Cognition 232 (2023) 105320

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Do individual differences in lexical reliance reflect states or traits?

Nikole Giovannone [a,b], Rachel M. Theodore [a,b,*]

[a] Department of Speech, Language, and Hearing Sciences, University of Connecticut, 2 Alethia Drive, Unit 1085, Storrs, CT 06269-1085, USA
[b] Connecticut Institute for the Brain and Cognitive Sciences, University of Connecticut, 337 Mansfield Road, Unit 1272, Storrs, CT 06269-1272, USA

ARTICLE INFO

Keywords:
Speech perception
Individual differences
Lexical processing

ABSTRACT

Research suggests that individuals differ in the degree to which they rely on lexical information to support speech perception. However, the locus of these differences is not yet known; nor is it known whether these individual differences reflect a context-dependent "state" or a stable listener "trait." Here we test the hypothesis that individual differences in lexical reliance are a stable trait that is linked to individuals' relative weighting of lexical and acoustic-phonetic information for speech perception. At each of two sessions, listeners ($n = 73$) completed a Ganong task, a phonemic restoration task, and a locally time-reversed speech task – three tasks that have been used to demonstrate a lexical influence on speech perception. Robust lexical effects on speech perception were observed for each task in the aggregate. Individual differences in lexical reliance were stable across sessions; however, relationships among the three tasks in each session were weak. For the Ganong and locally time-reversed speech tasks, increased reliance on lexical information was associated with weaker reliance on acoustic-phonetic information. Collectively, these results (1) provide some evidence to suggest that individual differences in lexical reliance for a given task are a stable reflection of the relative weighting of acoustic-phonetic and lexical cues for speech perception in that task, and (2) highlight the need for a better understanding of the psychometric characteristics of tasks used in the psycholinguistic domain to build theories that can accommodate individual differences in mapping speech to meaning.

1. Introduction

A long-standing challenge in the domain of speech perception is to explain how listeners reliably map the speech signal to meaning given the lack of invariance between speech acoustics and the linguistic representations that support meaning. For example, individual talkers differ in how they instantiate speech sounds (i.e., consonants and vowels), reflecting factors such as gender, age, dialect, and idiolect (Allen, Miller, & DeSteno, 2003; Chodroff & Wilson, 2017; Hillenbrand, Getty, Clark, & Wheeler, 1995; Lisker & Abramson, 1964; Newman, Clouse, & Burnham, 2001; Peterson & Barney, 1952; Theodore, Miller, & DeSteno, 2009). Even for a given talker, acoustic variation across productions of a given speech sound is typical (e.g., Newman et al., 2001; Theodore et al., 2009) and can be influenced by factors including the surrounding phonemes within a word (e.g., Delattre, Liberman, & Cooper, 1955) and speaking rate (e.g., Miller & Baer, 1983; Summerfield, 1981). Consequently, there is no one-to-one relationship between speech acoustics and speech sound categories.

In mapping speech to meaning, listeners use contextual cues to mitigate variability in the speech signal. Lexical context is one source of information that helps listeners disambiguate potential ambiguity in the input. For example, if a talker produces a speech sound with a voice-onset-time (VOT) midway between that expected for a canonical English /g/ and a canonical English /k/ in the context *-oat*, it may be unclear to a native English speaker whether the talker intended to say *goat* or *coat*. However, if this VOT were instead produced in the context *-ift,* it may be clear that the talker intended to say *gift* because *kift* is not an English word (Ganong, 1980). Using lexical information to guide the interpretation of speech acoustics, also known as the Ganong effect, is thus one mechanism that helps listeners solve the lack of invariance problem for speech perception (e.g., Drouin, Theodore, & Myers, 2016; Ganong, 1980; Norris, McQueen, & Cutler, 2003; Samuel & Kraljic, 2009; Tzeng, Nygaard, & Theodore, 2021).

A growing body of literature has demonstrated that individuals differ in the extent to which they use lexical information to facilitate speech perception, suggesting that some individuals are "more lexical" than

* Corresponding author at: Department of Speech, Language, and Hearing Sciences, University of Connecticut, 2 Alethia Drive, Unit 1085, Storrs, CT 06269-1085, USA.
E-mail address: rachel.theodore@uconn.edu (R.M. Theodore).

https://doi.org/10.1016/j.cognition.2022.105320
Received 6 June 2022; Received in revised form 8 September 2022; Accepted 30 October 2022
Available online 25 November 2022

others (Giovannone & Theodore, 2021a, 2021b; Ishida, Samuel, & Arai, 2016; Schwartz, Scheffler, & Lopez, 2013). For example, Ishida et al. (2016) tested participants on two tasks designed to elicit lexical effects on speech perception including (1) a phonemic restoration task and (2) a locally time-reversed speech (LTRS) task. In the phonemic restoration task, participants heard word (e.g., *accelerate*) and nonword (e.g., *vabbellerate*) items that contained one modified phoneme. For some items, the phoneme was entirely replaced by signal correlated white noise (replaced type); for other items, the signal correlated white noise was overlaid in time with the phoneme (added type). On each trial, participants heard an item with either replaced or added noise followed by the same item without noise and then rated the similarity of the two items on a scale from one to eight. In this paradigm, comparable similarity ratings for replaced and added items is taken as evidence that listeners perceptually restored the missing phoneme (Mattys, Barden, & Samuel, 2014; Samuel, 1981). The results of Ishida et al. (2016) in the aggregate showed a robust influence of lexical status on phonemic restoration; specifically, the difference in similarity ratings between replaced and added item types was smaller for words compared to nonwords, suggesting that lexical knowledge facilitated perceptual restoration of the missing phoneme (Ishida et al., 2016; Samuel, 1981).

The second task in Ishida et al. (2016) was a locally time-reversed speech task. The stimuli for this task consisted of acoustically manipulated words (e.g., *academic*) and matched nonwords (e.g., *acabemic*). These items were manipulated such that each item was parsed into segments of a specified length (e.g., 40 ms), the acoustic signal within each segment was temporally reversed, and then all segments were concatenated. Listeners were presented with pairs of stimuli in which the first item was a locally time-reversed token (i.e., the target) and the second item (spoken by a different talker) was an intact item (i.e., the standard). On each trial, listeners were asked to indicate whether the two talkers produced the same phonological string (i.e., the same word or nonword). The dependent measure was sensitivity (d'), which was calculated separately for word and nonword targets, with higher sensitivity interpreted as stronger phonological encoding of the locally time-reversed token. The results in the aggregate showed that sensitivity was higher for word compared to nonword targets, demonstrating that lexical information facilitated phonological encoding of the degraded speech.

Though robust lexical effects on perception were observed in the aggregate for both the phonemic restoration and LTRS tasks, Ishida et al. (2016) observed wide variability in the magnitude of the lexical effect for individual subjects in each task. Critically, Ishida et al. found a moderate, positive association ($r = 0.43$) between individual differences in lexical reliance across the two tasks. That is, the relative degree to which individuals relied on lexical information for phonemic restoration tracked with the relative degree to which they used lexical information when perceiving degraded speech int the LTRS task, suggesting that individual differences in lexical reliance may be a stable listener trait. Here we use the term "stable trait" to refer to consistent behavior in an individual over time and tasks. As described below and considered again in the Discussion section, a stable trait may reflect some aspect of the processing architecture that is fixed in an individual or may reflect a consistent approach that is habitually adopted in a given situation.

Ishida et al. (2016) provided critical insight into one aspect of speech processing that may reliably differ across individuals; however, they did not identify *why* some individuals may be "more lexical" than others. One possible contribution to individual differences in lexical reliance may be broader language and reading phenotype. For example, Schwartz et al. (2013) found that children with specific language impairment (SLI) showed a larger lexical effect in a Ganong task than their typically developing peers. Individuals with developmental dyslexia also show a larger Ganong effect compared to typically-developing peers (Derawi, Reinisch, & Gabay, 2022; Reed, 1989). Even among the range of typical receptive language ability (as indexed by standardized tests of language processing), weaker receptive language skills are associated with stronger use of lexical knowledge for speech perception (Giovannone & Theodore, 2021a, 2021b). A mediating factor that might drive the relationship between broad language phenotypes and lexical reliance is stability of speech sound processing. That is, increased lexical reliance may occur in tandem with weaker use of acoustic-phonetic cues for speech perception. For example, the children in Schwartz et al. (2013) who showed a large Ganong effect also showed deficits in speech sound discrimination, which is common in disorders associated with weaker receptive language ability such as developmental language disorder and SLI (e.g., Joanisse & Seidenberg, 1998, 2003). Larger Ganong effects in individuals with dyslexia (Derawi et al., 2022; Reed, 1989) can potentially be explained by the same factor, given known deficits in acoustic-phonetic processing in this population (e.g., Snowling, 1995, 1998). Moreover, individual differences in receptive language ability even within the typical range of ability are positively associated with graded sensitivity to acoustic-phonetic variation (Theodore, Monto, & Graham, 2019). Collectively, these studies raise the possibility that the reason why some individuals are "more lexical" than others may reflect the relative weakness of acoustic-phonetic processing such that more lexical individuals use lexical information to mitigate weaker acoustic-phonetic processing. Indeed, an inverse relationship between acoustic-phonetic processing and lexical reliance can be modeled in both interactive activation (e.g., McClelland & Elman, 1986) and modular models of spoken word recognition (e.g., Norris, McQueen, & Cutler, 2000).

The results of Ishida et al. (2016) suggest that individual differences in lexical reliance are stable across tasks. However, it is not yet known whether these differences are also stable across time. That is, do individual differences in lexical reliance as observed in phoneme restoration and LTRS tasks reflect a stable individual trait, or do they instead reflect a temporary state? Evidence of the association between individual differences in lexical reliance across these tasks, combined with the evidence from clinical populations (reviewed above), are wholly consistent with the hypothesis that individual differences in lexical reliance are in fact an internally consistent trait. However, a critical issue for individual differences research in the cognitive sciences domain is a lack of information about the psychometric properties of our tasks, including construct validity and test-retest reliability (Heffner et al., 2022; Heffner & Myers, 2021; Parsons, Kruijt, & Fox, 2019; Strand, Brown, Merchant, Brown, & Smith, 2018; Wilbiks, Brown, & Strand, 2022). Construct validity, or the degree to which a task measures what it is proposed to measure, is critical to the interpretation of individual differences. Without knowledge of the construct validity of a given task, the observed results are best interpreted within the scope of the measure used to obtain them, because the use of even a slightly different task could theoretically yield much different results. Heffner et al. (2022) investigated the construct validity and reliability of five tasks assumed to assess perceptual flexibility in speech perception. In their study, construct validity was quantified as the association between performance on the same task for two unique stimulus sets and reliability was quantified as the split-half reliability for performance in a single task. Though some tasks showed moderate association in performance across stimulus sets, others did not, indicative of low construct validity because variability in individual differences in these tasks may be partially attributable to differences in the task itself (i.e., stimuli). In contrast, split-half reliability was relatively high across their set of perceptual flexibility tasks.

Another way to assess whether a task is indeed measuring the assumed construct is to compare participants' performance on that task to a completely different task that is proposed to measure the same construct; that is, convergent validity can be measured as a subtype of construct validity. Tasks that are proposed to measure similar constructs – such as the Ganong, phonemic restoration, and LTRS tasks, which all are presumed to measure a lexical influence on speech perception (e.g., Ganong, 1980; Ishida et al., 2016; Samuel, 2011) – might assess lexical processing to different extents, and so might show differential construct

validity. However, if they demonstrate high convergent validity (i.e., listeners who show a large lexical effect for one task also show a large lexical effect for the other tasks), then researchers can have more confidence that these three tasks measure the same construct. Evidence of an association between lexical reliance on the phonemic restoration and LTRS tasks provided by Ishida et al. (2016) suggests some degree of convergent validity for these tasks. However, previous research also suggests that convergent validity across common tasks in the cognitive sciences is strikingly low. For example, Strand et al. (2018) tested seven tasks presumed to assess listening effort, and found that on average, performance across tasks was only weakly correlated ($r = 0.22$). In addition, Wilbiks et al. (2022) found *no* significant associations among four measures presumed to assess audiovisual integration. Findings such as these demonstrate the need to formally evaluate the assumption that our tasks assess similar constructs.

