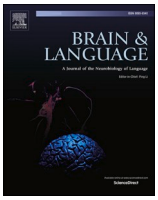
Brain & Language 226 (2022) 105070

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l



Reliability and validity for perceptual flexibility in speech

Christopher C. Heffner [a,b,c,*], Pamela Fuhrmeister [a,d], Sahil Luthra [b,e], Hannah Mechtenberg [e], David Saltzman [e], Emily B. Myers [a,b,e]

[a] Department of Speech, Language, and Hearing Sciences, University of Connecticut, Storrs, CT 06269, United States
[b] Institute for the Brain and Cognitive Sciences, University of Connecticut, Storrs, CT 06269, United States
[c] Department of Communicative Disorders and Sciences, University at Buffalo, Buffalo, NY 14214, United States
[d] Department of Linguistics, University of Potsdam, 11476 Potsdam, Germany
[e] Department of Psychological Sciences, University of Connecticut, Storrs, CT 06269, United States

ARTICLE INFO

Keywords:
Reliability
Construct validity
Individual differences
Speech perception
Phonetic identification
Phonetic learning
Phonetic adaptation

ABSTRACT

The study of perceptual flexibility in speech depends on a variety of tasks that feature a large degree of variability between participants. Of critical interest is whether measures are consistent within an individual or across stimulus contexts. This is particularly key for individual difference designs that are deployed to examine the neural basis or clinical consequences of perceptual flexibility. In the present set of experiments, we assess the split-half reliability and construct validity of five measures of perceptual flexibility: three of learning in a native language context (e.g., understanding someone with a foreign accent) and two of learning in a non-native context (e.g., learning to categorize non-native speech sounds). We find that most of these tasks show an appreciable level of split-half reliability, although construct validity was sometimes weak. This provides good evidence for reliability for these tasks, while highlighting possible upper limits on expected effect sizes involving each measure.

1. Introduction

Accurate speech perception demands flexibility on the part of listeners. It is seen in **native phonetic learning**, where listeners must adapt to phonetic signals in their native language that are different from those that are typically encountered, as when encountering a speaker with an unfamiliar accent (Xie et al., 2018) or someone speaking particularly fast (Adank & Devlin, 2010). And it can be seen in **non-native learning**, where listeners learn to categorize speech sounds not used in their native language (Heffner et al., 2019; Lively et al., 1994). Each task requires a level of flexibility on the part of a listener, and that flexibility may differ from individual to individual.

Indeed, individual differences are of key and growing interest in the field. In each case—native learning and non-native learning—unmasking the sources of individual variation can provide key insights into the mechanisms that facilitate that component of perceptual plasticity. Individuals vary in their native learning (Bent et al., 2016; Janse & Adank, 2012; Rotman et al., 2020) and their non-native learning (Chandrasekaran et al., 2010; Golestani & Zatorre, 2009; Kim et al., 2017; Schertz et al., 2015). Designs that rely on individual differences are particularly important in studies that focus on the neural architecture of perceptual flexibility, whether probing such questions through neuroimaging or through study of clinical populations. For example, researchers have explored whether individual differences in neuroanatomy can predict differences in non-native phonetic learning (Golestani, 2014) and have examined individual predictors of recovery in people with aphasia (Lazar & Antoniello, 2008). Yet these studies often leave unexplored the extent to which these individual differences are **reliable** and **valid**. To be able to successfully probe individual differences, it is necessary to understand the extent to which variation in tasks of perceptual flexibility is consistent within an individual.

For a task to be considered reliable, performance on that task must be relatively constant across different instances of that task. Common types of reliability include test–retest reliability (where a task is performed multiple times and the degree of variance is assessed) and split-half reliability (where a set of instances is split in half and performance on one half of the trials is compared to the other). Imagine a badminton player who wins a gold medal in seven consecutive Summer Olympic Games. This badminton player is clearly *reliably* good at badminton, at least in the Olympic context; despite some variation in the years,

* Corresponding author at: 122 Cary Hall, South Campus, Buffalo, NY 14214, United States.
E-mail address: ccheffne@buffalo.edu (C.C. Heffner).

https://doi.org/10.1016/j.bandl.2021.105070
Received 26 February 2021; Received in revised form 10 December 2021; Accepted 31 December 2021

stadiums, opponents, and contexts in which those games of badminton are played, that player is successful every time. Measures showing high reliability are typically required to use a task in clinical or applied contexts (Heaton et al., 2014); even some commonly used tasks such as the Stroop task have been shown to have surprisingly low reliability (Hedge et al., 2018).

For a task to be considered valid, the conclusions derived from it must map onto a consistent construct. In this paper, we focus on construct validity, the extent to which multiple measures that are purported to examine the same thing are actually tapping into a common mechanism. Decathlon is an excellent example of this in athletics. A gold medalist in the decathlon at the Summer Olympic Games has shown their prowess in a variety of athletic domains across running, jumping, and throwing. If "athleticism" is a valid construct, it makes sense to award it to the athlete who has shown consistently impressive skills across a great number of events. When designing a clinical measure, it is common to assess validity by comparing scores on that measure to already-existing tasks that are believed to tap into the same construct (Heaton et al., 2014). In this study, we probe construct validity using a relatively low bar, namely looking at the same task with two sets of well-matched stimuli. Using this approach, we are not aiming to make broader claims about the underlying constructs tapped by disparate measures, as in a different paper out of this group (Heffner & Myers, 2021), but rather ensuring that task performance can be generalized beyond one stimulus set.

Reliability is especially important for correlational studies, the approach taken by many studies of individual differences in speech perception. In experimental studies, poor reliability can complicate possible narratives that explain differences between conditions and will increase the noise present within each condition. In correlational studies, low reliability can be fatal, as reliability puts an upper limit on possible correlations between two measures of interest. This can be thought of in line with the Spearman formula (Spearman, 1904). The Spearman formula suggests that "true" correlations are attenuated by the square root of the products of the degree of measurement error in each individual measure. Under this idea, a "true" correlation of 0.7 between two measures could be observed to have a correlation as high as 0.56 when both measures have a reliability of 0.8, but as low as 0.21 when both measures have a reliability of 0.3. Lower reliability leads to more measurement error, and more measurement error leads to a lower observed correlation.

Validity also matters for the proper interpretation of perceptual flexibility studies. Having a valid measure ensures that the measure being taken references some broader concept in the field. Failing to assess validity means that some measures may offer phantom certainty about other processes or fail to generalize to a wide variety of contexts. This mistaken thinking has occurred within language processing. For instance, early studies assessing the influence of context speech rate on the perception of word-initial voicing contrasts led some researchers to propose that listeners do not use acoustic information more than a syllable away from an ambiguous segment (Summerfield, 1981). However, later studies related to segmentation, and even word-final voicing contrasts, indicates that the assumption of validity may have been mistaken. Unlike word-initial voicing, segmentation and word-final voicing are affected by context speech rate (Heffner et al., 2017).

Reliability is also related to statistical power. Therefore, measuring reliability for speech perception tasks has direct methodological implications, as reliability estimates can be used in power analyses to determine future sample sizes. Low reliability artificially depresses measured effect sizes; this, in turn, boosts the sample sizes required for adequate power. In personality psychology, where correlations are often already low, minimal sample sizes in correlational designs might be along the lines of 200 or 250 participants, depending on the requirements of the study and the assumptions made about the true underlying correlation (Schönbrodt & Perugini, 2013). Those sample sizes are almost never achieved in speech perception, where typical sample sizes within an individual population vary, ranging in a representative batch of individual differences tasks from 33 (Janse & Adank, 2012) to about 131 (Kapnoula et al., 2017), with contemporary studies often having around 60 participants (Bent et al., 2016; Golestani & Zatorre, 2009; Rotman et al., 2020). Numbers are smaller for studies of special populations, which average around 30 participants (Janse & Adank, 2012; Kim et al., 2018), and MRI studies, which typically have about 20 (Erb et al., 2012; Luthra et al., 2019). Sample sizes are often limited by practical constraints, such as money (particularly for MRI studies) or access to special populations. We therefore focus on finding highly reliable measurements that hold up even despite the low sample sizes typical of studies in speech perception.

