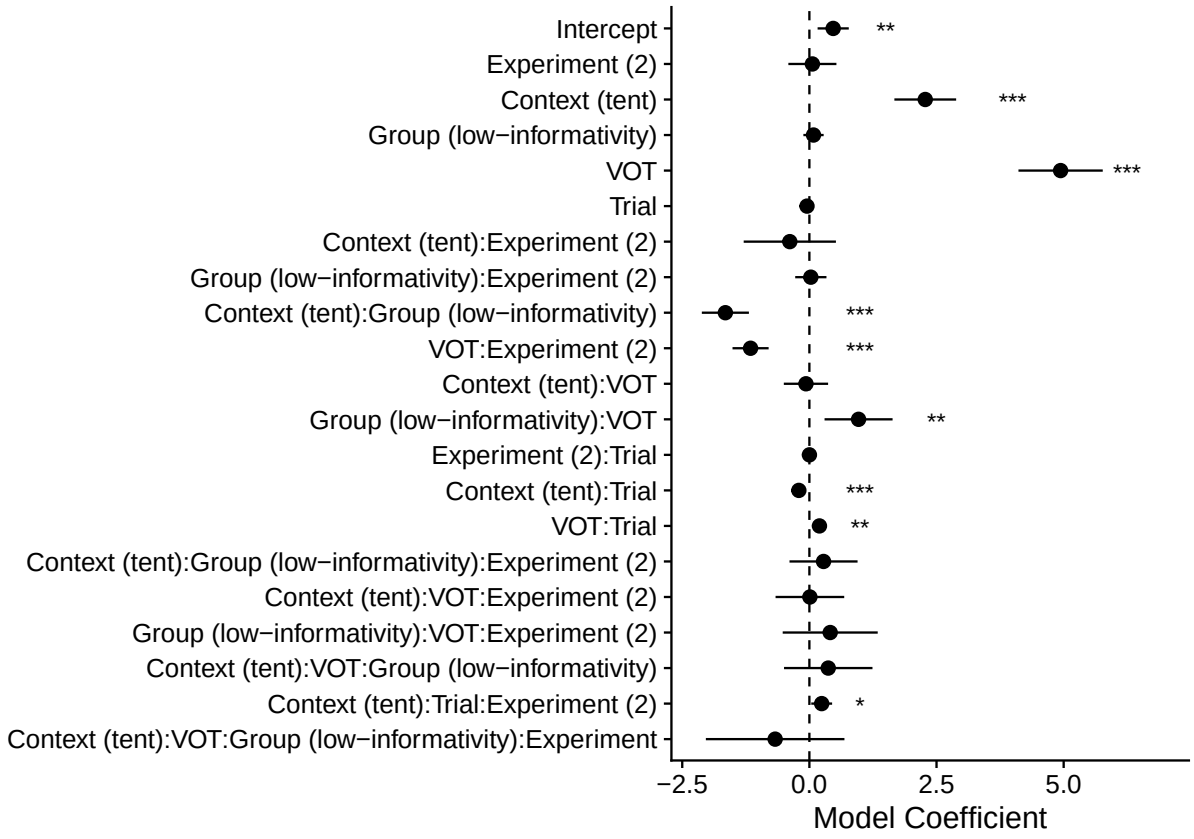

Figure S7: Model coefficients from the combined-experiment GLMM. Error bars are 95% confidence intervals.

VOT also both interacted with trial (context by trial: $\hat{\beta} = -0.21, z = -3.34, p < .001$; VOT by trial: $\hat{\beta} = 0.2, z = 3.03, p = .002$), such that the context effect got smaller and the VOT effect larger over the course of the test phase. In terms of differences between experiments, there was an interaction between VOT and experiment ($\hat{\beta} = -1.16, z = -6.38, p < .001$) such that the effect of VOT was smaller in Experiment 2, and a three-way interaction between experiment, context, and trial ($\hat{\beta} = 0.24, z = 2.32, p = .02$), such that the context effect was larger in Experiment 2 as the test phase unfolded. The three-way interaction between experiment, group, and context did not reach significance ($\hat{\beta} = 0.28, z = 0.82, p = .41$).

3.2 Discussion

When pooling data between Experiments 1 and 2, we replicated the major results reported in the main text: significant effects of context, VOT, and a context by group interaction in the test phase. Our between-experiment comparisons showed that there were two differences between Experiments 1 and 2. First, the effect of VOT was significantly smaller in Experiment 2; and second, the size of the context effect was larger in Experiment 2 across trials compared to Experiment 1. These results suggest that in general, participants in Experiment 2 may have relied more on context and less on VOT in their categorization responses. Critically, there was no three-way interaction be-

tween experiment, context, and group, suggesting that the context by group interaction was similar in both experiments.