Equally important for identifying whether individual differences in lexical reliance reflect stable traits or temporary states is the test-retest reliability of a given task (Parsons et al., 2019). Test-retest reliability refers to the degree to which consistent results are obtained on a given task each time it is administered. This psychometric property of a task is fundamental for understanding the degree to which individual variation may be attributable to the task itself. Like construct and convergent validity, reliability in speech perception tasks is chronically understudied, and highly variable. For example, Basu Mallick, Magnotti, and Beauchamp (2015) found exquisite test-retest reliability ($r = 0.91$) for a task that measured the McGurk effect in the same sample at two time points separated by one year. Acceptable test-retest reliability has also been observed for adults' spoken word recognition in an eye-tracking task using the visual world paradigm (Farris-Trimble & McMurray, 2013) and children's cerebral lateralization for receptive spoken language using a dichotic listening task (Harper & Kraft, 1986). Moreover, individual differences in the relative weighting of cues for stop category perception are stable over time (Idemaru, Holt, & Seltman, 2012). In contrast, Cristia and colleagues (Cristia, Seidl, Singh, & Houston, 2016) assessed the test-retest reliability of common tasks used in infant speech perception for 13 samples in which participants in each sample completed the same task at two time points. The results were dire; only three samples showed a significant, positive association over time. Without adequate test-retest reliability of a given task, drawing meaningful conclusions regarding individual differences in performance is extremely challenging. That is, if a task has poor test-retest reliability, we might misinterpret inconsistent patterns within individuals as evidence that performance reflects a temporary state and not an individual trait, when inconsistent patterns within individuals over time may instead reflect unstable tasks.

In this context, the goal of the current work is to test the hypothesis that individual differences in lexical reliance are a stable individual trait that reflects the relative use of lexical and acoustic-phonetic cues for speech perception. Participants completed a Ganong, phonemic restoration, and LTRS task at two time points. Robust lexical influences for these three tasks have repeatedly been observed when considering performance of a sample in the aggregate (e.g., Ganong, 1980; Giovannone & Theodore, 2021a, 2021b; Ishida et al., 2016; Mattys et al., 2014; Samuel, 1981) and, at a single point in time, an association between individual differences in the phonemic restoration and LTRS tasks has been observed (Ishida et al., 2016). In the current work, we quantified lexical and acoustic-phonetic reliance scores in each task for each listener at each time point. If individual differences in lexical reliance reflect stable individual traits, then we will observe a positive association between individuals' lexical reliance scores over time for each task in addition to positive associations between individuals' lexical reliance scores across tasks. If individual differences in lexical reliance are linked to relative use of acoustic-phonetic cues, then individuals who show strong lexical reliance scores will also show weaker acoustic-phonetic scores for a given task. A failure to observe these predicted patterns would suggest that individual differences in lexical reliance may reflect temporary listener states and/or poor psychometric characteristics of the selected tasks.

2. Methods

2.1. Participants

Participants ($n = 142$) were recruited from the Prolific participant pool (https://www.prolific.co; Palan & Schitter, 2018). All participants were monolingual English speakers born in and currently residing in the United States with no previous history of language-related disorders. Twenty-two participants were excluded due to failure to comply with task instructions as described in the procedure section below. The final sample included 120 participants at session one and 73 participants who also completed session two. Invitations to complete session two were issued two weeks after the completion of session one. The mean time between sessions was 17 days ($SD = 4$ days; *range* = 14–35 days). The session 1 sample size ($n = 120$) was determined in reference to the sample size used in experiment 1 of Ishida et al. (2016, $n = 60$), which was the higher of two samples sizes tested in that study ($n = 52$ in their experiment 2). Specifically, we tested twice their sample size in our session 1 with the goal of ensuring that the number of people who returned for session 2 would meet or exceed the Ishida et al. sample size. All data were collected inclusively between October 26, 2021 and May 17, 2022.

Given that testing the primary hypotheses for the current work requires examining performance across the two sessions, the analyses presented here consider the participants who completed both sessions ($n = 73$).[1] All participants were between 18 and 35 years of age (*mean* = 26 years, $SD = 5$ years). The sample included 50 women, 21 men, and 2 individuals who preferred not to report gender. In terms of race, participants were White (59), Black or African American (8), Asian (4), American Indian/Alaska Native (1), or more than one race (1). Ethnicity of the sample included 1 Hispanic or Latino participant and 72 participants who were not Hispanic or Latino.

2.2. Stimuli

2.2.1. Ganong task

Stimuli for the Ganong task consisted of two voice-onset-time (VOT) continua, one that perceptually ranged from *gift* to *kift* and one that perceptually ranged from *giss* to *kiss*. Stimuli were created from natural recordings of a male talker producing the items *gith, kith, gift, kift, giss*, and *kiss*.[2] The CV portion of the *gith* and *kith* productions were extracted and used to create a 15-step VOT continuum in Praat (Boersma, 2002) using the script developed by Winn (2020). Specifically, VOTs ranged between 10 and 100 ms in equal steps, and fundamental frequency at voicing onset was held constant across continuum steps (105 Hz, reflecting the average fundamental frequency at voicing onset for the natural *gith* and *kith* productions). The continuum perceptually ranged from /gɪ/ – /kɪ/. To create the *gift* – *kift* continuum, the /ft/ portion of

[1] As described in the main text, 120 participants completed session one and 73 participants returned to also complete session two. The analyses presented in the main text examine performance for the participants who completed both experimental sessions ($n = 73$). In the supplementary material, we report analyses for the full sample ($n = 120$) that were conducted to parallel the analyses for session one presented in the main text. Qualitatively identical patterns were observed in all cases except one; namely, there was a statistically significant relationship between lexical reliance scores on the Ganong and LTRS tasks in session one for the full sample ($n = 120$) but not the subset ($n = 73$), as described in detail in the supplemental material.

[2] Distinct talkers were used for stimulus creation across the three tasks. That is, the stimuli for the Ganong task were produced by a different talker than the stimuli for the phonemic restoration task, and both of these talkers were different than the two talkers who produced the stimuli for the LTRS task.

the natural *gift* token was spliced to each of the 15 steps of the /gɪ/ – /kɪ/ continuum. To create the *giss* – *kiss* continuum, the /s/ portion of the natural *giss* token was spliced to each of the 15 steps of the VOT continuum; the /s/ portion was equal in duration (224 ms) to the /ft/ portion of the *gift* – *kift* continuum. Using this method ensured that the CV portion (i.e., VOT and vowel) for each step were acoustically identical across the two continua; that is, the only physical difference between continua for a given step was the coda portion of the syllable (i.e., /ft/, /s/).

Pilot testing of the /gɪ/ – /kɪ/ continuum appended to a lexically neutral context (i.e., the /θ/ portion of the natural *gith* token) revealed that an eight-step subset including tokens with VOTs of 29, 36, 42, 49, 55, 61, 68, and 74 ms yielded a continuum that perceptually ranged from *gith* to *kith*, with the average voicing boundary centered in the continuum. Accordingly, these eight steps of the *gift* – *kift* and *giss* – *kiss* continua were used as stimuli for the current study.

2.2.2. Phonemic restoration task

Stimuli for the phonemic restoration task were a subset of those used in experiment 2 of Ishida et al. (2016), originally constructed by Mattys et al. (2014). Stimuli for this task consisted of 20 word-nonword pairs. In each pair, the word item contained either a liquid or a nasal phoneme in the onset position of the final syllable (e.g., *accelerate*); this phoneme will be referred to as the critical phoneme. The nonword item of each pair was matched to the real word member with respect to stress pattern and the final two syllables (e.g., *vabbellerate*), thus preserving the same critical phoneme across the word and nonword members of each pair. In addition, 10 filler word-nonword pairs were interleaved with the critical word-nonword pairs. Filler pairs contained a liquid or nasal phoneme in their first syllable (e.g., *skulduggery*), which served as the critical phoneme to encourage participants to listen to the whole word on each trial rather than just the final syllable. The nonword member of each filler pair was matched to the word member with respect to stress pattern, initial syllable, and initial phoneme of the second syllable (e.g., *skulldassipye*). Thus, a total of 60 items were used as stimuli for the phonemic restoration task (20 pairs × 2 items/pair +10 filler pairs x 2 items/pair).

Two types of each of the 60 items were created that contained signal correlated white noise at 0 dB SNR in the same temporal position as the critical phoneme. For the added item type, signal correlated white noise was added to the stimulus to occur simultaneously with the critical phoneme. For the replaced item type, the signal correlated white noise replaced the critical phoneme. A subset of the stimuli used in Ishida et al. (2016) was selected by sampling one third of their items in each category (liquid-target words/matched nonwords, nasal-target words/matched nonwords, filler words/matched non-words), thus preserving the relative distribution of items in each category in our subset. A full list of the stimuli used in the current phonemic restoration task can be found in Appendix A.

2.2.3. Locally time-reversed speech (LTRS) task

Stimuli for the LTRS task were a subset of those used in experiment 2 of Ishida et al. (2016). The LTRS stimuli created by Ishida and colleagues consisted of stop-dominant words and matched pseudowords that each ranged between three to five syllables in length. Matched pseudowords were created by changing the place of articulation in one phoneme for each word item; for example, for the word *academic*, the matched nonword was *acabemic*. Each item was recorded by both a male and female talker. The locally time-reversed stimuli were created by segmenting each token produced by the male talker into windows of equal sizes (20, 40, 60, and 80 ms). Each segment was reversed in time, then the segments of each token were concatenated to create a locally time-reversed stimulus. The tokens produced by the female talker remained unaltered.

The LTRS stimulus set used in the current study consisted of 24 stop-dominant words and 24 matching pseudowords. This subset was created by sampling one third of the items provided by Ishida and colleagues. The stimulus set used in the current study was further reduced from that used in experiment 2 of Ishida et al. by only using the 40 and 60 ms reversal windows. These windows were chosen to reduce task length while preserving a degree of variability present in the Ishida et al. task. A full list of the stimuli used in the current locally time-reversed speech task can be found in Appendix B.

2.3. Procedure

The procedure for both sessions was identical. The experiment was deployed as a web-based study hosted on the Gorilla Experiment Builder platform (https://gorilla.sc; Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020). After providing informed consent, participants completed two tasks designed to screen for headphone compliance on web-based platforms (Milne et al., 2021; Woods, Siegel, Traer, & McDermott, 2017). The main experimental tasks were then completed in the following order: Ganong, one block of phonemic restoration (word or nonword), LTRS, and a second block of phonemic restoration (word or nonword). All listeners completed both the word and nonword phonemic restoration blocks, with block order counterbalanced across listeners. This order was used to promote a more direct replication of the procedure used in Ishida et al. (2016), where participants completed one block of phonemic restoration (words or nonword) followed by the LTRS task and then the second block of phonemic restoration (word or nonword), with order of the word and nonword phonemic restoration blocks counterbalanced across listeners. After the final experimental task, participants completed a brief demographic survey. Each session lasted approximately 35 min and participants were paid $5.83 for their participation at each session. Procedural details specific to each task are explicated below.

2.3.1. Ganong task

The Ganong task consisted of 96 trials of phonetic identification (2 continua × 8 VOT steps x 6 repetitions). All stimuli were presented in a single block (i.e., stimuli from the two continua were mixed) in a different randomized order for each participant. Each trial began with the presentation of an auditory stimulus, after which participants were directed to indicate the initial sound by clicking on one of two buttons labeled “g” or “k.” The ISI was 1000 ms, timed from the participant’s response to the onset of the next auditory stimulus. Participants who showed inverse response functions (indicative of inverting the button assignments) or a flat response function (indicative of failure to perform the task as directed) were excluded, as reported in the Participants section.

2.3.2. Phonemic restoration task

As in Ishida et al. (2016), the phonemic restoration task was divided into two blocks, one for word items and one for nonword items. Each block contained 60 trials (30 replaced items +30 added items, including fillers). The task structure was the same for both blocks. On each trial, participants heard two items separated by 400 ms. The first item was a manipulated stimulus (i.e., a replaced or added item type) and the second item was the original, unmanipulated version of the stimulus. Then, participants were asked to judge how similar the two items were on a scale from 1 (not similar) to 8 (very similar), following the rating scale procedure developed for the phonemic restoration task (Samuel, 1996). Participants responded by clicking on one of eight appropriately labeled buttons. As in Ishida et al. (2016), participants were instructed to ignore the white noise to the best of their ability and to use the full range of the scale in their responses. The ISI was 1000 ms, timed from the participant’s response to the onset of the next auditory pair. Participants who did not provide at least three unique ratings across the word and nonword blocks (indicative of failure to perform the task as directed) were excluded, as reported in the Participants section.