In the current paper, we evaluate several tasks of perceptual flexibility to determine their reliability and validity, seeking highly reliable tasks. We do this both within single sets of stimuli (split-half reliability) and across two closely matched sets (construct validity). First, each task is described in terms of its background, relationship to other aspects of speech perception, and its methods. Each task was performed twice, with slight variations between each performance; say, by using different talkers or different phonetic contrasts. Consistency within a set of stimuli is assessed using split-half reliability, where the dataset is split into even and odd trials and performance on one half is correlated with the other. As our tasks often involve change over time, we decided to consistently split trials into odd and even halves, to ensure that the speech perception measure being assessed was reliable within the experiment. Consistency across these stimulus sets is assessed to test the construct validity of each measure. To the extent that performance is similar across two sets of stimuli for an identical methodology, it suggests that those methods are reliably indexing some construct that can explain individual differences in perceptual flexibility.

What does it mean to show "excellent evidence of reliability"? For the purposes of this paper, we will consider evidence for reliability to be "excellent" if an observed split-half reliability for a task is greater than 0.7, and the 95% confidence interval for the correlation does not include values lower than 0.5. Most authors suggest that correlation coefficients of at least 0.7 or 0.8 are sufficient for "good reliability", though for certain clinical purposes, a higher threshold may be needed (Hedge et al., 2018; Nunnally, 1978; Parsons et al., 2019). Perhaps surprisingly, this criterion does not require particularly large sample sizes. Indeed, some estimates suggest that it might require as few as 20 participants to produce an estimate of a 0.7 correlation with precision +/- 0.2 (Schönbrodt & Perugini, 2013). Other estimates have suggested a sample size of approximately 30 is necessary to uncover an estimate of a 0.7 correlation with precision +/- 0.2, or a 0.8 correlation with precision +/- 0.15 (Moinester & Gottfried, 2014). As such, we largely confine ourselves to sample sizes around 30 participants.

This decision to keep sample sizes small reflects the overall goal of the paper. We do not hope to estimate every correlation, regardless of its magnitude, with great precision. Nor do we make strong claims about comparing relative reliability estimates across tasks. We instead want to find the *most reliable* tasks, the ones that can lead to reasonable estimates of individual differences in correlational studies across a variety of different sample sizes. Additionally, each task will have the validity calculated, as tasks will be performed twice within the same participant using slightly different materials. Validity, in contrast to reliability, has had less written about it, and so we cautiously interpret the values that we obtain. However, validity estimates of 0.33 were considered "moderate" and of 0.74 were considered "good" for the UK Biobank dataset (Fawns-Ritchie & Deary, 2020).

2. Experiment 1

As listeners receive information about how a particular talker produces their speech sounds, they continually update their set of beliefs about how acoustic information maps onto phonetic categories (Saltzman & Myers, 2021; Tzeng et al., 2021). Previous work has suggested

that this belief-updating process is relatively talker-specific, with listeners maintaining separate *generative models* (sets of beliefs) for individual talkers (Kraljic & Samuel, 2007; Luthra et al., 2021) or groups of similar talkers (Kleinschmidt, 2019). Here, participants completed two sessions of a lexically guided perceptual learning task, hearing different talkers on each day. We assessed construct validity by testing whether the size of an individual's learning effect for one talker's voice was comparable to the size of their learning effect for the other talker's voice. We also assessed split-half reliability by examining the learning effect within each session.

2.1. Materials and methods

2.1.1. Procedure

During each session, participants completed two exposure blocks and two test blocks, presented in alternating order (exposure/test/exposure/test). During exposure blocks, participants heard words with an ambiguous fricative (an /s/-/ʃ/ blend) presented in lexically biased contexts, with the specific bias held constant throughout a block. For instance, in an /s/-biased exposure block, participants heard 16 words with an ambiguous fricative (noted "?") in /s/-biased contexts (e.g., ab?ent) and 16 words with an unambiguous fricative from the contrastive category (e.g., friendship). Participants completed both an /s/-biased block and an /ʃ/-biased block, with the order counterbalanced across participants. During each exposure block, participants were asked to indicate whether each item contained the /n/ phoneme to encourage participants to attend to phoneme-level cues. After each exposure block, participants completed a test block where they performed phonetic categorization of stimuli from a 7-step sign-shine continuum. Participants heard 10 repetitions of each step in random order for a total of 70 trials.

We used two talkers' voices for this experiment, with talker held constant throughout the entire session. A week later, eligible participants were invited back to complete a second session, which differed only in the voice of the talker. The specific talker heard during each session was counterbalanced across participants. The order of exposure blocks was held constant across sessions, but the specific order was counterbalanced across subjects; thus, if a participant received the /s/-biased exposure block before the /ʃ/-biased exposure block in Session 1, they also received the /s/-biased block first in Session 2. Response mappings were counterbalanced across participants but held constant across sessions for a given participant. Our experiment was programmed using the Gorilla online experiment builder (Anwyl-Irvine et al., 2020).

2.1.2. Participants

We recruited 77 participants from Prolific. All participants were native speakers of North American English, reported no hearing or vision impairment, currently lived in the United States, had no language or neurological disorders, had a participant rating on Prolific above 90 (maximum of 100), used a desktop computer, and had not participated in previous studies from our group using these stimuli. All participants reported being monolingual. Following previous work (Luthra et al., 2021), we rejected participants if they failed to respond on 10% of trials during either the exposure or test blocks, reached less than 70% accuracy on classifying the endpoints of the *sign-shine* continuum during test, or failed a headphone screening test (Woods et al., 2017) twice. This resulted in 45 participants with usable data in Session 1. These participants were invited to participate in Session 2 of the experiment. 32 participants (15 female, 17 male; mean age = 29, age range = 19–45) completed a second session, doing so roughly one week after their first session (M = 8 days; SD = 3 days; range = 7–22 days). The same data quality checks (including the headphone screener) used in Session 1 were applied to Session 2, and all 32 participants who completed the second session had usable data. All participants provided informed consent at the start of Session 1 via a digital information sheet per guidelines set by the University of Connecticut's institutional review board. Payment was set to $3.33 per session, with a maximum of $6.66 given if both sessions were completed. Participants generally required 15–20 min to complete each session.

2.1.3. Materials

Stimuli were taken from Luthra et al. (2021). For the exposure task, there were a total of 32 total items, 16 with a medial /s/ (e.g., *absent*) and 16 with a medial /ʃ/ (e.g., *friendship*); approximately two-thirds of these items contained the target phoneme /n/. There were two versions of each item: one with an unambiguous fricative (*absent*) and one with an ambiguous fricative (*ab?ent*). Stimuli for the phonetic categorization task consisted of a 7-step continuum from *sign* – *shine*.

Stimuli had initially been recorded by a female native speaker of English and were edited in STRAIGHT (Kawahara et al., 2008). The resultant stimuli were morphed to male tokens using the "Change Gender" tool in Praat (Boersma & Weenink, 2001), altering both pitch and formant ratio to yield a convincingly male talker. Note that the acoustics of the fricatives therefore differed between talkers; for instance, the spectral center of fricative for the female talker had a mean value of 7415 Hz (*SE*: 103 Hz) while the spectral center of the fricative for the male talker had a mean value of 5623 Hz (*SE*: 82 Hz).

2.1.4. Analysis

Data were analyzed using R (R Core Team, 2020). We first analyzed phonetic categorization data with a logit mixed effects model that tested for fixed effects of Session (Session 1/Session 2, sum-coded with a [1, −1] contrast), Bias (s-bias/sh-bias, sum-coded with a [1,-1] contrast), and Step (scaled). We used the maximal random effects structure (Barr et al., 2013), which included random by-subject slopes and interactions for Session, Bias and Step as well as random by-subject intercepts. This was also the most parsimonious model, as a simpler random effect structure resulted in poorer fit (Matuschek et al., 2017). Models were fit using the *glmer* function in the "lme4" package (Bates et al., 2016), and we used the *mixed* function in the "afex" package (Singmann et al., 2021) to evaluate the significance of each fixed effect using likelihood ratio tests.