4 Repeated exposure does not seem to reduce context effect during test

As noted in the main text, one concern about the design of Experiments 1-2 is that half of the lists contained repeated items between the exposure and test phases. However, due to the nature of our manipulation, that means that half of the high-informativity exposure participants received repeated items between exposure and test, but no low-informativity exposure participants received repeated exposure ("items" were repeated, but never identical sentences). It is thus possible that subjects in the high-informativity condition who heard repeated items exhibited larger effects of context, driving the context by exposure group interaction. We alleviated this concern by eliminating repetition in Experiment 3, but we wanted to estimate how much of a role this concern may have played in Experiments 1-2. We thus conducted an additional analysis of the high-informativity exposure group during the test phase of the experiment, asking whether item repetition interacted with context. We fit a GLMM of the form `response == /t/ ~ Context x Repetition`. We also included a random intercept and slope for context by subject and item, and random slope by repetition and the context by repetition interaction by item. In both experiments, the context by repetition interaction was not significant (Experiment 1: $\hat{\beta} = -.34, z = -.55, p = .58$; Experiment 2: $\hat{\beta} = -.08, z = -.15, p = .88$), suggesting that the effect of context was similar regardless of whether or not items were repeated between exposure and test.

5 Experiment 3 analyses confirming robustness of results to exclusion

As noted in the main text, participants in Experiment 3 were clearly less attentive than in Experiment 2 (and likely Experiment 1). Although qualitatively the results replicated Experiments 1 and 2, we were concerned since that data contained a high proportion of subjects who were likely paying little (if any) attention to the task. Thus, we decided to conduct additional analyses successively excluding participants according to their mean accuracy on filler trials. For each unique mean accuracy score, we excluded all participants below that score; then we fitted the exposure and test GLMMs to the resulting data. This gave us a total of 35 data subsets/models ranging from including all 237 subjects (mean accuracy of at least 32%) to 42 subjects (100% accuracy). Figure S8 shows the coefficient estimates and 95% confidence intervals from the models for each of the exclusion intervals.

As reported in the main text, we found significant main effects of context and VOT, and a significant interaction between context and group prior to any additional exclusions. Figure S8 shows that significance of these effects were stable across exclusion thresholds. Indeed, the coefficient estimates for the effects of context, VOT, and the context by group interaction all increased as more and more inattentive participants were excluded. Indeed, when we compare model coefficients from Experiment 3 with maximum exclusions (100% accuracy) to the first two experiments,

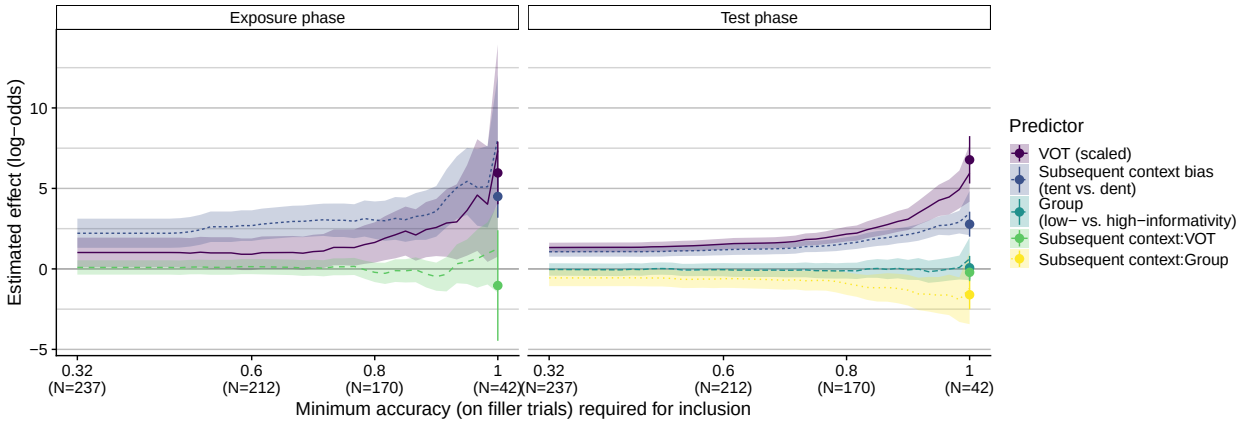

Figure S8: Coefficient estimates with 95% confidence intervals (y-axis) for GLMMs fitted to different subsets of the participants, successively excluding participants who did not achieve a minimum response accuracy on filler trials(x-axis). The points on the right hand-side of each panel indicate coefficient estimates with 95% confidence intervals from Experiment 1 for comparison. Note how coefficient estimates in Experiment 3 tended to converge towards the estimates obtained in Experiment 1, as more and more inattentive participants were excluded.