2.3.3. Locally time-reversed speech task

The LTRS task consisted of 96 trials and followed the procedure outlined in Ishida et al. (2016). On each trial, listeners first heard two items separated by 400 ms. The first item (the target) was a locally time-reversed stimulus produced by the male talker and the second item (the standard) was an intact stimulus provided by the female talker. For a given trial, the target (reversed) could either be a word or a nonword and the standard (intact) could either be the same item or the different matched item, yielding four trial types: word-word, word-nonword, nonword-word, and nonword-nonword. For example, when *academic* was presented as the target and *acabemic* was presented as the standard, this yielded a word-nonword trial type. Likewise, when *acsheptable* was presented as the target and *acceptable* was presented as the standard, this yielded a nonword-word trial type. Each of the 24 stimulus pairs (Appendix B) was presented in each of the four trial structures, yielding 96 trials for the LTRS task. For half of the trials, the target stimulus used a 40 ms reversal window; for the other half of the trials, the target stimulus used a 60 ms reversal window. For a given participant, reversal window was held constant across the four trial structures for a given item pair. Across listeners, we counterbalanced the assignment of item pairs to reversal window. After the stimulus pair was presented, participants were asked to indicate whether the two items contained the same phonological string (i.e., the same word or the same nonword) by clicking on one of two buttons labeled "Same" or "Different." The ISI was 1000 ms, timed from the participant's response to the onset of the next auditory pair. Participants who pressed a single button for every trial (indicative of failure to perform the task as directed) were excluded, as reported in the Participants section.

3. Results

We confirm that this manuscript reports all measures, conditions, data exclusions, and – as described in the Participants section – how the sample size was determined. Trial-level data and a script (in R) to reproduce all analyses and figures presented here are available at the Open Science Framework (OSF) repository for this manuscript: https://osf.io/dhybk/. Five sets of analyses were conducted. First, we confirmed the lexical effect of interest in each task in the aggregate. Second, we examined the stability of individual differences in lexical reliance over time for each task. Third, we examined the association of lexical reliance between tasks at each session. Fourth, we examined the relationship between individuals use of lexical and acoustic-phonetic cues in each task at each session. Finally, we present a reanalysis of Ishida et al. to promote more direct comparison between the results observed in their sample and the current sample. Each analysis is presented in turn below.

3.1. Confirmation of effects of interest

Three analyses were conducted to confirm that the primary lexical effects of interest were observed in the aggregate. For the Ganong task, the primary effect of interest is evidence that proportion /k/ responses are higher for the *giss–kiss* continuum compared to the *gift–kift* continuum, consistent with the influence of lexical context on perception of acoustic-phonetic ambiguity. For the phonemic restoration task, the primary effect of interest is evidence of increased phonemic restoration in the word block compared to the nonword block, consistent with lexical context biasing the perceptual recovery of portions of the speech signal that are replaced with noise. For the LTRS task, the primary effect of interest is evidence of higher sensitivity in discriminating the phonological content in reversed and intact speech for word compared to nonword targets, consistent with the interpretation that lexical information is used to guide perception of degraded speech. Each analysis is presented in turn, below.

3.1.1. Ganong task

To visualize performance, mean proportion /k/ responses was calculated separately for each participant for all cells formed by crossing VOT, continuum, and session. Grand means were then calculated over by-subject means, which are shown in Fig. 1, panel A. Visual inspection suggests a robust lexical effect in each session such that there are more /k/ responses for the *giss–kiss* compared to the *gift-kift* continuum. Visual inspection also suggests that the lexical effect is weakened across sessions. To analyze these patterns statistically, trial-level responses were analyzed using a generalized linear mixed effects model (GLMM) with the binomial response family (i.e., a logistic regression) as implemented using the glmer() function of the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in R. The dependent variable was category response (0 = /g/, 1 = /k/). The model contained fixed effects of continuum, VOT, session, and their interactions. VOT was entered into the model as a continuous variable scaled and centered around the mean. Continuum (*gift–kift* = −0.5, *giss–kiss* = 0.5) and session (session one = −0.5, session two = 0.5) were each entered as sum-coded contrasts. The random effects structure consisted of random intercepts by participant and random slopes by participant for VOT, continuum, session, and all interactions.

As expected, the model results showed a significant effect of VOT ($\widehat{\beta}$ = 3.159, SE = 0.135, z = 23.403, $p < 0.001$), indicating that /k/ responses increased as VOT increased. Critically, the model also showed a significant effect of continuum ($\widehat{\beta}$ = 2.564, SE = 0.269, z = 9.538, $p < 0.001$), with the direction of the beta estimate indicating more /k/ responses for the *giss-kiss* continuum compared to the *gift-kift* continuum. The interaction between VOT and continuum was also reliable ($\widehat{\beta}$ =0.325, SE = 0.146, z = 2.219, p = 0.027), indicating that the Ganong effect differed across continuum steps. The model did show a main effect of session ($\widehat{\beta}$ = −0.378, SE = 0.132, z = −2.856, p = 0.004), with /k/ responses decreasing from session one to session two, and an interaction between continuum and session ($\widehat{\beta}$ = −1.178, SE = 0.233, z = −5.059, $p < 0.001$), indicating that the Ganong effect (i.e., the difference between the two continua) was attenuated in session two compared to session one. The interaction between VOT, continuum, and session was not reliable ($\widehat{\beta}$ = −0.081, SE = 0.261, z = −0.310, p = 0.757).

To confirm that the interaction between continuum and session reflected an attenuation of the Ganong effect across sessions and not, for example, the extinction of the Ganong effect in session two, separate models were constructed for each session following the structure outlined for the omnibus model except for removing session from the fixed and random effects structure. A robust effect of continuum was observed in both session one ($\widehat{\beta}$ = 3.079, SE = 0.269, z = 11.446, $p < 0.001$) and session two ($\widehat{\beta}$ = 1.925, SE = 0.290, z = 6.637, $p < 0.001$).

3.1.2. Phonemic restoration task

To visualize performance, mean similarity rating was calculated separately for each participant for all four cells formed by crossing block and item type. Grand means were then calculated over by-subject means, which are shown in Fig. 1, panel B. Recall that complete phonemic restoration would manifest as identical similarity ratings between replaced and added items, indicating that listeners judged the replaced item to be as similar in phonological form to an unmodified item as they did the added item. Visual inspection suggests the expected lexical effect such that phonemic restoration is stronger for the word block compared to the nonword block. That is, the difference in similarity ratings between replaced and added item types is *smaller* for words compared to

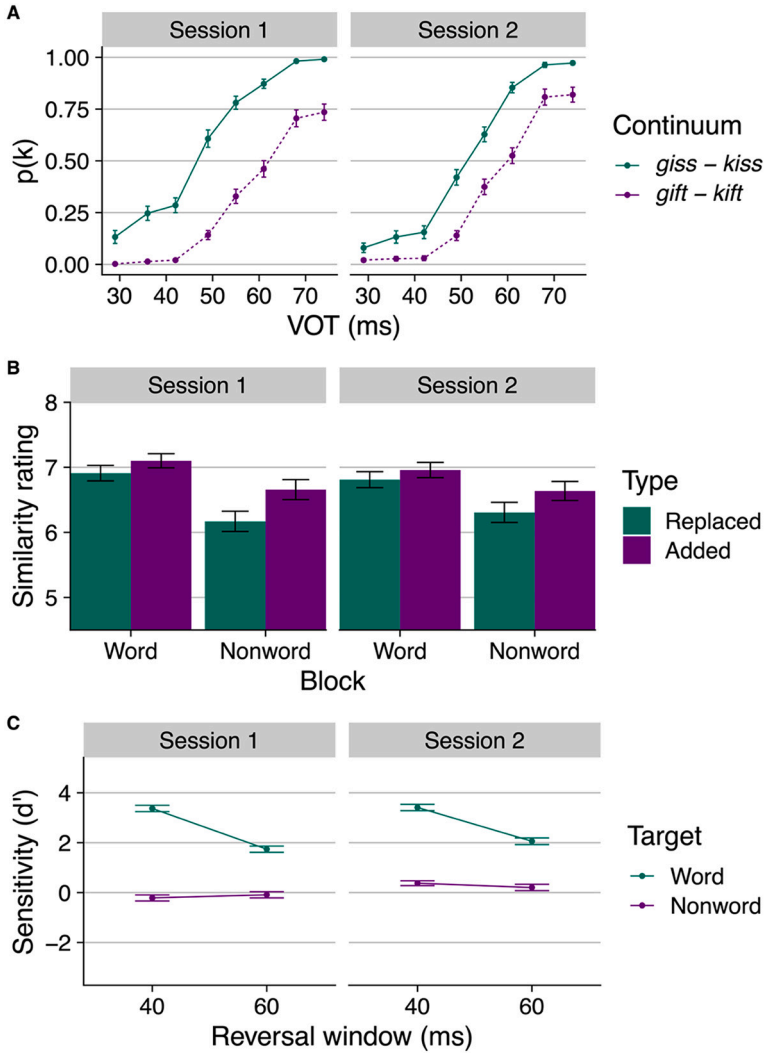

Fig. 1. Aggregate performance for the Ganong (A), phonemic restoration (B), and LTRS (C) tasks at each session. Panel A shows mean proportion /k/ responses in the Ganong task as a function of VOT and continuum. Panel B shows mean similarity ratings as a function of block and item type. Panel C shows mean sensitivity as a function of reversal window and target. In all panels, means reflect grand means calculated over by-subject averages; error bars indicate standard error of the mean.

nonwords, suggesting more complete phonemic restoration for words compared to nonwords.

To examine these patterns statistically, trial-level responses were analyzed using a linear mixed effects model (LMM) with the binomial response family (i.e., a logistic regression) as implemented using the lmer() function of the lme4 package (Bates et al., 2015) in R.[3] The dependent variable was trial-level similarity rating. The model contained fixed effects of type, block, session, and their interactions; type (replaced = −0.5, added = 0.5), block (word = −0.5, nonword = 0.5), and session (session one = −0.5, session two = 0.5) were entered into the model as sum-coded contrasts. The random effects consisted of random intercepts by subject and random slopes by subject for type, block, session, and their interactions.

The results showed a main effect of type ($\widehat{\beta} = 0.289$, $SE = 0.036$, $t = 7.946$, $p < 0.001$), reflecting higher ratings for added compared to replaced items, a main effect of block ($\widehat{\beta} = -0.502$, $SE = 0.106$, $t = -4.728$, $p < 0.001$), reflecting lower similarity ratings for nonwords compared to words, and no main effect of session ($\widehat{\beta} = -0.032$, $SE = 0.075$, $t = -0.421$, $p = 0.675$). The model also showed an interaction between type and session ($\widehat{\beta} = -0.100$, $SE = 0.047$, $t = -2.118$, $p = 0.036$), indicating that the difference between replaced and added items was attenuated at session two compared to session one. The interaction between block and session was not reliable ($\widehat{\beta} = 0.179$, $SE = 0.108$, $t = 1.656$, $p = 0.102$).

Critically, the results showed a reliable interaction between type and block ($\widehat{\beta} = 0.240$, $SE = 0.054$, $t = 4.423$, $p < 0.001$), indicating that the magnitude of the difference between item types was smaller in the word block compared to the nonword block, consistent with previous evidence indicating a lexical effect on phonemic restoration (e.g., Ishida et al., 2016). That is, the type by block interaction is the critical effect of interest because it reflects decreased phonemic restoration for nonwords compared to words. Simple slopes analyses showed higher ratings for added compared to replaced items for both the word ($\widehat{\beta} = 0.169$, $SE = 0.036$, $t = 4.662$, $p < 0.001$) and nonword blocks ($\widehat{\beta} = 0.409$, $SE = 0.054$, $t = 7.619$, $p < 0.001$); thus, both blocks showed incomplete restoration, though phonemic restoration was more complete for words compared to nonwords. The interaction between type, block, and session was not

[3] To promote comparison to the analyses conducted by Ishida et al. (2016), both the phonemic restoration and locally time-reversed speech tasks were also analyzed using ANOVA, which showed quantitively similar results to those observed in the mixed effects regression models presented in the main text. These analyses can be viewed by executing the script provided on the OSF repository for this manuscript.

reliable ($\widehat{\beta} = -0.116$, $SE = 0.101$, $t = -1.151$, $p = 0.253$). That is, there was no statistically significant interaction to suggest that the magnitude of the lexical effect (i.e., the type by block interaction) was weaker in session two compared to session one, in contrast to results from the Ganong task.