To assess construct validity and split-half reliability, we fit psychometric curves using the package *quickpsy* (Linares & López-Moliner, 2016) for each test block (e.g., the test block after the "s-biasing" exposure on the first session). We focused on the crossover point of the logistic function, which reflects the location of the phonetic category boundary. Construct validity was assessed by measuring the size of the learning effect (i.e., the difference between phonetic category boundaries in the "s-biasing" and "sh-biasing" conditions) across sessions. Split-half reliability was computed analogously, using data from within a single session.

2.2. Results

A mixed effects model analyzing the phonetic categorization data (Fig. 1) yielded a significant Bias × Session interaction, $\chi^2(1) = 8.93$, $p = 0.003$, as well as significant effects of Bias, $\chi^2(1) = 6.22$, $p = 0.01$, and of Step, $\chi^2(1) = 72.64$, p less than 0.0001; no other fixed effects were significant. This is in line with the idea that perceptual learning was stronger in Session 1 than Session 2.

To assess the stability of phonetic recalibration across sessions (i.e., construct validity), we measured the learning effect in each session by computing the difference in phonetic category boundary between each biasing condition ("s-biasing" minus "sh-biasing"). The correlation in the size of the learning effect across sessions reflects our measure of construct validity (results found in Fig. 1). There was no detectable relationship in boundary difference across sessions, $r = 0.079$ (95% CI: -0.278 - 0.416).

We also examined split-half reliability within each session by measuring the size of the learning effect using only odd trials or using only even trials. We found moderate split-half reliability in Session 1, r

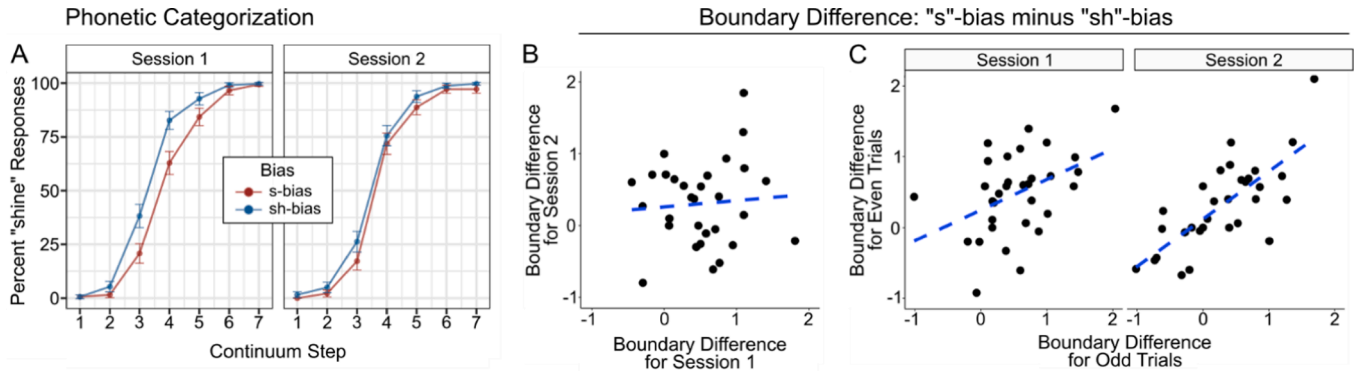


Fig. 1. Results from the lexically guided perceptual learning task, in which listeners heard a different talker on each session. (A) shows phonetic categorization functions for each session and biasing direction (red lines indicate categorization following s-biased talkers and blue lines following sh-biased talkers). x-axis lists each step from a clear "sign" (step 1) to a clear "shine" (step 7), and the y-axis indicates percent "shine" responses. (B) shows construct validity, and (C) shows split-half reliability. In (B) and (C), each black dot is a single participant, and the dashed blue line indicates the linear best-fit line.

= 0.436 (95% CI: 0.103 - 0.681), and strong reliability in Session 2, r = 0.743 (95% CI: 0.532 - 0.867).

2.3. Discussion

We found weak construct validity in a phonetic recalibration task, as an individual's degree of recalibration for one talker did not predict their degree of recalibration for another talker. We observed moderate evidence for split-half reliability in an initial experimental session and strong evidence for split-half reliability in a second session one week later.

The results parallel those of Saltzman and Myers (2021), who observed smaller perceptual learning effects in a second session that took place roughly a week after the first session. Critically, however, participants in Saltzman and Myers's study heard the same talker in both sessions, whereas participants in the current experiment heard a different talker in each session. The present results are also striking given previous work showing that listeners can maintain separate sets of beliefs for these two talkers (Luthra et al., 2021). The strong split-half reliability value for the second session suggests that phonetic recalibration effects can be reliably measured in the short-term, even if the size of these effects may change over a longer time period.

3. Experiment 2

Experiment 2 examined the perceptual flexibility necessary to perceive accented speech. Non-native speakers, even with high proficiency, continue to produce speech sound categories that differ from their native-speaking peers (Runnqvist et al., 2011). The ability of a listener to adapt to these differences is subject to variation (Clarke & Garrett, 2004). To study individual differences in these abilities, we used an adaptive staircase procedure, where the amount of noise in the background of a non-native talker's speech was systematically manipulated to determine the point at which a listener understood some (although not necessarily all) of the content words within the sentence.

3.1. Methods

3.1.1. Participants

47 participants were recruited for Experiment 2: 18 via Amazon's Mechanical Turk (MTurk) service, 16 from Prolific, and 13 from the University of Connecticut (UConn) in the United States (including two who were tested remotely). The sample from Prolific in this and subsequent studies was collected after initial analysis to raise sample sizes to be greater than 30, in line with reviewer feedback. Although when performing null hypothesis significance testing it can be problematic to recruit additional participants after initial analyses, due to the risk of inflated Type I error, the present study did not use this approach. Therefore, there is no additional risk of Type I error with this approach; adding the sample makes the estimates more precise. One participant from MTurk and one from UConn did not have usable data due to experimenter error; two participants from MTurk and two from Prolific were excluded for being non-monolingual; four participants from Prolific were excluded for learning another language before age 13; and one participant from Prolific was excluded for having lived in Italy for 9 months. This left a total of 36 (15 MTurk; 9 Prolific; 12 UConn) monolingual participants (13 male, 23 female). The average age was 25.5 years (range = 18–41). All reported being native speakers of English and no history of hearing impairment. In this study, we did not ask about a history of speech/language disorders. UConn participants were given course credit for participation; MTurk and Prolific participants were paid $10. In this and subsequent experiments, UConn participants wore headphones; Prolific participants were given a headphone check; MTurk participants were asked to wear headphones, but no headphone check was performed.

3.1.2. Materials

In the accent learning task, participants completed an adaptive staircase procedure, which assessed how much noise a participant could tolerate while still comprehending accented speech. Participants heard sentences that were based on the Harvard sentences (Egan, 1948). The content words from the original set of sentences were scrambled, leaving sentences that were grammatical but semantically anomalous. For example, participants might hear a sentence such as "a first shore is the fresh day." Participants heard two talkers: one male native speaker of Italian, and one female native speaker of Belarusian. As a proxy of accented speech perception, we added speech-shaped noise to the signal, with the level of noise changing from trial to trial. Possible signal-to-noise (SNR) values ranged from −14 dB to + 24 dB. For each talker, each participant heard 15 practice sentences recorded by a native speaker of English before hearing six blocks of 15 sentences from one of the talkers; the task then restarted with 15 practice sentences and the other talker.

