Lexical entrainment reflects a stable individual trait

Lexical entrainment reflects a stable individual trait: Implications for individual differences in language processing

Anita Tobar-Henríquez, Hugh Rabagliati, Holly P. Branigan

The University of Edinburgh

Abstract

Language use is intrinsically variable, such that the words we use vary widely across speakers and communicative situations. For instance, we can call the same entity *refrigerator* or *fridge*. However, attempts to understand individual differences in how we process language have made surprisingly little progress, perhaps because most psycholinguistic instruments are better-suited to experimental comparisons than differential analyses. In particular, investigations of individual differences require instruments that have high test-retest reliability, such that they consistently distinguish between individuals across measurement sessions. Here, we established the reliability of an instrument measuring lexical entrainment, or the tendency to use a name that a partner has used before (e.g., using *refrigerator* after a partner used *refrigerator*), which is a key phenomenon for the psycholinguistics of dialogue. Online participants completed two sessions of a picture matching-and-naming task, using different pictures and different (scripted) partners in each session. Entrainment was measured as the proportion of trials on which participants followed their partner in using a low-frequency name, and we assessed reliability by comparing entrainment scores across sessions. The estimated reliability was substantial, both when sessions were separated by minutes and when sessions were a week apart. These results suggest that our instrument is well-suited for differential analyses, opening new avenues for understanding language variability.

Test-retest reliability, individual differences, lexical entrainment, language production

Lexical entrainment reflects a stable individual trait: Implications for individual differences in language processing

Language use is intrinsically variable: People express themselves differently across different social situations (Eckert & Rickford, 2001; Gregory & Carroll, 2018), and across their lifespans (e.g., Juncos-Rabadán, Facal, Rodríguez, & Pereiro, 2010; March, Wales, & Pattison, 2011). Language use also shows wide variation from individual to individual, depending on demographic characteristics such as gender (e.g., Leaper & Ayres, 2007). This variability shows up even in references to commonplace objects, where speakers might variously refer to the same object as a *fridge*, *refrigerator* or even just *thing*. A range of theories have drawn on psycholinguistic experiments using group-level approaches to elucidate the cognitive and social factors that might inform and constrain how we make lexical choices, and in doing so they have also cast light on factors that can affect how we use language more generally. In this paper, we take a novel approach to investigating variability in language use, seeking to establish whether the way in which we make lexical choices reflects stable individual differences in how we process language. We focus on one particular determinant of lexical choices: **lexical entrainment**[i]**,** or speakers' overt use of names that their conversational partner has used before (e.g., to use *refrigerator* after hearing a partner use *refrigerator*; Garrod & Anderson, 1987; Brennan & Clark, 1996; Branigan, Pickering, Pearson, McLean, & Brown, 2011).

Research on lexical entrainment has revealed that lexical choices are affected by cognitive factors, such as memory, as well as social factors, such as the identity of a partner. However, there are still competing theories as to what underlies the phenomenon, and as to how cognitive and social factors interact in language production. For instance, Pickering & Garrod (2004) suggested that we reuse our partner's words mainly because we have recently processed them; recent processing makes the words more accessible in memory, and

therefore facilitates retrieval and reuse (e.g. Neely, 1976; Meyer, 1996). By contrast, audience design accounts suggest that our tendency to reuse our partner's lexical choice occurs because we tend to adopt their perspective during conversation in order to facilitate mutual comprehension (Clark, 1996). Evidence for this account comes from demonstrations that we reuse lexical choices in a partner-specific fashion (Brennan & Clark, 1996; Horton & Gerrig, 2002, 2005). Other theories have suggested that reusing our partner's lexical choices can also be considered a pro-social behavior, as it makes us more likeable to our interlocutors (Van Baaren, Holland, Steenaert, van Knippenberg, 2003); consistent with this, sociolinguistic studies have shown that social hierarchies and community membership correlate with our tendency to reuse a partner's choices. For instance, members of low status groups are more likely to entrain than members from high status groups (Palomares, Giles, Soliz, & Gallois, 2016). Similarly, people are more likely to entrain to more prototypical community members than to non-prototypical community members (Gallois & Callan, 1991).