3.1.3. Locally time-reversed speech task

Performance for the LTRS task was analyzed following the procedure outlined in Ishida et al. (2016). Specifically, sensitivity (*d'*) was calculated for each participant at each session for the four cells formed by crossing target and reversal window. For word targets (e.g., *academic*), hits were defined as "same" responses when the standard was the same word (e.g., *academic*) and false alarms were defined as "same" responses when the standard was the matched nonword (e.g., *acabemic*). For nonword targets (e.g., *acabemic*), hits were defined as "same" responses when the standard was the same nonword (e.g., *acabemic*) and false alarms were defined as "same" responses when the standard was the matched word (e.g., *academic*). If a participant had hit or false alarm rates at either floor (0) or ceiling (1), then these values were corrected to 0.01 and 0.99, respectively, so that *d'* could be calculated. Grand means were then calculated over by-subject means, which are shown in Fig. 1, panel C. Visual inspection suggests a robust lexical influence on sensitivity such that *d'* is higher for word compared to nonword targets, an effect that appears to be slightly attenuated in session two compared to session one. Visual inspection also suggests an interaction between target and reversal window, with a stronger lexical effect for the 40 ms compared to the 60 ms reversal window.

To examine these patterns statistically, sensitivity scores (*d'*) were submitted to repeated-measures ANOVA with the factors of target (word vs. nonword), reversal window (40 ms vs. 60 ms), and session (session one vs. session two). Critically, the results showed a main effect of target [$F(1,72) = 258.24$, $p < 0.001$], reflecting higher sensitivity for word compared to nonword targets and thus confirming the expected lexical influence on sensitivity. The ANOVA also showed a main effect of reversal window [$F(1,72) = 103.09$, $p < 0.001$], reflecting higher sensitivity for the 40 ms window compared to the 60 ms window, and a main effect of session [$F(1,72) = 40.24$, $p < 0.001$], reflecting slightly higher sensitivity in session two compared to session one. The ANOVA also revealed significant interactions between target and reversal window [$F(1,72) = 83.52$, $p < 0.001$], target and session [$F(1,72) = 4.23$, $p = 0.004$], and target, reversal window, and session [$F(1,72) = 6.38$, $p = 0.014$]. The reversal window by session interaction was not reliable [$F(1,72) < 1.00$, $p = 0.932$].

To explicate the three-way interaction, separate repeated-measures ANOVAs were conducted for each session with the factors of target and reversal window. The session one ANOVA showed significant main effects of target [$F(1,72) = 211.32$, $p < 0.001$] and reversal window [$F(1,72) = 78.05$, $p < 0.001$] in addition to a significant interaction between target and reversal window [$F(1,72) = 103.82$, $p < 0.001$]. Paired *t*-tests confirmed higher sensitivity for word compared to nonword targets for both the 40 ms [$t(72) = 18.62$, $p < 0.001$] and 60 ms reversal windows [$t(72) = 8.43$, $p < 0.001$]. For word targets, sensitivity was higher for the 40 ms compared to the 60 ms window [$t(72) = 11.96$, $p < 0.001$]; for nonword targets, no reliable difference between these two reversal windows was observed [$t(72) = -1.20$, $p = 0.234$]. The same pattern of results was observed for session two. Specifically, there was a significant main effect for both target [$F(1,72) = 240.77$, $p < 0.001$] and reversal window [$F(1,72) = 66.44$, $p < 0.001$] and a significant interaction between these two factors [$F(1,72) = 29.15$, $p < 0.001$]. Paired t-tests confirmed higher sensitivity for word compared to nonword targets for both the 40 ms [$t(72) = 18.816$, $p < 0.001$] and 60 ms reversal windows [$t(72) = 8.503$, $p < 0.001$]. For word targets, sensitivity was higher for the 40 ms compared to the 60 ms window [$t(72) = 8.27$, $p < 0.001$]; for nonword targets, no reliable difference between these two reversal windows was observed [$t(72) = 1.42$, $p = 0.161$]. Collectively, these results confirm the presence of the expected lexical effect on sensitivity scores in both sessions and provide no evidence to suggest that the 3-way interaction observed in the omnibus ANOVA reflects an attenuation of the lexical effect at session two compared to session one.

3.2. Individual differences in lexical reliance over time (test-retest reliability)

All three tasks exhibited the expected influence of lexical status on performance in the aggregate. The next set of analyses was conducted to examine whether individual differences in each task were stable over time. To do so, we first calculated a lexical reliance score for each participant for each task at each session. For the Ganong task, recall that performance at each session was analyzed using a GLMM that included random intercepts by subject and random slopes by subject for VOT, continuum, and their interaction. By-subject random slopes for continuum from each model were used as the lexical reliance score at each session, respectively. Accordingly, a by-subject random slope for continuum equal to zero reflects weak lexical reliance (i.e., no difference in /k/ responses between the two continua), slopes greater than zero reflect a lexical influence in general (i.e., more /k/ responses for the *giss–kiss* compared to the *gift–kift* continuum), and the magnitude of by-subject random slopes for continuum can be interpreted as a continuous measure of lexical reliance (i.e., lower slope values indicate a weaker effect of continuum compared to higher slope values).

A similar approach was taken to calculate a lexical reliance score for the phonemic restoration task at each session. Recall that separate LMMs were constructed for each session. The dependent variable was trial-level similarity rating. Each model contained fixed effects of type, block, and their interaction; type (replaced = −0.5, added = 0.5) and block (word = −0.5, nonword = 0.5) were entered into the model as sum-coded contrasts. The random effects structure consisted of random intercepts by participant and random slopes by participant for type, block, and their interaction. The lexical reliance score for each participant in each session was quantified by the by-subject random slope coefficient for the type by block interaction. With this metric, a slope coefficient of zero indicates no lexical effect on phonemic restoration (i.e., the effect of item type is constant across word and nonword blocks) whereas a slope coefficient greater than zero indicates that the difference between added and replaced items is smaller for the word block compared to the nonword block (i.e., the effect of type increases from word to nonword blocks). As described for the Ganong task, the magnitude of by-subject random slope coefficients can be interpreted as a continuous measure of lexical reliance.

Lexical reliance scores for the LTRS task were calculated according to the metric used in Ishida et al. (2016). Sensitivity (*d'*) for each participant at each session was first calculated separately for word and nonword targets for each reversal window (as described above). Then, mean sensitivity was calculated for word and nonword targets as the average *d'* across the two reversal windows. Finally, the lexical reliance score was calculated as the difference in *d'* between word and nonword targets. With this metric, a lexical reliance score of zero indicates no lexical effect (i.e., sensitivity for word targets is the same as sensitivity for nonword targets), a positive score indicates a lexical effect (i.e., sensitivity for word targets is greater than sensitivity for nonword targets), and the lexical reliance score can be interpreted as a continuous measure of lexical reliance.

Using these scores, we calculated the association between lexical reliance across sessions for each of the three tasks using Pearson's *r*. To control family-wise error rate, we applied the conservative Bonferroni correction to guide interpretation of *p*-values, which adjusted alpha to 0.017 given family-wise alpha of 0.05 and three comparisons. As shown in Fig. 2, each task showed an association between lexical reliance scores in each session. Specifically, we observed $r = 0.72$ ($p < 0.001$) for the Ganong task, $r = 0.37$ ($p = 0.001$) for the phonemic restoration task, and $r = 0.74$ ($p < 0.001$) for the LTRS task. Because the magnitude of the

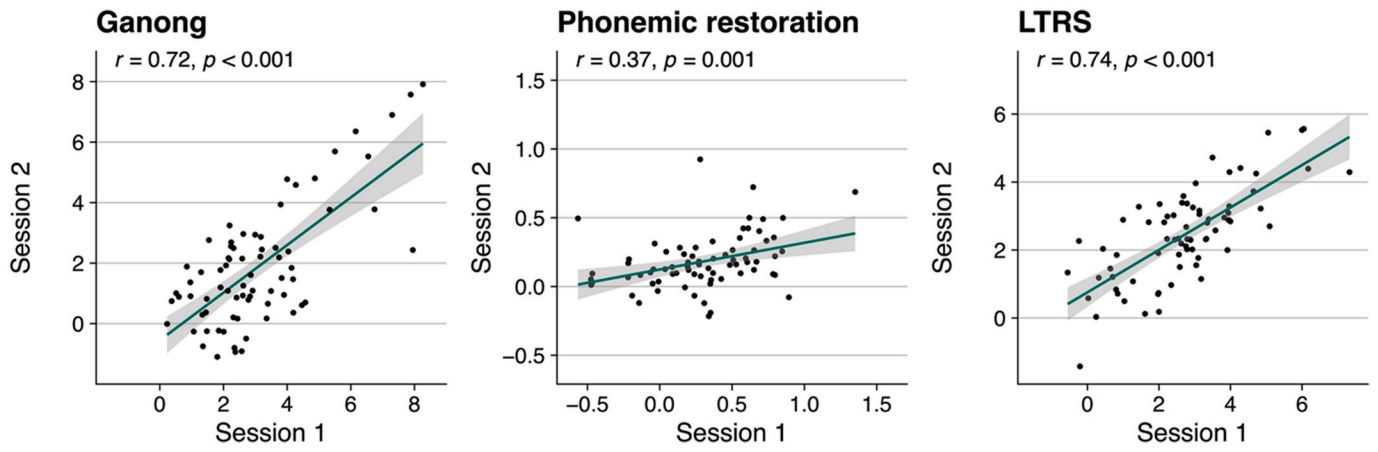

Fig. 2. Association between lexical reliance scores across the two test sessions for the Ganong (left), phonemic restoration (middle), and LTRS (right) tasks. In all plots, black circles indicate individual participants, the green function indicates the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

correlation was numerically larger for the Ganong and LTRS tasks compared to the phonemic restoration task, we performed a test of the significance for the difference between two correlations based on dependent groups as implemented with the cocor.dep.groups.non-overlap() function of the cocor package (Diedenhofen & Musch, 2015). The results showed that the magnitude of the correlations observed for the Ganong and LTRS tasks were significantly stronger than the magnitude of the correlation observed for the phonemic restoration task ($z = 3.050$, $p = 0.002$ and $z = 3.260$, $p = 0.001$, respectively).

Finally, recall that the time between each test session showed some variability across participants (*mean* = 17 days, *SD* = 4 days; *range* = 14–35 days). To examine whether individuals' consistency over time was linked to the time between the two test sessions, three additional correlations were calculated, one for each task. For each task, we examined the association between the difference in lexical reliance score across the two sessions and the number of days between sessions. There