3.1.3. Design and procedure

Participants heard a sentence and were asked to judge which of three sets of paired content words were in the sentence they heard (for example, for the sentence "a first shore is the fresh day", the listener might have to choose between "first" and "force", "shore" and "store", and "fresh" and "press"). The noise level started at a signal-to-noise ratio (SNR) of +8 dB. If participants were accurate for all word pairs on two consecutive trials, the noise level went up; if they were incorrect on at least one pair, the noise level went down. Step sizes started at 6 dB; after two reversals, the step size decreased to 2 dB. The staircase then reset

itself at the beginning of each 15-trial block. After six blocks of one talker, the participant then heard the second talker; the order of talkers was counterbalanced across participants. The task lasted approximately 60 min total.

3.1.4. Analysis

In theory, participants who could tolerate more noise in the signal were adapting better than participants who could not tolerate as much. Thus, the average reversal point for noise across the experiment was used as an index of comprehension. Participants who were able to tolerate more noise in the signal had reversal points that reflected a low SNR and more noise (and, thus, a higher score on accent learning), while participants who required lower noise levels to comprehend had reversal points that reflected the high SNR necessary for understanding. Of course, by using noise as a proxy to measure accent perception, it should be noted that participant responses also index a more-basic ability to withstand noise when understanding a signal. We mirror previous studies in the accent learning literature when making this assumption (Adank & Janse, 2010).

3.2. Results

Fig. 2 shows the split-half reliability and construct validity for the accent learning task, as shown in terms of the average reversal point for the SNR for each talker. It appears that the male native speaker of Italian was generally harder to understand than the female native speaker of Belarusian, given that listeners stabilize at a lower SNR (higher noise) for the female talker than the male talker. Both talkers had roughly equivalent split-half reliabilities: $r = 0.759$ (95% CI: 0.573 - 0.871) for the Italian native speaker and $r = 0.808$ (95% CI: 0.652 - 0.898) for the Belarusian native speaker. The estimate of construct validity using Pearson's r across the native Italian and Belarusian talkers was 0.531 (95% CI: 0.246 - 0.732).

3.3. Discussion

Reliability was excellent for both talkers, with estimates for each talker being higher than 0.7 and with neither confidence interval including 0.5. The estimate for construct validity (0.53) suggests that at least some of the variance in the noise level tolerated for the native speaker of Italian can be explained by the variance in the noise level tolerated for the native speaker of Belarusian.

4. Experiment 3

Experiment 3 also tested the reliability of native learning; in this case, rate learning. From talker to talker and sentence to sentence, there is variation in the rate of speech (Bosker, 2017; Schwab, 2011); despite this variation, listeners are able to flexibly track the speech rate of others to understand what they say. A frequent approach to study this particular type of perceptual flexibility is to examine the perception of speech that has been artificially and uniformly compressed in rate; studies have found that some listeners can still accurately perceive speech information with a duration as short as 25% of the originally recorded rate (Adank & Janse, 2009; Banks et al., 2015; Peelle & Wingfield, 2005). Like Experiment 2, this study used an adaptive staircase procedure to identify average reversal points for each participant that represent their individual rate learning abilities.

4.1. Methods

4.1.1. Participants

36 participants were recruited for Experiment 3: 27 from MTurk, 5 from Prolific, and 3 from UConn. One participant from MTurk dropped out between the two sections of the study. Two participants from MTurk and one from Prolific were excluded due to being non-monolingual, while one participant from UConn was excluded for learning another language before the age of 13. This left a final total of 29 (24 MTurk; 4 Prolific; 2 UConn) monolingual participants (17 male, 13 female). All participants self-reported being native English speakers with no history of hearing loss; in this study, we did not ask about a history of speech/language disorders. UConn participants were given course credit for participation; Prolific participants were paid $10.00; and MTurk participants were paid $6.67. Pay was slightly different between MTurk and Prolific due to differences in the exact time taken across both platforms. One Prolific participant left their age blank on the demographics form; the remaining participants had an average age of 29.8 years (range = 19–49).

4.1.2. Materials

In the rate learning task, sentences were recorded by two native speakers of American English, one male and one female; 105 sentences from each speaker were used for the lists below, with 15 forming a practice block before participants heard compressed sentences. Sentences were time-compressed using the Pitch Synchronous Overlap and Add (PSOLA) method in Praat (Boersma & Weenink, 2001).

4.1.3. Design and procedure

Participants listened to a series of speeded sentences. After hearing each sentence, participants were given three sets of paired content words and were asked which member of each pair they heard. The amount of compression for each sentence was determined through a staircase procedure based on logarithmic steps. Compression levels along a logarithmically spaced 20-step continuum between 20% and 56% time compression for the male speaker and between 24% and 67% time compression for the female speaker (ranges based on pilot testing).

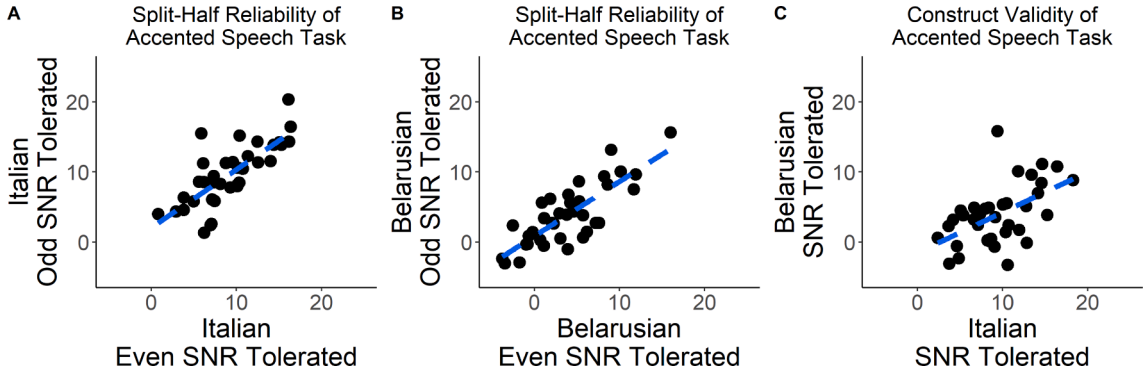


Fig. 2. Average reversal point, in SNR ratio, across the experiment for different subsets of items. (A) shows split-half reliability for the native speaker of Italian, (B) shows split-half reliability for the native speaker of Belarusian, and (C) shows construct validity for the native Italian and native Belarusian talkers, within participants. Each point is a single participant. The dashed blue line shows the linear best fit line.

The logarithmic spacing ensured the ratio of the duration of one step to the next was constant; each step was approximately 95% the duration of the next longest step. If all probe words were reported correctly for two trials in a row, the degree of compression was increased, while if any word was reported incorrectly on a single trial, the amount of compression was decreased. Changes in the level of compression started at four continuum steps but were decreased to one after two reversals. This procedure repeated across six 15-trial blocks before the talker switched; the order of talkers was counterbalanced across participants. The task lasted approximately 40 min in total.

4.1.4. Analysis

In line with some common procedures for adaptive staircase studies (Amitay et al., 2006; Kollmeier et al., 1988), the average reversal point across the experiment was used to evaluate the amount of learning done over the course of the study. This was calculated across blocks. We used this rather than, for example, the endpoint at which participants converged at, because average reversal level ended up being more reliable within individual participants. Participants with poorer learning of speeded speech had reversal points at a relatively slow rate of speech, while participants who could tolerate more compression had reversal points at a relatively fast rate of speech.

4.2. Results

Fig. 3 shows split-half reliability and construct validity for the average reversal level. Construct validity, as measured using Pearson's *r*, was 0.657 (95% CI: 0.388 - 0.822). Participants who are better able to tolerate rate compression for one talker were also better able to tolerate it for the other. Split-half reliability for the male talker was 0.856 (95% CI: 0.716 - 0.929), while it was 0.766 (95% CI: 0.560 - 0.882) for the female talker.

4.3. Discussion

Reliability and construct validity were both high for the compressed rate learning study. For both the male talker and female talker, reliability was high, and there was a strong correlation in average reversal points across the different talkers.