the estimates are quite similar. Table 1 shows the model coefficients for the test phase of Experiments 1 and 2, and Experiment 3 at three exclusion levels. Notably, the correlation between participant-specific random effects also flips from positive to negative with increasing exclusion rates, mirroring the negative correlations found in Experiments 1-2. Figures S9 and S10 replicate Figures 7-8 from the main text for an 85% exclusion threshold.

One point of interest worth noting is the correlations between random effects (see Figure S10.) As noted in the main text, the correlation between participant-specific random effects was negative across Experiments 1-2 for both the exposure and the test phase. This exact pattern does not hold as strongly in Experiment 3, even with exclusions: while the correlations in both the exposure and test phase trend more negative with higher exclusion rates, the exposure phase correlations do not converge to the estimates from Experiments 1-2. Indeed, even within Experiments 1 and 2, the exposure phase correlation estimates are less strongly negative than the test phase correlations. What exactly to make of this pattern is not entirely clear, though it may point to differences in the cue distributions between exposure and test for the high-informativity group. Recall that in the exposure phase, the high-informativity exposure group is not exposed to strongly clashing VOT/context combinations (following results from Bushong & Jaeger, 2019, which showed that reducing cue conflict results in larger context effects in responses.) Perhaps the cue conflict introduced in the test phase of the experiment drives participants to favor one cue or the other, producing the strong negative correlations observed across experiments. Exploring these possibilities is a task for future work.

Predictor	Experiment 1	Experiment 2	Experiment 3 ≥32% acc. n = 237	Experiment 3 ≥85% acc. n = 150	Experiment 3 100% acc. n = 42
VOT	6.78^{***}	4.94^{***}	1.89^{***}	2.62^{***}	5.92^{***}
Context	2.78^{***}	2.3^{***}	$.93^{***}$	1.89^{***}	3.43^{***}
Context:Group	-1.61^{*}	$-.81^{+}$	$-.55^{*}$	-1.2^{**}	-1.23

*** p $<$ .001 ** p $<$ .01 * p $<$ .05 $^{+}$ p $<$ .1

Table 1: Coefficient estimates and significance for Experiments 1-3 with varying levels of participant exclusion for Experiment 3. In Experiment 2, average participant accuracy was 98%, with all but one participant having $\geq 85\%$ accuracy.