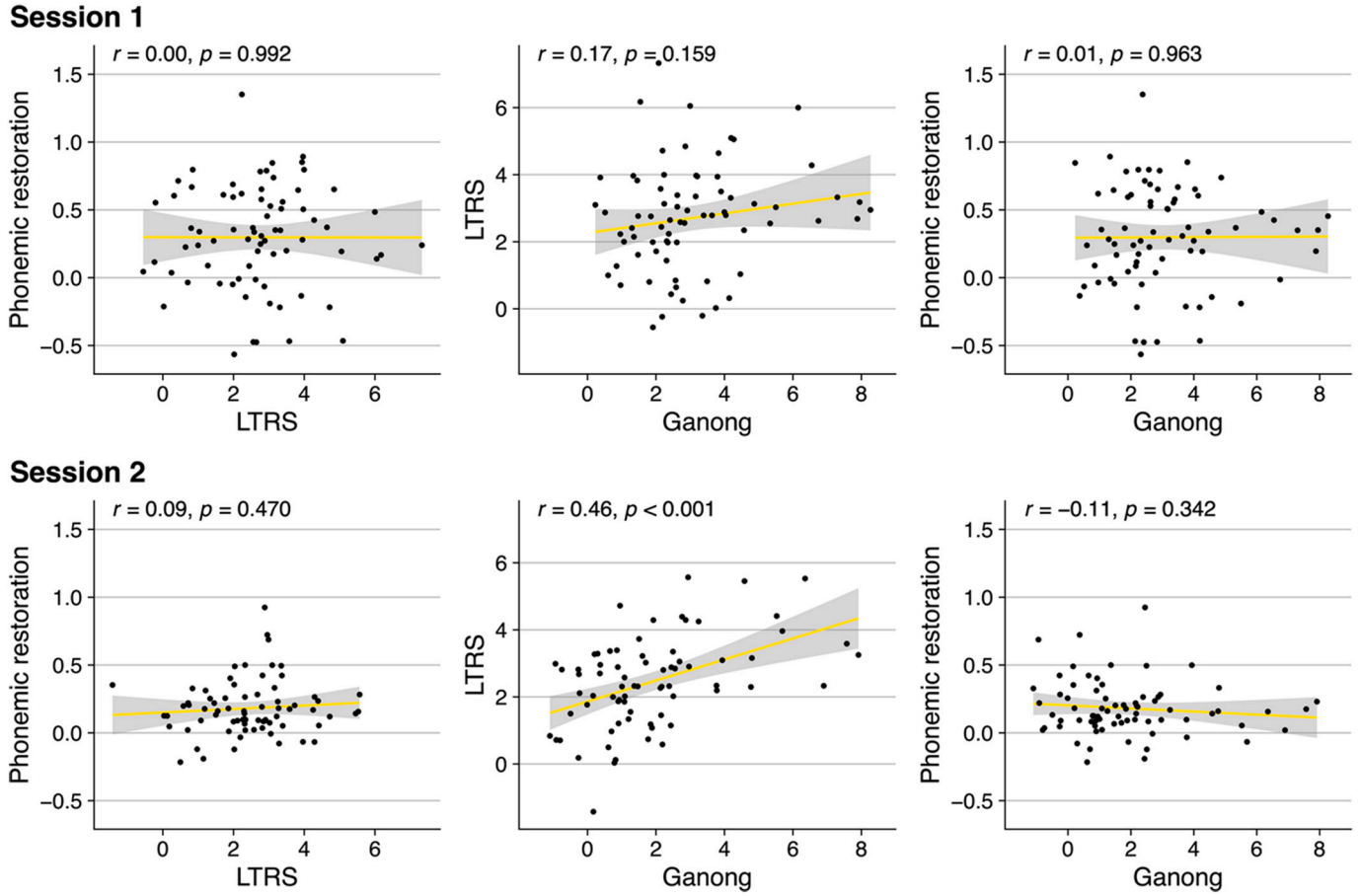

Fig. 3. Association between lexical reliance scores across tasks for session one (top) and session two (bottom). In all plots, black circles indicate individual participants, the yellow function indicates the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

was no significant association between the difference in lexical reliance scores across sessions and the number of days between sessions for the Ganong task ($r = 0.17$, $p = 0.139$), the phonemic restoration task ($r = 0.00$, $p = 0.986$), or the LTRS task ($r = 0.19$, $p = 0.107$).

Collectively, these analyses suggest that individual differences in lexical reliance *within a given task* were stable over time, more so for the Ganong and LTRS tasks compared to the phonemic restoration task.

3.3. Individual differences in lexical reliance across tasks (convergent validity)

In the next set of analyses, we examined the degree to which individual differences in lexical reliance were stable across tasks. Recall that Ishida et al. (2016) observed that individual differences in the lexical influence on a phonemic restoration task were associated with individual differences in the lexical influence on an LTRS task. That is, they found evidence that "more lexical" individuals exhibited relatively stronger lexical effects on both tasks. Here we examined whether the same pattern would be observed in the current sample at each session and, moreover, if individuals' lexical effects for the phonemic restoration and LTRS tasks were also associated with lexical effects on the Ganong task.

Individual lexical reliance scores for each task were identical to those described above. Using these scores, we calculated the association between each task at each session using Pearson's *r*. To control family-wise error rate, we applied the conservative Bonferroni correction to guide interpretation of *p*-values, which adjusted alpha to 0.008 given family-wise alpha of 0.05 and six comparisons. As shown in Fig. 3, there was very limited evidence that individuals who demonstrate a large lexical effect in one task also do so in the other tasks. The association between the LTRS and phonemic restoration tasks showed $r = 0.00$ ($p = 0.992$) at session one and $r = 0.09$ ($p = 0.470$) at session two. Likewise, the association between the Ganong and phonemic restoration tasks showed $r = 0.01$ ($p = 0.963$) at session one and $r = -0.11$ ($p = 0.342$) at session two. At session one, the relationship between the Ganong and LTRS tasks was also weak ($r = 0.17$, $p = 0.159$); at session two, this relationship remained moderate in magnitude though it was statistically reliable ($r = 0.46$, $p < 0.001$). Collectively, these results provide no substantial evidence to suggest that individual differences in lexical reliance were stable across the three tasks examined here.

3.4. Relationship between individuals' use of acoustic-phonetic and lexical cues

Our fourth analysis examined whether individual differences in lexical reliance reflect a trade-off in the use of lexical and acoustic-phonetic cues for speech perception given previous evidence suggesting that increased lexical recruitment may compensate for relatively weaker acoustic-phonetic encoding. To do so, we derived a metric of acoustic-phonetic reliance for each of the three tasks and then assessed the association between individuals' lexical and acoustic-phonetic reliance scores. Analyses and results for each task are described in turn below. In total, six associations were calculated for this analysis using Pearson's *r*. To control family-wise error rate, we applied the conservative Bonferroni correction to guide interpretation of *p*-values, which adjusted alpha to 0.008 given family-wise alpha of 0.05 and six comparisons.

3.4.1. Ganong task

For the Ganong task, the lexical reliance score was identical to that described previously (i.e., by-subject random slopes for continuum as derived from the GLMM for each session). An acoustic-phonetic reliance score was calculated as follows. First, we fit trial-level data to a GLMM with the fixed effect of VOT separately for each continuum and each session. The random effects structure in each model included random intercepts by subject and random slopes by subject for VOT. Then, we used the coefficients of the by-subject random slopes for each continuum to calculate a mean VOT slope across continua for each subject at each session. This metric thus reflects the steepness of the psychometric function linking VOT to /k/ responses. Higher values indicate more categorical responses, and thus increased consistency in mapping VOT to a phonetic category, whereas lower values indicate less consistent mapping between VOT and phonetic category. As shown in Fig. 4 (top), there was a strong, inverse association between the lexical and acoustic-phonetic reliance scores in both session one ($r = -0.84$, $p < 0.001$) and session two ($r = -0.50$, $p < 0.001$). Specifically, individuals with larger lexical effects showed weaker use of VOT for category identification. Data from three representative participants are shown in Fig. 4 to illustrate this relationship. Subject 68 shows a ceiling lexical effect and minimal use of VOT, subject 97 shows a moderate lexical effect along with moderately consistent use of VOT, and subject 119 shows a floor lexical effect and near perfect consistency in using VOT for category identification.

As the adage goes, "it didn't have to be this way." Recall that the metric of acoustic-phonetic reliance was calculated independently from the lexical effect. In principle, a participant could show a large lexical effect (i.e., functions for the two continua that are greatly displaced along the x-axis) and a large acoustic-phonetic effect (i.e., perfectly categorical response functions for each continuum). Likewise, a participant could show a small lexical effect (i.e., functions for the two continua that are aligned along the x-axis) and a small acoustic-phonetic effect (i.e., shallow response functions for each continuum). Instead, we observed a robust association between these two measures, suggesting that individual differences in lexical reliance occur in tandem with relatively weaker reliance on acoustic-phonetic cues for phonetic identification.

3.4.2. Phonemic restoration task

For the phonemic restoration task, the lexical reliance score was identical to that described previously (i.e., by-subject random slopes for the type by block interaction from the GLMM for each session). The acoustic-phonetic score was quantified as the average phonemic restoration effect in each session, as measured by the by-subject random slopes for type in the GLMM. That is, by-subject random slopes for type index the relationship between replaced and added items across both word and nonword blocks. A slope of zero indicates complete phonemic restoration (i.e., equal similarity ratings for replaced and added item types) where a slope greater than zero indicates incomplete phonemic restoration (i.e., lower ratings for replaced compared to added item types). We note that using by-subject random slopes for the type by block interaction cannot be considered as an isolated measure of lexical reliance; rather, the interaction coefficients might be better considered as indexing the added contribution of lexical information to phonemic restoration (indeed, this is an important feature of the task design). Critically, the derived measures of lexical and acoustic-phonetic reliance are mathematically independent. For example, the individual who shows a large type effect for both blocks *and* the individual who shows no type effect for both blocks would both show the same interaction effect; specifically, the interaction for both individuals would reflect a by-subject slope of zero given that the type effect is constant across blocks.

As shown in Fig. 4 (middle), there was a strong, positive association between the lexical and acoustic-phonetic reliance scores in both session one ($r = 0.65$, $p < 0.001$) and session two ($r = 0.88$, $p < 0.001$). Specifically, individuals with larger lexical effects showed stronger acoustic-phonetic sensitivity as indexed by less complete phonemic restoration in the aggregate. Data from three representative participants are shown in Fig. 4 to illustrate this relationship. Subject 96 shows a robust lexical effect such that near complete phonemic restoration is observed in the word block with minimal restoration in the nonword block; yet, this participant also shows high acoustic-phonetic sensitivity given the large effect of type on similarity ratings. Subject 106 shows a

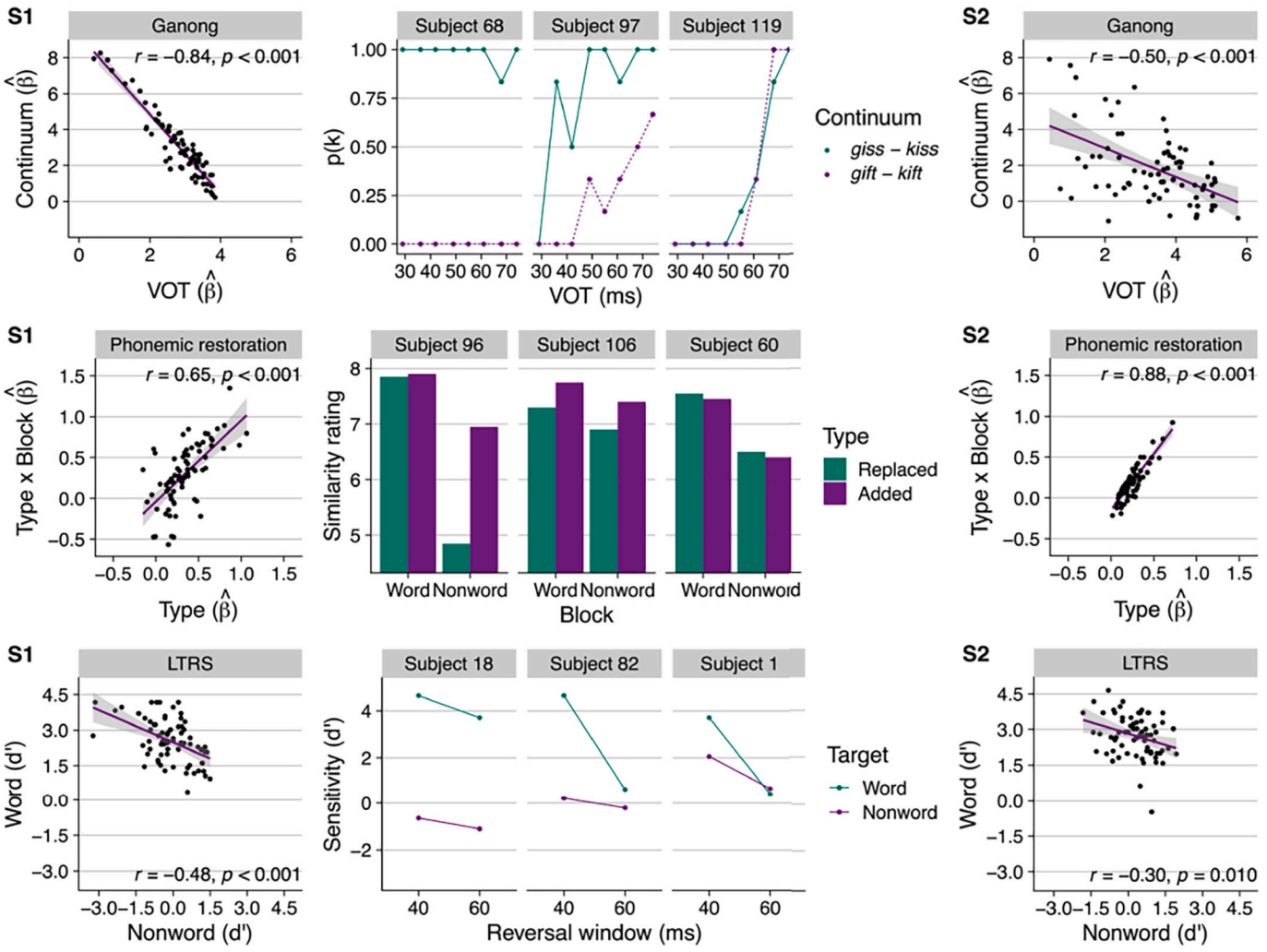

Fig. 4. Relationship between acoustic-phonetic and lexical reliance scores in the Ganong (top), phonemic restoration (middle), and LTRS (bottom) tasks for both sessions. Tasks are shown in separate rows. Within each row, the plot at left shows the association between acoustic-phonetic and lexical reliance scores at session one in addition to individual subject data for three representative subjects; the plot at right shows the association between acoustic-phonetic and lexical reliance scores at session two. In all scatterplots, black circles indicate individual participants, the purple function indicates the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit. In all representative subject plots, displayed values reflect means calculated across trials for each respective cell. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

moderate lexical effect given that the type effect is only moderately weaker for words compared to nonwords, which matches the moderate sensitivity to type in the aggregate. Finally, subject 60 shows a minimal lexical effect given that phonemic restoration for the word block is very similar to phonemic restoration for the nonword block; this subject also shows poor acoustic-phonetic sensitivity in the aggregate given near complete phonemic restoration overall.