5. Experiment 4

One situation in which perceptual plasticity is necessary is in the acquisition of new phonetic categories. Non-native phonetic learning requires listeners to flexibly adapt to new ways to categorize phonetic space, whether through explicit instruction or incidental acquisition of those categories. In Experiment 4, we examined the within-participant reliability of an incidental non-native phonetic learning task based on the Systematic Multimodal Associations Reaction Time (SMART) task of auditory category learning (Gabay et al., 2015). In the SMART task, learners are told their job is to press a button on a keyboard corresponding to the location of a visual target. What they are not told is that there is a lawful relationship between the location of the target (and, thus, the button press required to hit the target) and speech sounds that precede the appearance of those targets, such that phonetic categories can predict button presses. Learners may then incidentally pick up on that correlation. When those pairings are scrambled, learners should slow down, leading to an increase in reaction times. In Gabay et al. (2015), differences in the magnitude of reaction time costs across conditions were treated as indicative of differences in the quality of learning across those conditions; if this is true, the magnitude of reaction time costs could also be used as an index of the degree of perceptual flexibility within individuals.

5.1. Materials and methods

5.1.1. Participants

51 participants were recruited for Experiment 4: 15 from MTurk, 17 from Prolific, and 19 from UConn. All participants self-reported being native English speakers with no history of hearing loss; in this study, we did not ask about a history of speech/language disorders. 12 participants (3 from MTurk, 7 from Prolific, and 2 from UConn) were excluded due to their performance on the task, detailed in the Analysis section, and an additional 6 participants (3 from Prolific, 3 from UConn) were excluded for the following reasons: one UConn participant was missing a demographics form; two UConn participants and one Prolific participant were excluded for learning another language before the age of 13; and two Prolific participants were excluded for their experience with languages relevant to the non-native language contrast being learned. Of the remaining 33 (12 MTurk; 7 Prolific; 14 UConn) monolingual participants (11 male, 22 female), the mean age was 25.3 years (range = 18–44). UConn participants were given course credit for participation; MTurk and Prolific participants were paid $10.

5.1.2. Materials and design

Participants learned categories in one of two non-native continua: a fricative continuum ([x] to [ç]), taken from German and described in greater detail in Heffner et al. (2019), and a geminate continuum, taken from Arabic, where the length of a [t] sound between two vowels was successively increased to create intermediate steps. These continua were chosen to minimize the overlap in the acoustic information learned in each contrast, as the fricative contrast relies on a spectral shift, whereas the geminate contrast relies on detection of timing information. Each continuum used two endpoint tokens and seven intermediate steps, making it, all told, a nine-step continuum.

The incidental task was a version of the SMART task (Gabay et al.,

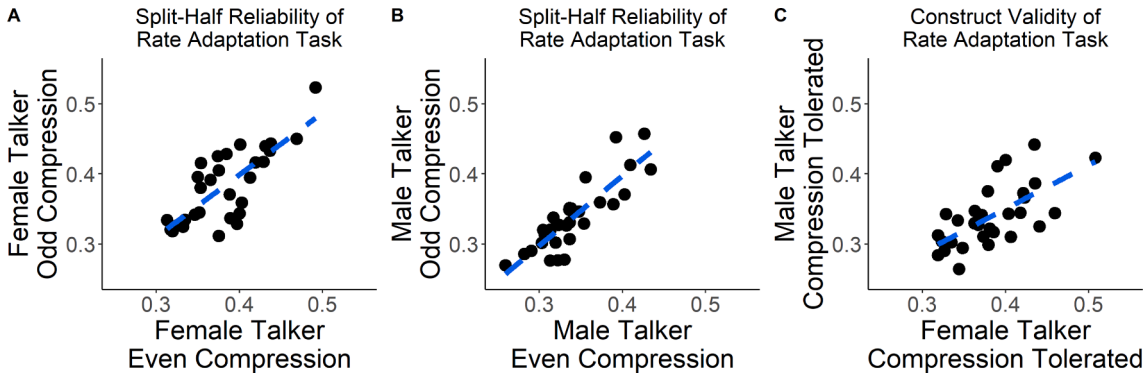


Fig. 3. Average reversal point for different subsets of items. (A) shows split-half reliability for the male talker, (B) shows split-half reliability for the female talker, and (C) shows construct validity for the male and female talkers, within participants. Compression level is indicated as a proportion of the original duration of the clip. Each point is a single participant. The dashed blue line shows the linear best fit line.

2015). Participants pressed buttons corresponding to the appearance of a zombie in one of three locations on the screen. Each visual target was preceded by six tokens sampled from one of three evenly spaced categories along the stimulus continuum. The first three stimuli along the continuum were mapped to a single location on the screen; the next three to a second location; and the final three to a third location. This pairing of sounds to locations was meant to match previous studies conducted using this paradigm and the explicit learning paradigm of Experiment 5 (Heffner et al., 2019), specifically the "Neapolitan" condition of the Heffner et al. (2019) paper. Unlike in German, where only two categories are found in this continuum with a perceptual allophonic boundary roughly corresponding to the boundary between the sixth and seventh stimuli on the continuum (Key, 2014), this meant that participants were learning three categories along this fricative continuum. The boundaries in the geminate continuum were chosen to match the configuration of the fricative continuum.

The mapping of sound categories to locations was counterbalanced on a by-participant basis. Because the six sounds were a reliable cue to the location where the zombie would appear, listeners in theory could take advantage of this information to anticipate the appearance of the target and facilitate a faster response. There were 24 trials in each of nine blocks, with opportunities for breaks between each block. To test whether listeners had learned the association between the sound category information and the zombie location, sounds in the seventh block occurred with a random relationship to the location. If listeners had been using the reliable relationship between sounds and zombie in the preceding blocks, reaction times should slow once the sound-to-location relationship is disrupted in the randomized block.

5.1.3. Procedure

Participants first performed the incidental task using one stimulus continuum (either German or Arabic), and then, after a short break, performed it again with the other continuum. The experiment lasted approximately 55 min total.

5.1.4. Analysis

The outcome measure for this task was the difference in reaction times between the randomized block and the immediately previous two blocks (the baseline blocks). To compensate for differences in baseline reaction times, this was expressed as a log-transformed ratio of the reaction time in the randomized block to the reaction time in the baseline blocks, which will be referred to as the "key ratio". A key ratio of 0 represented no increase in reaction times (and, thus, no evidence for learning), while increasing scores were meant to indicate increasing learning. Trials were excluded from further analysis if responses for that trial were incorrect (1.6%), faster than 100 ms (1.5%), or slower than 1500 ms (0.6%). Participants who had more than 10% of trials excluded during any single one of the randomized or baseline blocks were excluded from further data analysis, leading five participants to be excluded from subsequent data analysis.

5.2. Result

The mean key ratio across both stimulus types was positive (M = 0.116 for the fricatives and 0.0876 for the geminates), suggesting that, on balance, listeners did show some learning of the sound-to-location pairings in this task. This positive ratio reflects raw reaction time increase during the critical block (M = 425 ms for fricatives and 443 ms for geminates) compared to the baseline blocks (M = 384 ms for fricatives and 404 ms for geminates). The estimate of the construct validity of this measure (depicted in Fig. 4) was 0.263 (95% CI: -0.088 - 0.556), with split-half reliability being excellent for the fricative stimuli (0.869; 95% CI: 0.750 - 0.934) but only middling for the geminate stimuli (0.648; 95% CI: 0.391 - 0.811).

5.3. Discussion

The task of incidental learning showed middling (for the geminates) to excellent (for the fricatives) split-half reliability for this measure, but construct validity was poor. The correlation observed between performance on each continuum was weak. Responses during the participants' debriefings, as well as the reasons for the exclusion of individual trials enumerated earlier, suggests one rationale for why the construct validity of this measure may have been relatively low. Some participants seemed to be paying little attention given the simplicity of the task, while others indicated that they used explicit strategies to learn the pairings. For example, some participants said that they noticed that hearing some sounds would allow them to predict where a visual target appeared on the screen. This combination may have meant that different participants were using different strategies across the two sections of the study.