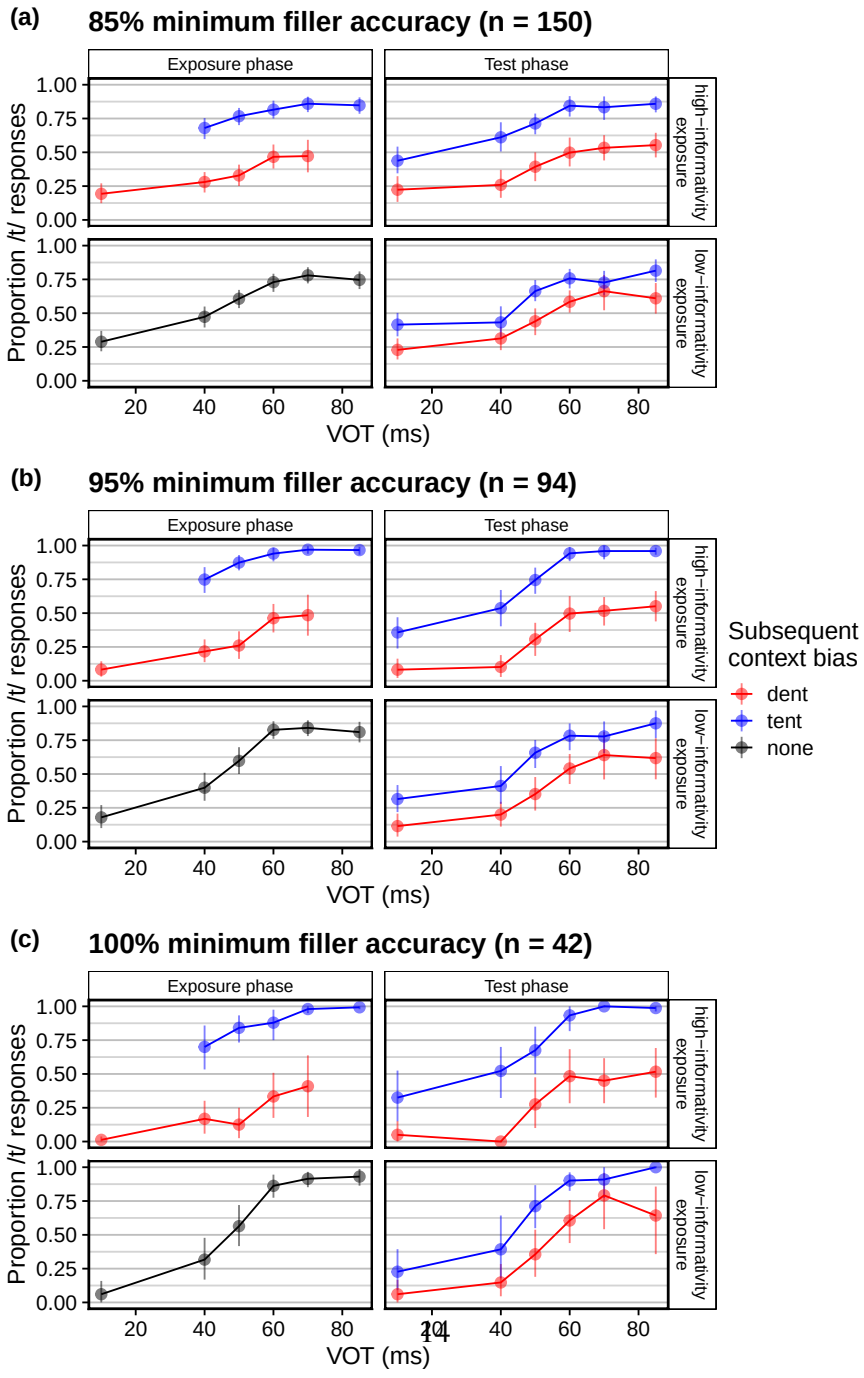

Figure S9: Figure 7 from the main text, repeated for filler accuracy thresholds between 85-100%.