3.4.3. *Locally time-reversed speech task*

Like the phonemic restoration task, the LTRS task does not lend itself to a strict separation of acoustic-phonetic and lexical contributions to perception. Recall that the lexical reliance score for this task is the difference in sensitivity (*d'*) for word and nonword targets. In principle, sensitivity for nonword targets – where lexical information is not available – is a reasonable metric of acoustic-phonetic sensitivity independent of lexical reliance. However, if we were to examine the relationship between sensitivity to nonword targets and the lexical reliance score (as calculated for the previous analyses), then an artificial association would emerge given the mathematical contingency between the acoustic-phonetic score (*d'* for nonword targets) and the lexical reliance score (*d'* for word targets – *d'* for nonword targets). To mathematically dissociate acoustic-phonetic and lexical scores, here we considered sensitivity for nonword targets as the acoustic-phonetic score and sensitivity for word targets as the lexical reliance score, on the logic that sensitivity for nonwords reflects performance when lexical information is not available and sensitivity for words reflects performance when lexical information is available to guide perception. Accordingly, sensitivity for word targets does not reflect an isolated lexical influence, but rather sensitivity given the added contribution of lexical information.

As shown in Fig. 4 (bottom), there was a moderate, inverse association between the acoustic-phonetic and lexical reliance scores in session one ($r = -0.48$, $p < 0.001$). In session one, individuals with larger lexical effects, as indexed by sensitivity for word targets, showed weaker acoustic-phonetic sensitivity as indexed by sensitivity for nonword targets. Data from three representative participants are shown in Fig. 4 to illustrate this relationship. Subject 18 shows high *d'* for word targets and extremely low *d'* for nonword targets, subject 82 shows moderate *d'* for both word and nonword targets, and subject 1 shows high *d'* for both word and nonword targets. In session two, a qualitatively similar pattern is observed; however, this association did not meet threshold for statistical significance following correction for multiple comparisons in

session two ($r = -0.30$, $p = 0.010$).

3.5. Comparisons between Ishida et al. (2016) and the current sample

In contrast to previous research demonstrating a reliable association between individual differences in lexical reliance for the phonemic restoration and LTRS tasks (Ishida et al., 2016), no association between these two measures was observed in the current sample at either session. Recall that the metric of lexical reliance in the current study – while conceptually identical to the method used in Ishida et al. – was calculated slightly differently. Specifically, we calculated this as by-subject random slopes for the type by block interaction from a mixed effects model that operated on trial-level data. Accordingly, the lexical reliance metric used in the current work is influenced by trial-level variability. In contrast, Ishida et al. calculated the lexical metric as follows. First, four means were calculated for each subject, one for each cell formed by crossing type and block, collapsing across all trials within each cell. Second, phoneme restoration in each block was calculated for each participant as the difference between the mean rating for replaced items and added items (i.e., mean rating for replaced items – mean rating for added items). Third, the lexical reliance score was calculated as the difference of the differences; that is, phoneme restoration in the word block minus phoneme restoration in the nonword block. With this procedure, larger lexical reliance scores indicate greater difference between the phonemic restoration effect for the word and nonword blocks.

To examine whether the null associations we observed between lexical reliance for the phonemic restoration and LTRS tasks in the current sample may reflect our mathematically (but not conceptually) different approach to calculating lexical reliance for the phonemic restoration task, a final analysis was performed to compare the current results more directly to previous work. In total, six associations were calculated for this analysis using Pearson's *r*. To control family-wise error rate, we applied the conservative Bonferroni correction to guide interpretation of *p*-values, which adjusted alpha to 0.008 given family-wise alpha of 0.05 and six comparisons.

Lexical reliance scores for the phonemic restoration task were calculated separately for each participant in the current sample at each session following the exact method used in Ishida et al. (2016). Fig. 5, panel A shows the association between lexical reliance scores on the phonemic restoration and LTRS tasks for the Ishida et al. sample (provided as supplementary material to their manuscript) and the current sample at each session. No reliable association was observed between lexical reliance scores for these two tasks at either session one ($r = 0.12$, $p = 0.308$) or session two ($r = 0.17$, $p = 0.148$).

Visual inspection of Fig. 5, panel A reveals a striking challenge for the interpretation of lexical reliance scores across the range of observed scores. Namely, there are many individuals who show *negative* lexical reliance scores, including 40% of the Ishida et al. sample, 27% of the current sample at session one, and 37% of the current sample at session two. Interpreting lexical reliance scores continuously is reasonable when they are bounded by zero at the floor; that is, for both metrics, a score of zero indicates no lexical effect, which is the lowest interpretable lexical effect for each metric. Consider the metric for phonemic restoration. A lexical reliance score of zero indicates that phonemic restoration was equal between the word and nonword blocks. A negative lexical reliance score would indicate greater phonemic restoration for *nonwords* compared to words. It is not clear what theory would predict that such a pattern of results is consistent with the interpretation that an individual is using lexical information less than someone who shows a lexical reliance score of zero. Likewise, interpretation of lexical reliance scores on the LTRS task is bounded by a floor. A score of zero indicates that sensitivity was equal between word and nonword targets, indicating no

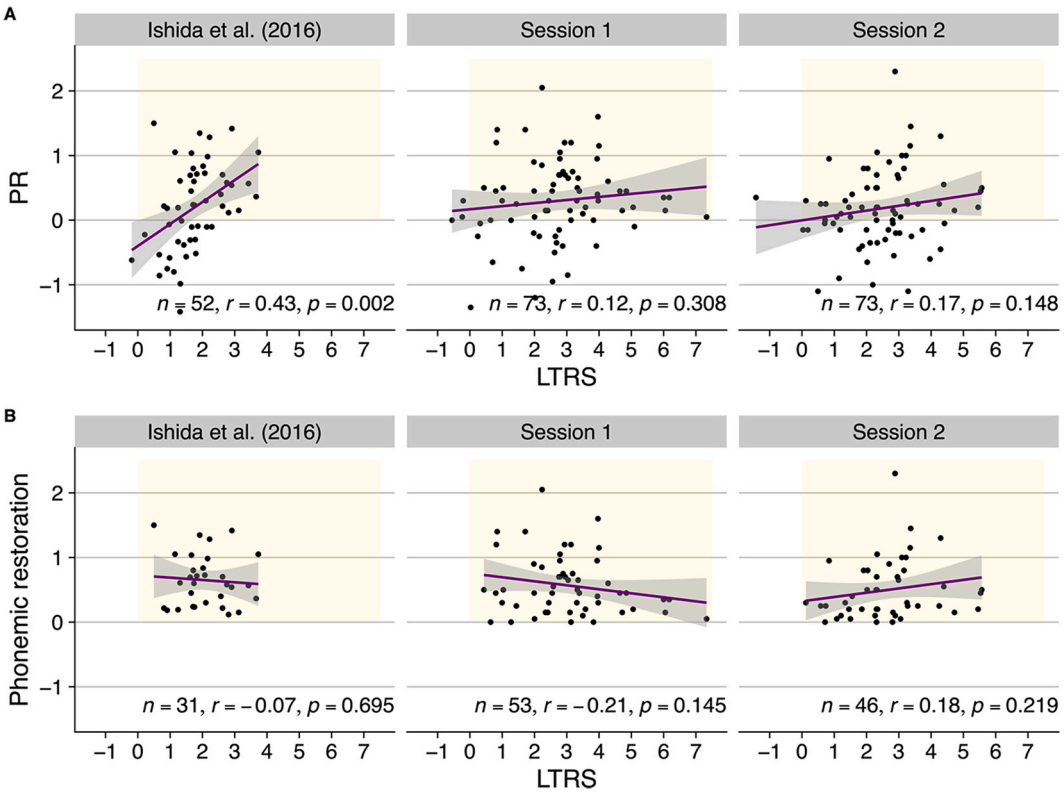

Fig. 5. Association between lexical reliance scores across the phonemic restoration and LTRS tasks for the Ishida et al. (2016) sample and the current sample at both session one and session two. Panel A shows all participants in each sample; panel B shows participants who had lexical reliance scores greater than or equal to zero for both tasks in each sample. In all plots, the region shaded in yellow indicates positive lexical reliance scores, black circles indicate individual participants, the purple function indicates the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

lexical influence on sensitivity. A negative score indicates lower sensitivity to word compared to nonword targets, which does not reasonably map to the interpretation of a weaker lexical influence compared to a score of zero. Though *interpretation* of negative lexical reliance scores is challenged (for the reasons described above), we note that their emergence in a large sample of participants (as in the Ishida et al. and current samples) is not surprising given that the scores that we observe are estimates of the true score. That is, any observed score is best considered as a "true" score convolved with a noise distribution that reflects, perhaps among other things, a degree of measurement error.

Given that interpretation of the lexical reliance score for both metrics is logically bound to zero as a floor (i.e., no lexical influence), we examined the association between lexical reliance scores for each task in each sample for the participants who showed lexical reliance scores greater than or equal to zero for both tasks. The results converged across samples. Specifically, there was no association between lexical reliance scores on the phonemic restoration and LTRS tasks for the Ishida et al. sample ($r = -0.07$, $p = 0.695$), the current sample in session one ($r = -0.20$, $p = 0.145$), or the current sample in session two ($r = 0.18$, $p = 0.220$).[4]

4. Discussion

The goal of the current work was to test the hypothesis that individual differences in lexical reliance are a stable individual trait that reflects individuals' relative use of lexical and acoustic-phonetic cues for speech perception. Participants completed three tasks designed to elicit a lexical effect on speech perception at each of two sessions separated in time. For each task at each session, we quantified individuals' reliance on lexical and acoustic-phonetic cues for speech perception. In the aggregate, strong lexical effects were observed for each task at each session, consistent with past research (Ganong, 1980; Giovannone & Theodore, 2021a, 2021b; Ishida et al., 2016; Samuel, 1981). Also consistent with past research, a wide range of individual differences in the magnitude of the lexical effect were observed (Giovannone & Theodore, 2021a; Ishida et al., 2016). For a given task, individual differences in lexical recruitment were significantly associated across sessions. Individual differences in lexical reliance showed a negative association with acoustic-phonetic reliance for two of the three tasks, consistent with the hypothesis; however, one task revealed a positive association between lexical reliance and acoustic-phonetic scores. There was no evidence to suggest that individual differences in lexical reliance were associated across tasks.

As reviewed in the introduction, one challenge for building theories that can account for individual differences in behavior is a relatively poor understanding of the psychometric properties of the tasks that are used to assess individual differences (Hedge, Powell, & Sumner, 2018; Heffner et al., 2022; Parsons et al., 2019; Rouder & Haaf, 2019; Strand et al., 2018; Wilbiks et al., 2022). The results from the current study help to address this challenge. Specifically, the Ganong and LTRS tasks showed strong test-retest reliability, and the phonemic restoration task showed moderate test-retest reliability. As discussed by Heffner et al. (2022), correlation coefficients in the range of 0.7 to 0.8 are often considered a marker of "adequate" reliability. By this metric, the Ganong and LTRS tasks demonstrate adequate test-retest reliability, and the PR task demonstrates lower than adequate test-retest reliability. Establishing that the Ganong and LTRS tasks have adequate test-retest reliability is key not only for interpreting findings of the current investigation, but also for their future use in the domain of speech perception research, as we now know that these tasks appear adequate to consistently measure individual differences. The observed associations across time for each task suggest that, at least for a given task, a listener who is more dependent on lexical information at an initial time point is also more dependent on lexical information at a later time point, providing one piece of evidence to suggest that lexical reliance is a stable individual trait.