6. Experiment 5

Experiment 5 looked at the reliability of an explicit non-native learning task (Heffner et al., 2019). In explicit learning tasks, learners are told their task straightforwardly: their job is to match speech sounds with colored squares, which stand in for different categories. They then get explicit feedback that should help to guide their future responses. Although further from non-native learning in naturalistic contexts, these tasks are also far more common in the literature (Chandrasekaran et al., 2010; Golestani & Zatorre, 2009; Perrachione et al., 2011), and still give insight into perceptual flexibility in non-native learning. This task used the same phonetic continua as the incidental task (Experiment 4).

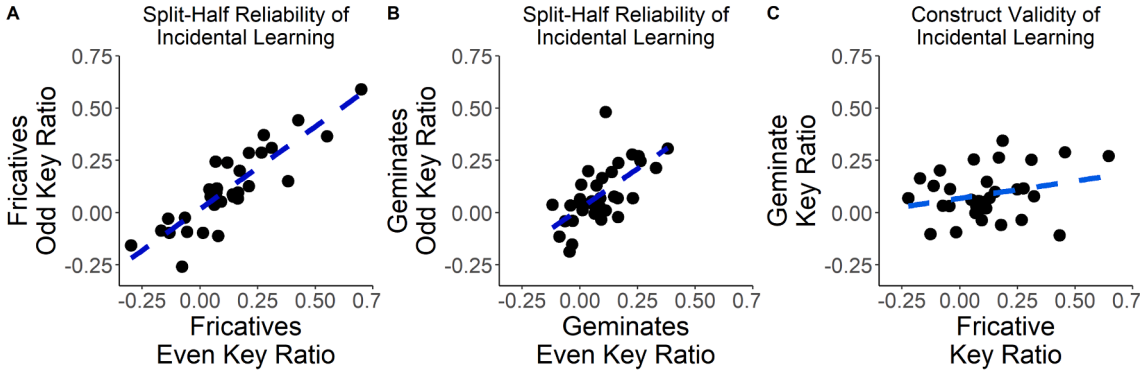


Fig. 4. Relationships in the key ratio for specific groups of items. (A) shows split-half reliability for the fricative stimuli, (B) shows split-half reliability for the geminate stimuli, and (C) shows construct validity using the fricative stimuli and geminate stimuli within the same participants. Each dot is an individual participant. The dashed blue lines indicate a linear best fit line.

6.1. Methods

6.1.1. Participants

A total of 47 participants were run in Experiment 5, including 14 from UConn, 17 from Prolific, and 16 from MTurk. All participants self-reported being native English speakers with no history of hearing loss; in this study, we did not ask about a history of speech/language disorders. Two UConn participants and six Prolific participants were excluded for learning another language before the age of 13; two MTurk participants were excluded for being non-monolingual; two Prolific participants were excluded for exposure to languages relevant to the phonetic contrasts being learned; and one MTurk participant was excluded for a missing continuum and demographics form. The remaining 34 (12 UConn, 9 Prolific, 13 MTurk) monolingual participants (11 male, 23 female) had an average age of 27.0 years (range = 18–48). UConn participants were given course credit for participation; MTurk participants were paid $10.00; and Prolific participants were paid $8.35. Pay was slightly different between MTurk and Prolific due to differences in the exact time taken across each platform.

6.1.2. Materials and design

The materials used were identical to those created for the incidental paradigm, save that there were 10 steps along the continuum in the present experiment rather than the 9 used in the incidental task. This difference in the number of steps reflects differences in the experimental design of previous studies using these tasks (Heffner et al., 2019). Participants were told to pair speech sounds to one of three colored squares displayed on the screen. Listeners first heard one of a set of ten speech sound tokens. The first three stimuli along the continuum were mapped to the first category, the middle four to a second category, and the final three to a third category. Participants then paired this token with one of three colored squares that served as stand-ins for category labels representing three separate categories. Explicit feedback in the form of a check or X, displayed for 750 ms, followed the participants' responses.

6.1.3. Procedure

Like for Experiment 4, the order of continua presented for the explicit task was counterbalanced across participants. The experiment lasted approximately 50 min total.

6.1.4. Analysis

For the explicit learning task, the outcome measure was the average accuracy across the whole experiment. This metric accounts for the learning speed and the endpoint of learning for each participant. Participants who learned slowly and those who would not reach a high level of accuracy would thus receive low scores, while only participants who learned quickly and maintained a high level of performance would receive high scores. Scores were expected to range from chance (0.33) to perfect performance (1.00).

6.2. Results

On average, as can be seen in Fig. 5, on average, participants successfully learned both the fricative categories ($M = 0.714$) and geminate categories ($M = 0.638$). The estimate of construct validity was 0.656 (95% CI: 0.408 - 0.814). Individual performance in this task is often very steady within a single participant across sound categories being acquired, especially within the same continuum: splitting each continuum into odd and even trials showed excellent split-half reliability for both fricatives (0.957; 95% CI: 0.916 - 0.979) and geminates (0.936; 95% CI: 0.875 - 0.968).

6.3. Discussion

Explicit non-native learning showed excellent reliability. For both fricatives and geminates independently of each other, participants showed reliable estimates of explicit learning. Construct validity was also good. This indicates that the simple explicit learning task here provides reliable information about individual differences.

7. General discussion

This set of studies set out to identify the reliability and validity of a variety of tasks used to assess perceptual flexibility in speech sound categories. A total of five studies were run and are summarized in Table 1. For each study, measures of both split-half reliability and construct validity were calculated. A variety of tasks of perceptual flexibility in speech, spanning a gamut of possible focuses, have reasonable split-half reliability, and some element of construct validity as well.

High reliability is good news. In broad strokes, tasks are said to be reliable when measures of reliability for them are higher than 0.7, with scores above 0.9 considered to be particularly meritorious (Hedge et al., 2018). We judged tasks to have excellent reliability if the reliability observed was greater than 0.7 and the 95% confidence interval did not include 0.5. With only two exceptions (out of 10 total estimates of split-half reliability studied), tasks of perceptual flexibility showed split-half reliabilities greater than 0.7, with two measures showing reliability greater than 0.9. Put another way, all but one measure—the degree of perceptual learning on the first session of the phonetic recalibration task (Experiment 1)—showed a stronger reliability than the reliability shown (0.64) in one paper for average RT costs on the Stroop task (Hedge et al., 2018). It is worth noting that the measures of the two tasks showing the weakest reliability involved difference scores, which were found to be generally less reliable than other scores by Hedge et al. (2018), although the confidence intervals for estimates from both tasks also included

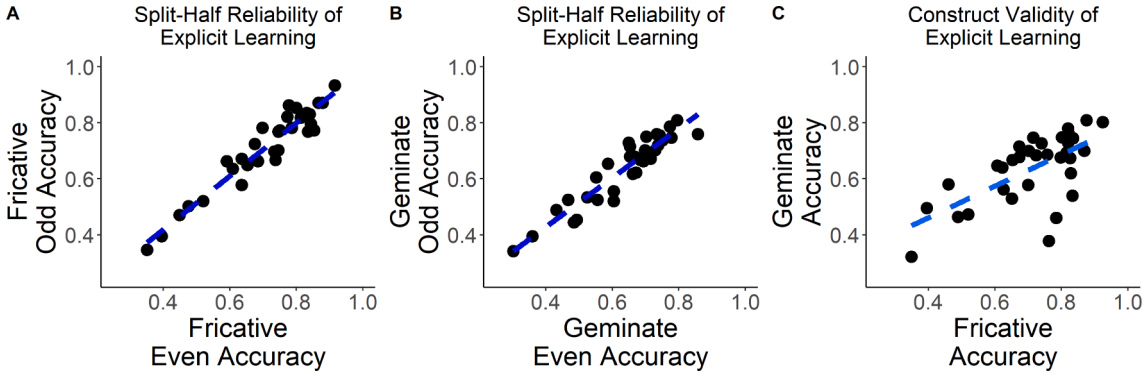


Fig. 5. Relationships in the accuracy for specific groups of items. (A) shows split-half reliability for the fricative stimuli, (B) shows split-half reliability for the geminate stimuli, and (C) shows construct validity using the fricative stimuli and geminate stimuli within the same participants. Each dot is an individual participant. The dashed blue lines indicate a linear best fit line.