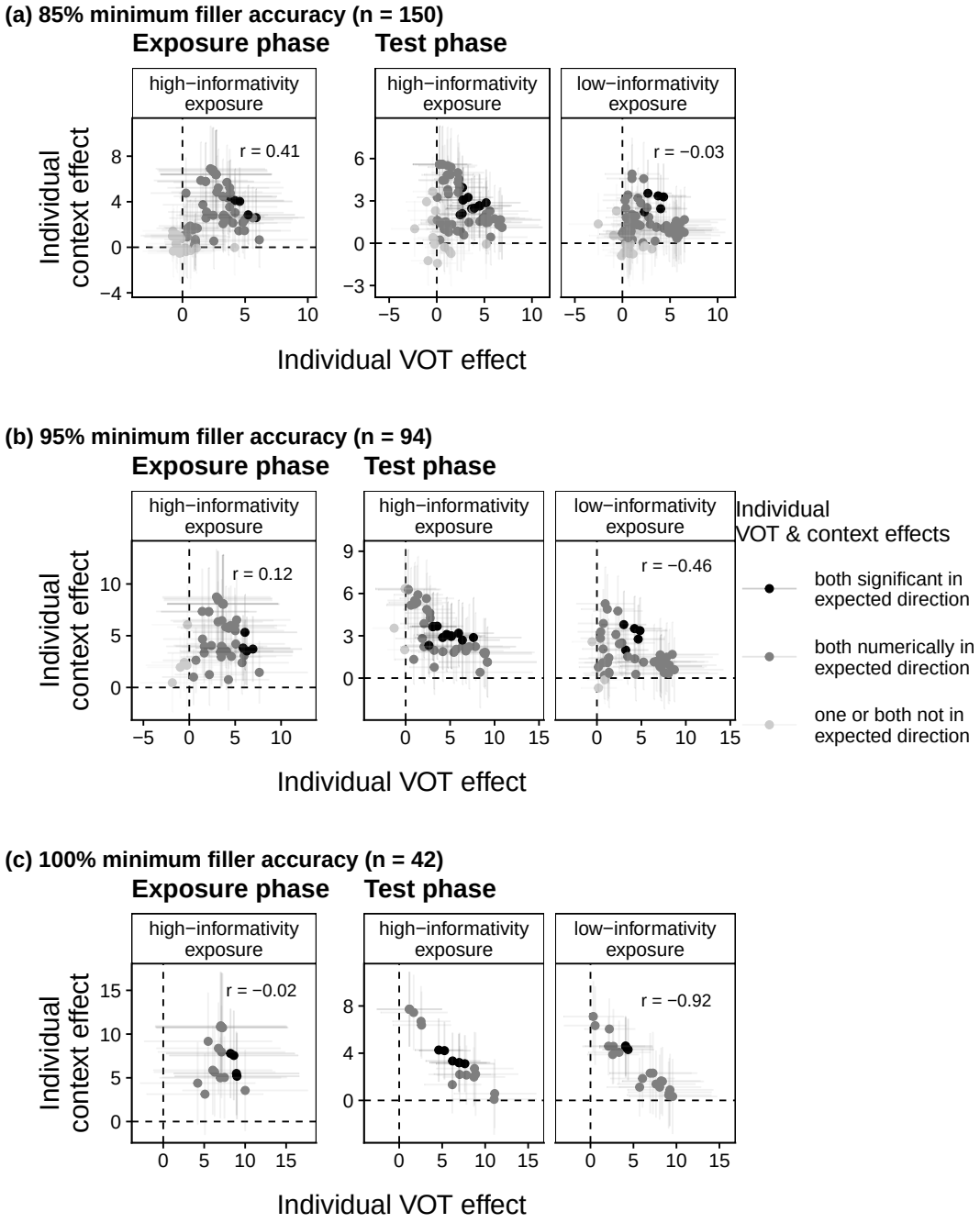

Figure S10: Figure 8 from the main text, repeated for filler accuracy thresholds between 85-100%. Comparison to Figure 8 in the main text shows that the number of participants with significant effects for both VOT and context has increased after exclusions. This is a consequence of reduced shrinkage of the participant-specific effects towards 0: as more inattentive participants are excluded (who have VOT and context effects centered around zero), the population mean that participant-specific effects are shrunk towards increases for both VOT and context effects.

5.1 Similar issues with data quality on Mechanical Turk

Even prior to conducting Experiment 3, we had worried about conducting experiments using Amazon's Mechanical Turk given recent reports of reduced data quality on this platform, compared to other online platforms Peer, Brandimarte, Samat, and Acquisti (2017). After weighing the pros and cons, we decided to complete the series of experiments on Mechanical Turk, with the (likely ill-guided) goal to facilitate comparison across experiments.

Here we summarize additional evidence from our lab that points to increasing issues with data quality on Mechanical Turk over recent years. The second author has been involved in three large-scale projects on speech perception that suggested a drastic drop in the Mechanical Turk's data quality some time after 2018. Each of these projects on its own could be due to random variability in the data, but taken together may be indicative of a larger issue with the Amazon Mechanical Turk crowdsourcing platform. We summarize the three projects briefly, as they form the background to how we approached the interpretation of Experiment 3.

In 2018, Xie, Liu, and Jaeger (2021) conducted a large-scale direct replication (N = 320 participants) of their own experiment on accent adaptation that was originally conducted in 2015 (also N = 320). Both experiments employed Mechanical Turk, the same recruitment criteria, the same stimuli, the same procedure, and the same instructions. While the replication confirmed all qualitative results, participants in the 2018 experiment exhibited overall lower accuracy (across all conditions), substantially more variability, and (thus) reduced effect sizes (see discussion in Xie et al., 2021). In a second project, the same trend—qualitatively replicating differences between conditions but overall reduced accuracy—was observed for a series of five experiments on perceptual recalibration, conducted between 2015 and 2018 (see Figure 2 in Liu & Jaeger, 2019). Finally, a third ongoing project conducted experiments on causal inference during speech perception between 2019-2021, and again found decreasing data quality with Mechanical Turk (as assessed through catch trials). Once the project switched to the Prolific crowdsourcing platform all of these reliability issues subsided.

References

Bushong, W., & Jaeger, T. F. (2017). Maintenance of perceptual information in speech perception. In *Proceedings of the Thirty-ninth Annual Conference of the Cognitive Science Society* (pp. 186–181).

Bushong, W., & Jaeger, T. F. (2019). Dynamic re-weighting of acoustic and contextual cues in spoken word recognition. *The Journal of the Acoustical Society of America*, *146*(2), EL135–EL140.

Liu, L., & Jaeger, T. F. (2019). Talker-specific pronunciation or speech error? discounting (or not) atypical pronunciations during speech perception. *Journal of experimental psychology. Human perception and performance*, *45*, 1562-1588. doi: 10.1037/xhp0000693

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3-36.

Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology: General*, *150*(11), e22.