However, we found no evidence to suggest that individual differences in lexical recruitment were associated across tasks. In session one, we observed no statistically significant correlations between the magnitude of the lexical effect across any of the three tasks. In session two, a moderate correlation between the magnitude of the lexical effect emerged only for the Ganong and LTRS tasks. Thus, in contrast to the results observed in Ishida et al. (2016), we found no evidence of convergent validity among the phonemic restoration and LTRS tasks, nor between either of these two tasks and the Ganong task. On the one hand, the lack of an association between lexical reliance scores across tasks in the current study suggests that perhaps some people are *not* more lexical than others; that is, lexical reliance may not reflect a stable individual trait. Instead, the high test-retest reliability we observed for a given task alongside poor convergent validity across tasks may suggest that lexical reliance is a flexible state that may vary within an individual depending on the situation or task. On the other hand, we must consider the possibility that the lack of convergent validity might mean that our selected tasks are not all measuring the same thing – that is, they may have low convergent validity because they are not assessing the same construct, or because lexical information might play a different role in each task.

Consider first the Ganong task. In this task, listeners are asked to identify the initial speech sound of each stimulus, and thus the phonetic identification decision is an explicit report of the perceived category. In addition, the decision is made temporally local to the stimulus. Listeners have access to (at least) two sources of information that could be used to guide their decisions – the VOT at word onset and subsequent lexical context. Both sources of information are informative for making a phonetic decision. While some listeners might respond with high fidelity to VOT, other listeners might respond in line with lexical information, and yet other listeners may more equally integrate these two sources of information. In general, acoustic-phonetic and lexical information are in high conflict with each other in the Ganong task, which may on its own encourage listeners to rely on one source of information at the expense of the other source of information. Indeed, previous research shows that the magnitude of the Ganong effect is sensitive to the input-driven conflict between phonetic and lexical cues (Bushong & Jaeger, 2019; Giovannone & Theodore, 2021b).

Now consider the phonemic restoration task. Here, listeners are asked to judge the phonological similarity between an item with signal correlated white noise (either added to or completely replacing a phoneme in the signal) and the same item without noise. The lexical effect in this task (i.e., greater phonemic restoration for words compared to nonwords) occurs because lexical information facilitates the perceptual restoration of the missing phoneme in the replaced items. Signal correlated white noise that has replaced a phoneme presumably only weakly maps to speech sound representations, if at all. Top-down lexical information is available to guide interpretation of the noise in replaced segments when the replaced noise occurs in words; however, this is not the case when the replaced noise occurs in nonwords. On this view, a parallel may be made to the Ganong task, in which midpoint VOTs provide weak activation of phonetic categories and thus top-down lexical knowledge may exert a stronger influence on perception at the expense of the role of acoustic-phonetic information. However, the phoneme restoration task requires an increased memory load compared to the Ganong task given that listeners must hold two tokens in memory to make their similarity judgments. Moreover, the decision in the phonemic restoration task is a less direct measure of linguistic perception than the Ganong task because listeners do not explicitly report which

[4] For all correlations presented in the main text, a parallel version was run in which the sample was limited to individuals who showed lexical effects greater than or equal to zero for each task in the respective analysis. The qualitative results converged in all cases. These analyses can be viewed by executing the script provided on the OSF repository for this manuscript.

sounds or words were perceived.

Finally, consider the LTRS task. Here, listeners must discriminate between the phonological string in a locally time-reversed item and in an intact item. Locally time reversed speech produces a degraded signal that potentially disrupts listeners' ability to encode the phonetic information in the reversed stimulus. If the reversed token is not accurately encoded, then the listener will struggle to accurately discriminate between reversed and intact items, leading to low sensitivity. In the aggregate, listeners show higher sensitivity for word compared to nonword targets, which suggests that lexical information *facilitates* the encoding of phonetic information from the degraded signal. That is, the role of lexical information in the LTRS task is hypothesized to serve a rather distinct function compared to the Ganong and phonemic restoration tasks, serving to facilitate acoustic-phonetic sensitivity in the former and attenuate acoustic-phonetic sensitivity in the latter. As with the phonemic restoration task, the LTRS task also introduced a memory load compared to the Ganong task because listeners must hold two items in memory to make the discrimination decision. Thus, though all three tasks elicit strong, reliable lexical influences on speech perception, they differ in numerous ways including the type of decision, the memory burden, and the role of lexical information for the task at hand. These task differences may drive lexical information not only to be used in different ways across tasks, but also to different extents.

An additional factor to consider regarding the lack of convergent validity among the Ganong, phonemic restoration, and LTRS tasks examined here is the size of the sample under investigation. The analyses presented in the main text reflect performance of 73 participants who completed both test sessions. Ishida et al. (2016) examined performance of 52 participants, and the analyses presented in the Supplementary Material examined performance of all 120 participants who completed the first test session. These sample sizes are relatively large compared to most sample sizes used for individual differences research in the domain of speech perception (see Heffner et al., 2022, for a review), though they may be considered relatively small compared to other domains including personality psychology (e.g., Schönbrodt & Perugini, 2013). The observed correlation in any given sample can be influenced by both the sample size and the "true" correlation (e.g., Schönbrodt & Perugini, 2013). For example, the variability observed across a range of samples will be relatively high for small sample sizes when the "true" correlation is low compared to the variability observed across a range of large sample sizes when the "true" correlation is high.

To provide an existence proof of this point for the lexical effects considered here, we first combined the Ishida et al. (2016; $n = 52$) and session one ($n = 120$) samples. Both samples completed the phonemic restoration and LTRS tasks and collectively form one of the largest samples ($n = 172$) for individual differences research in the speech perception domain. We calculated the association between lexical reliance scores on these tasks in the combined sample. As shown in Fig. 6, panel A, the correlation was modest ($r = 0.20$, $p = 0.010$). We do not mean to imply that this association should be taken as the "true" association; instead, we consider it as the best estimate of the true association given the available data. Then, we generated 1000 random samples of 75 participants and, separately, 1000 random samples of 100 participants from the full sample of 172 participants. As noted above, a sample size of 75 or 100 participants meets or exceeds convention for sample sizes for individual differences research in the domain of speech perception, though these sample sizes are relatively low compared to the convention in other domains, such as personality psychology. For each random sample, we calculated the association between lexical reliance scores on the two tasks.

Fig. 6, panel B shows the distribution of *r* values obtained across the 1000 random samples of 75 participants (left) and the 1000 samples of 100 participants (right). Wide heterogeneity in the obtained *r* values is observed for both sets of samples. For example, the observed association across the 1000 samples of 75 participants includes multiple *r* values of 0.00 and an *r* that was approximately twice (0.42) the magnitude of the association observed in the full sample of 172 participants (0.20). Though heterogeneity across samples is attenuated for the 1000 samples of 100 participants, it is by no means ameliorated. This example illustrates one challenge associated with individual differences research in the cognitive sciences, among others that have recently been brought to light in the context of a failure to observe reliable associations across multiple tasks for individual differences in inhibition (Hedge et al., 2018; Rouder & Haaf, 2019; Rouder, Kumar, & Haaf, 2019).

To return to the central hypothesis – that individual differences in lexical reliance reflect a stable listener trait – the lack of association between lexical reliance scores across tasks provides no evidence to support this hypothesis. However, for the reasons described above, it is difficult to ascertain whether the absence of reliable associations across tasks is a characteristic of the listeners or a characteristic of the tasks themselves. That is, these three tasks may have poor convergent validity, as has been observed for tasks presumed to measure perceptual adaptation (Heffner et al., 2022), audiovisual integration (Wilbiks et al., 2022), and listening effort (Strand et al., 2018). Teasing apart the

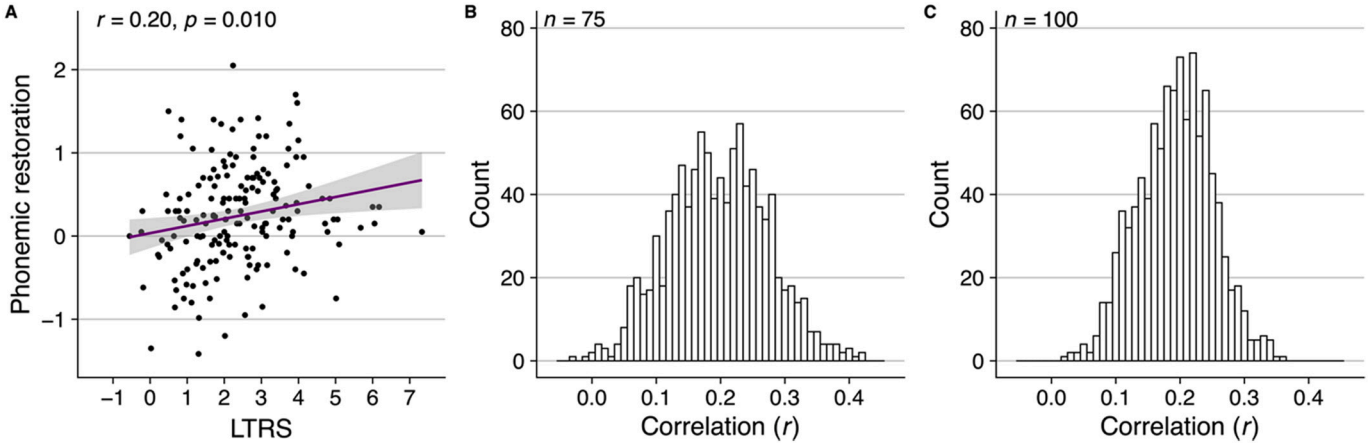

Fig. 6. Panel A shows the association between lexical reliance scores for the phonemic restoration and locally time-reversed speech (LTRS) tasks in the combined samples of Ishida et al. (2016) and session 1 of the current work. Black circles indicate individual participants, the purple function indicates the line of best fit, and the shaded region indicates the 95% confidence interval for the line of best fit. Panel B shows the distribution of correlation coefficients (*r*) reflecting the association between lexical reliance scores for these two tasks for 1000 random samples of 75 participants from the combined sample (left) and 1000 random samples of 100 participants from the combined sample (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

psychometric aspects of tasks used in the cognitive sciences from the role of individual variation is inherently challenging and completely intertwined. For example, high test-retest reliability is necessary to both establish a psychometrically sound task for use in individual differences research and to serve as evidence for stable individual differences. That is, to be interpreted as a stable individual difference, performance must remain reliably consistent across time, and consistently eliciting the same score over time is one index of high test-retest reliability. The current findings show extreme promise for the use of the Ganong and LTRS tasks in the domain of speech perception generally and for individual differences research more specifically given the extremely high test-retest reliability of these tasks. Though the current work observed no association between test-retest reliability and time between test sessions, we acknowledge that future research is needed to examine whether similarly high test-retest reliability is observed for these tasks when different stimuli are presented at each test session (noting that following Heffner et al. (2022), such a manipulation may also be considered a measure of construct validity). Though test-retest reliability is necessary, it may not be sufficient to conclude that individual differences reflect a stable trait outside of the specific task. Stronger evidence would also entail consistent performance across tasks that assess the same construct, which requires tasks that show strong convergent validity. Despite the challenges inherent in this line of investigation, we posit that a better understanding of the psychometric characteristics of our tasks is an exciting endeavor for future research because it will lay the foundation for theory that can account for the rich variability in individual behavior.

Finally, we hypothesized not only that individual differences in lexical reliance represent a stable listener trait, but that they arise due to individuals' relative use of lexical and acoustic-phonetic cues for speech perception. Consistent with this hypothesis, we found strong, inverse associations between acoustic-phonetic and lexical reliance scores for the Ganong and LTRS tasks. In both tasks, listeners who showed stronger lexical effects also showed weaker reliance on acoustic-phonetic information. In the Ganong task, ascertaining the relative influence of acoustic-phonetic versus lexical information on perception was more clearly separable than for the LTRS task, which may explain why the association between lexical and acoustic-phonetic reliance was stronger for the Ganong compared to the LTRS task. Moreover, the only evidence of convergent validity in any of the tasks was between these two tasks in session two. We do not want to overinterpret this relationship given that it was inconsistently observed, but we do draw it to the attention of the reader here.