Table 1
Summary of *r* values for reliability and construct validity. Each row indicates a different experiment, as well as its topic and the task used to assess it.

Exp.	Construct	Stimulus Set	Reliability	Excellent Evidence?	Validity
1	Perceptual learning	Session 1	0.436 (0.103–0.681)	X	0.079 (−0.278–0.416)
		Session 2	0.743 (0.532–0.867)	✓	
2	Accent learning	Italian	0.759 (0.573–0.871)	✓	0.531 (0.246–0.732)
		Belarusian	0.808 (0.652–0.898)	✓	
3	Rate learning	Male	0.856 (0.716–0.929)	✓	0.657 (0.388–0.822)
		Female	0.766 (0.560–0.882)	✓	
4	Incidental non-native learning	Fricatives	0.869 (0.750–0.934)	✓	0.263 (−0.088–0.556)
		Geminates	0.648 (0.391–0.811)	X	
5	Explicit non-native learning	Fricatives	0.957 (0.916–0.979)	✓	0.656 (0.408–0.814)
		Geminates	0.936 (0.875–0.968)	✓	

higher levels of reliability. This provides support for the idea that many tasks of phonetic plasticity show excellent levels of reliability.

When resources are limited, then, researchers may want to consider using tasks with excellent levels of reliability. The observed correlations are capped by the reliability each task showed, which in turn leads to different recommended sample sizes based on reliabilities (Parsons et al., 2019). Imagine running an experiment examining the correlation between performance on the explicit non-native learning task (for fricatives) and the rate learning task (for the male talker). The current paper suggests that the reliability of both tasks is high: 0.96 for the fricatives and 0.86 for the rate learning task. Parsons et al. (2019) suggest that a sample size of approximately 40 would be sufficient to obtain 80% power to detect a true correlation of 0.5 between the two tasks, given those levels of reliability. However, in a world in which the rate learning task is much less reliable (say, 0.4), the sample size needed to obtain that power level and true correlation would rise (approximately 100 for this example). The present paper includes many studies that were constructed as a precursor to assess a broader conceptualization of individual differences in perceptual flexibility (Heffner & Myers, 2021). In that paper, the authors ran participants in multiple tasks of perceptual flexibility, seeking the extent to which the tasks shared common variance; assessing the extent to which the observed correlations reflected true common variance was impossible without real measures of reliability and validity. Issues around sample size are of particular urgency when using cognitive neuroscientific designs, many of which are relatively underpowered (Button et al., 2013). Similarly, in tasks using clinical populations, which often depend on small sample sizes, it may be helpful to focus on tasks that generate consistent results within individuals, particularly for designs that rely on longitudinal observations of individual participants.

For validity, meanwhile, the results are more nuanced. Many correlations were greater than zero, often estimated at levels greater than 0.7. However, other correlations were small; for instance, the correlation between the two sessions for the perceptual learning task was just 0.08. Unlike for reliability, there is no standard accepted level at which measures of validity such as these should be considered "good" or "excellent", but it is worthwhile noting the variability present in the levels in of construct validity. As with any other correlation, the possible correlation coefficient for validity was contingent on the reliabilities for each individual component of the task, but in some cases the validity was much lower than the individual reliability components would have suggested. Tasks with a high split-half reliability but low validity may indicate that the nature of the stimuli used in that particular task may drive quite a bit of the variation between listeners in performance on that task. However, these estimates may truly reflect something broader about the underlying psychological processes that these tasks tap. Further tests would be necessary to explore which of these possibilities seems more likely.

These conclusions, of course, are limited by the populations and sample sizes used. Reliability, after all, is not just a function of the task used; it is also a function of the populations used. In these experiments, we used a variety of different populations, from two online sources (Mechanical Turk and Prolific) and in-person participants at UConn. In our final analyses here, we decided to combine across all three populations, reasoning that providing more potential variability would better allow for a sample of participants that spans the experience of younger adults. A lack of variability can act against estimates of reliability if the population observed is too homogeneous, as college participant pools often are (Henrich et al., 2010).

Yet this may also limit the ability to make inferences about the populations we drew from in this study, as reliability can differ between populations for the same task. Imagine sending beach volleyball teams from the Summer Olympic Games to the bonspiel for curling at the Winter Olympic Games. It is unlikely they will perform *consistently* at curling in the same way that the designated curlers would have. Similarly, in behavioral testing, it may be the case that, say, older adults would show different reliability for each of these tasks from younger adults. Thus, a weakness of our study is that we pool across these three populations to get our samples; in Supplementary Material, we include observed reliability and validity for the online participants on their own in Experiments 2, 3, 4, and 5. The patterns observed in the online participants alone are strikingly similar to the ones including UConn participants in the present manuscript (perhaps unsurprisingly, considering that UConn participants did not make up a large fraction of the participants here). The sample sizes used in the current paper were also relatively small; they were sufficient to assess whether any one task showed excellent reliability, but were insufficient for smaller reliabilities and/or for comparison of reliability or validity across tasks. As such, our conclusions are limited primarily to whether any one task showed a strong level of reliability rather than nuanced discussions of precise magnitudes of reliability or validity.

In sum, this project assessed the internal consistency of a variety of measures of perceptual flexibility in speech. The split-half reliability and construct validity of five tasks were assessed. In general, these tasks showed excellent split-half reliability. Correlations between stimulus sets indicated the possibility of construct validity, although correlations were weaker between stimulus sets than they were within stimulus sets. This shows that many tasks of perceptual flexibility show consistency within individuals, but the weaker construct validity seen here could have implications for subsequent studies investigating individual differences using similar tasks (for example, researchers may want to be cautious in their conclusions about generalizability).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank the members of the UConn Language and Brain Lab for their assistance in data collection and discussion of results. Portions of this research were presented at the annual meetings of the

Psychonomic Society in 2018 and 2019 and the Society for Neurobiology of Language annual conference in 2019. Funding for this research was provided by the National Science Foundation (grant SMA 1714858 awarded to Heffner, grant BCS 1554510 to Myers, grant DGE 1747486 to the University of Connecticut), and National Institutes of Health (R01 DC013064 to EBM). CCH was also supported by a Hunt Fellowship awarded by the Acoustical Society of America. SL was supported by a Graduate Research Fellowship from the National Science Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bandl.2021.105070.

References

Adank, P., & Devlin, J. T. (2010). On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. *NeuroImage, 49*(1), 1124–1132. https://doi.org/10.1016/j.neuroimage.2009.07.032

Adank, P., & Janse, E. (2009). Perceptual learning of time-compressed and natural fast speech. *Journal of the Acoustical Society of America, 126*(5), 2649–2659. https://doi.org/10.1121/1.3216914

Adank, P., & Janse, E. (2010). Comprehension of a novel accent by young and older listeners. *Psychology and Aging, 25*(3), 736–740. https://doi.org/10.1037/a0020054

Amitay, S., Irwin, A., Hawkey, D. J. C., Cowan, J. A., & Moore, D. R. (2006). A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners. *Journal of the Acoustical Society of America, 119*(3), 1616–1625. https://doi.org/10.1121/1.2164988

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*, 388–407. https://doi.org/10.1101/438242

Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015). Cognitive predictors of perceptual adaptation to accented speech. *Journal of the Acoustical Society of America, 137*(4), 2015–2024. https://doi.org/10.1121/1.4916265

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2016). *lme4: Linear mixed-effects models using Eigen and S4*. http://cran.r-project.org/package=lme4.

Bent, T., Baese-Berk, M., Borrie, S. A., & McKee, M. (2016). Individual differences in the perception of regional, nonnative, and disordered speech varieties. *The Journal of the Acoustical Society of America, 140*(5), 3775–3786. https://doi.org/10.1121/1.4966677

Boersma, P., & Weenink, D. (2001). Praat: Doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Bosker, H. R. (2017). How our own speech rate influences our perception of others. *Journal of Experimental Psychology: Learning Memory and Cognition, 43*(8), 1225–1238. https://doi.org/10.1037/xlm0000381

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews: Neuroscience, 14*, 365–376.