In contrast, though a significant association between lexical and acoustic-phonetic reliance was observed in the phonemic restoration task, it was in the opposite direction as predicted by our hypothesis. That is, listeners who showed a larger lexical effect in the phonemic restoration task also showed *stronger* sensitivity to acoustic-phonetic information. In the phonemic restoration task, we considered random slopes by participant for the type by block interaction (or lexical effect) as our measure of lexical reliance, and random slopes by participant for the main effect of type (or overall phonemic restoration) as our measure of acoustic-phonetic reliance. Participants who showed a larger lexical effect in this task showed weaker phonemic restoration. Weaker phonemic restoration is indicative of higher sensitivity to acoustic-phonetic information because it indicates that the listener did not completely restore the replaced phoneme; that is, they detected its absence. Like in the LTRS task, it is difficult to isolate acoustic-phonetic cue use from lexical cue use in the phonemic restoration task. Moreover, phonemic restoration is an incredibly strong effect; indeed, phonemic restoration was near ceiling in the word block of the present study for both sessions when considered in the aggregate. The ceiling effects observed in the word block attenuate the nature of individual variability that can be observed, which may have forced the observed relationship between acoustic-phonetic reliance (i.e., the main effect of type) and lexical reliance (i.e., the interaction between type and block) even though they are in principle mathematically independent. Given the high rate of ceiling effects in this task for the word block – that is, because phonemic restoration for words is so robust – we suggest that this task may be poorly suited to assess individual differences on the level of granularity that we had intended. Future research using tasks that support a clearer separation between acoustic-phonetic and lexical cue use will be helpful to explain why some individuals may be more lexical than others.

The inverse relationship between acoustic-phonetic and lexical reliance within the Ganong task, and, to a lesser extent, in the LTRS task, has implications for clinical populations who may show increased reliance on lexical information for speech perception, including individuals with developmental language disorder, specific language impairment, and developmental dyslexia (Derawi et al., 2022; Giovannone & Theodore, 2021a, 2021b; Reed, 1989; Schwartz et al., 2013). These disorders have been etiologically associated with deficits in acoustic-phonetic processing (Joanisse & Seidenberg, 2003; Snowling, 1995, 1998). The strong trading relationship between acoustic-phonetic and lexical information observed in the typical population examined here suggests a plausible mechanism for increased lexical reliance in these clinical populations; specifically, increased lexical reliance may be a compensatory mechanism for deficits in acoustic-phonetic processing. Future research is needed to test this hypothesis directly.

In conclusion, the current study yields three primary findings regarding the role of individual differences in lexical reliance for speech perception. First, individual differences in lexical reliance are stable over time for a given task, suggesting that performance *on these tasks* reflects a stable listener trait. Second, individual differences across tasks were only weakly associated, perhaps indicative of poor convergent validity across tasks. Finally, for two of the three tasks, increased reliance on lexical information was associated with weaker reliance on acoustic-phonetic information. Collectively, these results (1) provide some evidence to suggest that individual differences in lexical reliance for a given task are a stable reflection of the relative weighting of acoustic-phonetic and lexical cues for speech perception in that task, and (2) highlight the need for a better understanding of the psychometric characteristics of tasks used in the psycholinguistic domain to build theories that can accommodate individual differences in mapping speech to meaning.

CRediT authorship contribution statement

Nikole Giovannone: Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Rachel M. Theodore:** Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

None.

Data availability

Data have been made available on the OSF (https://osf.io/dhybk/).

Acknowledgements

This work was supported by NIH NIDCD grant R21DC016141 to RMT, NSF grants DGE-1747486 and DGE-1144399 to the University of Connecticut, and by the Jorgensen Fellowship (University of Connecticut) to NG. The views expressed here reflect those of the authors and not the NIH, the NIDCD, or the NSF. Gratitude is extended to Mako Ishida, Arthur Samuel, and Takayuki Arai for generously providing their stimuli for use in the current work. Gratitude is also extended to Anne Marie Crinnion for lively and fruitful discussion on the themes presented

here.

Appendix A. Stimuli used in the phonemic restoration task. The critical phoneme in each word (i.e., the phoneme to which signal correlated white noise was added, or signal correlated white noise completely replaced) is marked in bold.

Word	Nonword	Filler word	Filler nonword
accele**r**ate	vabbelle**r**ate	**a**pprenticeship	**a**ppraggisstent
accompani**m**ent	ishimpni**m**ent	**b**ilingualism	**b**yelemmuddizzan
apoca**l**ypse	twestoka**l**ips	**c**onfederate	**c**onfizzunut
astrono**m**er	yiplono**m**er	do**m**esticity	do**m**upsezzuppee
binocu**l**ars	yaboku**l**ars	**e**xtravaganza	**e**xtrottaparta
conventio**n**al	hakzenshu**n**nel	ina**r**ticulate	ina**r**kaddussess
conside**r**ate	bupwide**r**ate	**m**agnificent	**m**agnippennust
curricu**l**um	hotricu**l**um	obs**t**inacy	obs**t**ilunzoo
develop**m**ent	sikwellup**m**ent	sku**ll**duggery	sku**ll**dassipye
disfigure**m**ent	pashigu**mm**ent	**u**nemployment	**u**nemsarlint
equiva**l**ent	mestiva**l**ent		
educatio**n**al	voplortatio**n**al		
illitera**t**e	ossitera**t**e		
enlighten**m**ent	sinfighten**m**ent		
immobi**l**ise	kwatobi**l**ise		
monoga**m**y	stantoga**m**y		
molecu**l**ar	jotteku**l**ar		
receptio**n**ist	rudipshu**n**nist		
recove**r**y	stroppuvve**r**y		
vindictive**n**ess	kwamsholtive**n**ess		

Appendix B. Stimuli used in the locally time-reversed speech task. The phoneme that was changed to create matched nonwords is marked in bold.

Word	Nonword	Word	Nonword
academic	aca**b**emic	establishment	es**p**ablishment
acceptable	ac**sh**eptable	extraordinary	extraordi**m**ary
appointment	a**t**ointment	identical	i**b**entical
capacity	ca**t**acity	incapable	in**t**apable
category	ca**p**egory	interrupted	in**p**errupted
certificate	cer**p**ificate	nobody	no**d**ody
competition	compe**p**ition	particular	par**p**icular
contemporary	con**p**emporary	predicament	pre**b**icament
decorated	de**t**orated	propaganda	propa**d**anda
department	de**t**artment	remarkable	rema**rt**able
documentary	docu**n**entary	satisfactory	sa**p**isfactory
electricity	electri**sh**ity	understandable	understan**b**able

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105320.

References

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America, 113*(1), 544–552.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*, 388–407, 1–20 https://doi.org/10.3758/s13428-019-01237-x.

Basu Mallick, D., Magnotti, F., & Beauchamp, S. M. (2015). Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review, 22*(5), 1299–1307.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

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.

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics, 61*, 30–47.

Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test–retest reliability in infant speech perception tasks. *Infancy, 21*(5), 648–667.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America, 27*(4), 769–773.

Derawi, H., Reinisch, E., & Gabay, Y. (2022). Increased reliance on top-down information to compensate for reduced bottom-up use of acoustic cues in dyslexia. *Psychonomic Bulletin & Review, 29*(1), 281–292.

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS One, 10*(4), Article e0121945.

Drouin, J. R., Theodore, R. M., & Myers, E. B. (2016). Lexically guided perceptual tuning of internal phonetic category structure. *The Journal of the Acoustical Society of America, 140*(4), EL307–EL313.

Farris-Trimble, A., & McMurray, B. (2013). Test–retest reliability of eye tracking in the visual world paradigm for the study of real-time spoken word recognition. *Journal of Speech, Language, and Hearing Research, 56*(4), 1328–1345.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*(1), 110–125.

Giovannone, N., & Theodore, R. M. (2021a). Individual differences in lexical contributions to speech perception. *Journal of Speech, Language, and Hearing Research, 64*(3), 707–724.

Giovannone, N., & Theodore, R. M. (2021b). Individual differences in the use of acoustic-phonetic versus lexical cues for speech perception. *Frontiers in Communication: Language Sciences, 6*, Article 691225.

Harper, L. V., & Kraft, R. H. (1986). Lateralization of receptive language in preschoolers: Test–retest reliability in a dichotic listening task. *Developmental Psychology, 22*(4), 553–556.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*(3), 1166–1186.

Heffner, C. C., Fuhrmeister, P., Luthra, S., Mechtenberg, H., Saltzman, D., & Myers, E. B. (2022). Reliability and validity for perceptual flexibility in speech. *Brain and Language, 226*, Article 105070.

Heffner, C. C., & Myers, E. B. (2021). Individual differences in phonetic plasticity across native and nonnative contexts. *Journal of Speech, Language, and Hearing Research, 64*(10), 3720–3733.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099–3111.

Idemaru, K., Holt, L. L., & Seltman, H. (2012). Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *The Journal of the Acoustical Society of America, 132*(6), 3950–3964.

Ishida, M., Samuel, A. G., & Arai, T. (2016). Some people are "more lexical" than others. *Cognition, 151*, 68–75.

Joanisse, M. F., & Seidenberg, M. S. (1998). Specific language impairment: A deficit in grammar or processing? *Trends in Cognitive Sciences, 2*(7), 240–247.

Joanisse, M. F., & Seidenberg, M. S. (2003). Phonology and syntax in specific language impairment: Evidence from a connectionist model. *Brain and Language, 86*(1), 40–56.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*(3), 384–422.

Mattys, S. L., Barden, K., & Samuel, A. G. (2014). Extrinsic cognitive load impairs low-level speech perception. *Psychonomic Bulletin & Review, 21*(3), 748–754.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1–86.

Miller, J. L., & Baer, T. (1983). Some effects of speaking rate on the production of/b/and/w. *The Journal of the Acoustical Society of America, 73*(5), 1751–1755.

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods, 53*(4), 1551–1562.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America, 109*(3), 1181–1196.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23*(3), 299–325.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Palan, S., & Schitter, C. (2018). Prolific. Ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27.

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science, 2*(4), 378–395.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 24*(2), 175–184.

Reed, M. A. (1989). Speech perception and the discrimination of brief auditory cues in reading disabled children. *Journal of Experimental Child Psychology, 48*(2), 270–292.

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review, 26*(2), 452–467.

Rouder, J. N., Kumar, A., & Haaf, J. M. (2019). Why most studies of individual differences with inhibition tasks are bound to fail. *PsyArXiv.*. https://doi.org/10.31234/osf.io/3cjr5

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110*(4), 474–494.

Samuel, A. G. (1996). Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General, 125*(1), 28–51.

Samuel, A. G. (2011). Speech perception. *Annual Review of Psychology, 62*, 49–72.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics, 71*(6), 1207–1218. https://doi.org/10.3758/APP.71.6.1207

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609–612.
Schwartz, R. G., Scheffler, F. L., & Lopez, K. (2013). Speech perception and lexical effects in specific language impairment. *Clinical Linguistics & Phonetics, 27*(5), 339–354.
Snowling, M. (1995). Phonological processing and developmental dyslexia. *Journal of Research in Reading, 18*(2), 132–138.
Snowling, M. (1998). Dyslexia as a phonological deficit: Evidence and implications. *Child Psychology and Psychiatry Review, 3*(1), 4–11.
Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech, Language, and Hearing Research, 61*(6), 1463–1486.
Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 7*(5), 1074–1095.
Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America, 125*(6), 3974–3982. https://doi.org/10.1121/1.3106131
Theodore, R. M., Monto, N. R., & Graham, S. (2019). Individual differences in distributional learning for speech: What's ideal for ideal observers? *Journal of Speech, Language, and Hearing Research*, 1–13.
Tzeng, C. Y., Nygaard, L. C., & Theodore, R. M. (2021). A second chance for a first impression: Sensitivity to cumulative input statistics for lexically guided perceptual learning. *Psychonomic Bulletin & Review, 28*, 1003–1014.
Wilbiks, J. M., Brown, V. A., & Strand, J. F. (2022). Speech and non-speech measures of audiovisual integration are not correlated. *Attention, Perception, & Psychophysics*, 1–11.
Winn, M. B. (2020). Manipulation of voice onset time in speech stimuli: A tutorial and flexible Praat script. *The Journal of the Acoustical Society of America, 147*(2), 852–866.
Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064–2072.