Chandrasekaran, B., Sampath, P. D., & Wong, P. C. M. (2010). Individual variability in cue-weighting and lexical tone learning. *Journal of the Acoustical Society of America, 128*(1), 456–465. https://doi.org/10.1121/1.3445785

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America, 116*(6), 3647–3658. https://doi.org/10.1121/1.1815131

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope, 58*(9), 955–991.

Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2012). Auditory skills and brain morphology predict individual differences in adaptation to degraded speech. *Neuropsychologia, 50*(9), 2154–2164. https://doi.org/10.1016/j.neuropsychologia.2012.05.013

Fawns-Ritchie, C., & Deary, I. J. (2020). Reliability and validity of the UK Biobank cognitive tests. *PLoS ONE, 15*(4), 1–24. https://doi.org/10.1371/journal.pone.0231627

Gabay, Y., Dick, F. K., Zevin, J. D., & Holt, L. L. (2015). Incidental auditory category learning. *Journal of Experimental Psychology: Human Perception and Performance, 41*(4), 1124–1138. https://doi.org/10.1037/xhp0000073

Golestani, N. (2014). Brain structural correlates of individual differences at low-to high-levels of the language processing hierarchy: A review of new approaches to imaging research. *International Journal of Bilingualism, 18*(1), 6–34. https://doi.org/10.1177/1367006912456585

Golestani, N., & Zatorre, R. J. (2009). Individual differences in the acquisition of second language phonology. *Brain and Language, 109*(2–3), 55–67. https://doi.org/10.1016/j.bandl.2008.01.005

Heaton, R. K., Akshoomoff, N., Tulsky, D., Mungas, D., Weintraub, S., Dikmen, S., Beaumont, J., Casaletto, K. B., Conway, K., Slotkin, J., & Gershon, R. (2014). Reliability and validity of composite scores from the NIH toolbox cognition battery in adults. *Journal of the International Neuropsychological Society, 20*(6), 588–598. https://doi.org/10.1017/S1355617714000241

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. https://doi.org/10.3758/s13428-017-0935-1

Heffner, C. C., Idsardi, W. J., & Newman, R. S. (2019). Constraints on learning disjunctive, unidimensional auditory and phonetic categories. *Attention, Perception, and Psychophysics, 81*(4), 958–980. https://doi.org/10.3758/s13414-019-01683-x

Heffner, C. C., & Myers, E. B. (2021). Individual differences in phonetic plasticity across native and non-native contexts. *Journal of Speech, Language, and Hearing Research, 64*(10), 3720–3733. https://doi.org/10.1044/2021_jslhr-21-00004

Heffner, C. C., Newman, R. S., & Idsardi, W. J. (2017). Support for context effects on segmentation and segments depends on the context. *Attention, Perception, and Psychophysics, 79*(3), 964–988. https://doi.org/10.3758/s13414-016-1274-5

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences, 33*(2–3), 111–135. https://doi.org/10.1017/S0140525X10000725

Janse, E., & Adank, P. (2012). Predicting foreign-accent adaptation in older adults. *Quarterly Journal of Experimental Psychology, 65*(8), 1563–1585. https://doi.org/10.1080/17470218.2012.658822

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance, 43*(9), 1594–1611. https://doi.org/10.1037/xhp0000410

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 3933–3936*. https://doi.org/10.1109/ICASSP.2008.4518514

Key, M. (2014). Positive expectation in the processing of allophones. *Journal of the Acoustical Society of America, 135*(6), EL350–EL356. https://doi.org/10.1121/1.4879669

Kim, D., Clayards, M., & Goad, H. (2017). Individual differences in second language speech perception across tasks and contrasts: The case of english vowel contrasts by Korean learners. *Linguistics Vanguard, 3*(1), 1–11. https://doi.org/10.1515/lingvan-2016-0025

Kim, D., Clayards, M., & Goad, H. (2018). A longitudinal study of individual differences in the acquisition of new vowel contrasts. *Journal of Phonetics, 67*, 1–20. https://doi.org/10.1016/j.wocn.2017.11.003

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience, 34*(1), 43–68. https://doi.org/10.1080/23273798.2018.1500698

Kollmeier, B., Gilkey, R. H., & Sieben, U. K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. *Journal of the Acoustical Society of America, 83*(5), 1852–1862. https://doi.org/10.1121/1.396521

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*, 1–15. https://doi.org/10.1016/j.jml.2006.07.010

Lazar, R. M., & Antoniello, D. (2008). Variability in recovery from aphasia. *Current Neurology and Neuroscience Reports, 8*(6), 497–502. https://doi.org/10.1007/s11910-008-0079-x

Linares, D., & López-Moliner, J. (2016). quickpsy: An R package to fit psychometric functions for multiple groups. *The R Journal, 8*(1), 122–131.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tokhura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America, 96*(4), 2076–2087. https://doi.org/10.1121/1.410149

Luthra, S., Fuhrmeister, P., Molfese, P. J., Guediche, S., Blumstein, S. E., & Myers, E. B. (2019). Brain-behavior relationships in incidental learning of non-native phonetic categories. *Brain and Language, 198*, Article 104692. https://doi.org/10.1016/j.bandl.2019.104692

Luthra, S., Mechtenberg, H., & Myers, E. B. (2021). Perceptual learning of multiple talkers requires additional exposure. *Attention, Perception, and Psychophysics, 83*(5), 2217–2228.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Moinester, M., & Gottfried, R. (2014). Sample size estimation for correlations with pre-specified confidence interval. *The Quantitative Methods for. Psychology, 10*(2), 124–130. https://doi.org/10.20982/tqmp.10.2.p0124

Nunnally, J. C. (1978). *Psychometric Theory* ((2nd ed.)). McGraw-Hill.

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. https://doi.org/10.1177/2515245919879695

Peelle, J. E., & Wingfield, A. (2005). Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech. *Journal of Experimental Psychology: Human Perception and Performance, 31*(6), 1315–1330. https://doi.org/10.1037/0096-1523.31.6.1315

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America, 130*(1), 461–472. https://doi.org/10.1121/1.3593366

R Core Team. (2020). *R: A language and environment for statistical computing*. https://www.r-project.org/.

Rotman, T., Lavie, L., & Banai, K. (2020). Rapid perceptual learning: A potential source of individual differences in speech perception under adverse conditions. *Trends in Hearing, 24*, 1–16. https://doi.org/10.1177/2331216520930541

Runnqvist, E., Strijkers, K., Sadat, J., & Costa, A. (2011). On the temporal and functional origin of L2 disadvantages in speech production: A critical review. *Frontiers in Psychology, 2*(DEC), 1–8. https://doi.org/10.3389/fpsyg.2011.00379

Saltzman, D. I., & Myers, E. B. (2021). Listeners are initially flexible in updating phonetic beliefs over time. *Psychonomic Bulletin & Review, 28*, 1354–1364.

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics, 52*, 183–204. https://doi.org/10.1016/j.wocn.2015.07.003

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schwab, S. (2011). Relationship between speech rate perceived and produced by the listener. *Phonetica, 68*(4), 243–255. https://doi.org/10.1159/000335578

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. S., Højsgaard, S., Fox, J., Lawrence, M. A., Mertens, U., Love, J., Lenth, R., & Christensen, R. H. B. (2021). *afex: Analysis of Factorial Experiments*. https://cran.r-project.org/web/packages/afex/index.html.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*(1), 72–101.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 7*(5), 1074–1095. https://doi.org/10.1037/0096-1523.7.5.1074

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*(3), 1003–1014. https://doi.org/10.3758/s13423-020-01840-6

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics, 79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Xie, X., Earle, F. S., & Myers, E. B. (2018). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience, 33*(2), 196–210. https://doi.org/10.1080/23273798.2017.1369551