In this paper, we make the case that studying individual differences in lexical entrainment can be informative not only about this specific phenomenon, but also about theories of language use more generally. For instance, given that lexical entrainment could be explained as a consequence of audience design, an individual differences' approach could be used to test the hypothesis that an individual's perspective-taking skills might predict the degree to which they engage in lexical entrainment (cf. Hopkins, Yuill, & Branigan, 2017), which would in turn cast light on the effects of perspective-taking in language processing more generally. Similarly, individual variability in lexical entrainment can also cast light on how situational factors might interact with individual traits factors, such as personality, in language use. For example, are individuals who are more prone to exhibit pro-social behavior (agreeableness) more likely to lexically entrain to a non-prototypical community member than individuals who are less agreeable? Finally, individual differences in how people entrain

at the lexical level can inform theories of our general tendency to reuse a partner's linguistic choice at other levels of structure (e.g., phonetics, syntax, etc.; Pickering & Branigan, 1998; Pickering & Garrod, 2004). If entrainment at different levels of linguistic structure is supported by a domain-general mechanism for imitation, then our tendency to reuse a partner's lexical choice should correlate with, for example, our tendency to reuse a partner's syntactic choice (e.g., passive versus active structures; see also Horton, 2014, for related research on individual differences in syntactic entrainment).

However, a pre-requisite for studying individual differences in phenomena such as lexical entrainment is possessing instruments and protocols that are calibrated to allow us to reliably measure these behaviors at the individual level. One important aspect of this is ensuring that instruments have high test-retest reliability, such that they can consistently distinguish between individuals across two measurement sessions (Shrout & Fleiss, 1979; Polit, 2015; Berchtold, 2016). Test-retest reliability is usually quantified by measuring the correlation coefficient between two sets of measurements from the same group of individuals; classically, the reliability of a test is said to be excellent if that coefficient is above .8, substantial if between .8 and .6, moderate if between .6 and .4, and poor if below .4 (Cicchetti & Sparrow, 1981; Landis & Koch, 1977). Understanding the test-retest reliability of an instrument is critical because it is a key determinant of the statistical power of a study. If studies are conducted using instruments that have a low test-retest reliability, then their ability to detect relationships with other constructs will be compromised by their inability to consistently distinguish between individuals on the dimension being measured. For instance, in order to investigate the relationship between agreeableness and lexical entrainment, we would need instruments that consistently distinguish between individuals in terms of both their degree of agreeableness and their propensity towards lexical entrainment.

Reliable instruments exist for testing certain individual differences, such as personality traits (e.g., Big Five; John & Srivastava, 1999). But recent work has suggested that many of the most well-known paradigms for assessing cognitive processing actually have poor test-retest reliability. Hedge, Powell, & Sumner (2018) demonstrated that a range of classic tasks, such as the Flanker and Stroop tasks, which reliably elicit effects at the group level, do not reliably measure individual variation, producing test-retest reliability scores that often fail to reach even a moderate level. The reason for these low scores, Hedge et al. argue, is that these instruments produce a distinctive restricted range of responses, which minimise variability between respondents (e.g., almost all participants show a Stroop cost, and this cost is similarly-sized across participants). Although this feature is highly desirable for experimental research, it is problematic for correlational studies, which need to elicit a large enough range of scores to capture individual variation. The fact that reliable experimental tasks elicit minimally different effect sizes between individuals compromises the instruments' ability to distinguish between these individuals, and thus leads to low test-retest reliability.

This claim has important potential consequences for the study of language processing, because it seems quite likely that many experimental tasks that have been used for studying individual variation in this field are actually ill-suited for that purpose (Kidd, Donnelly, & Christiansen, 2018). For example, a number of studies have assessed individual variation in statistical learning, i.e., the ability to learn statistical co-occurrences of features in our environment (Siegelman, Bogaerts, & Frost, 2017). However, most of the tasks that have been used to assess individual differences in statistical learning were actually designed for studying group-level comparisons, and so it is not clear that they reliably measure individual variation. Indeed, there is some evidence that test-retest reliability in these tasks is compromised (Arnon, 2019; Pardo, Urmanche, Wilman, & Wiener, 2017; Siegelman & Frost, 2015). This point is important because, in principle, it casts doubt on whether prior

findings are likely to replicate, as correlations drawn from instruments with low test-retest reliability are likely to be either false positives or negatives. Thus, it implies that we should be wary about using those studies to draw theoretical conclusions.

These considerations highlight the importance of establishing the test-retest reliability of an instrument before using it to study individual differences. Here we aim to establish the test-retest reliability of an instrument for measuring the phenomenon of lexical entrainment, both as a necessary step in the development of sound correlational studies of the phenomenon, and also as a way to evaluate previous work that has examined individual variation in entrainment. For example, Hopkins et al. (2017) studied lexical entrainment in a sample of typically developing and autistic children, and found that individuals' tendency to entrain did not correlate with measures of theory of mind or inhibitory control. However, without knowing the test-retest reliability of the lexical entrainment instrument, it is hard to interpret these null results. If the instrument has low reliability, then we should not typically expect to find reliable correlations between measures of entrainment and inhibitory control, even if the underlying factors are indeed associated. By contrast, if test-retest reliability is high, then these null findings are more likely to be indicative of true null associations. Thus, understanding the test-retest reliability of lexical entrainment instruments presents an important goal.

Branigan et al. developed a lexical entrainment instrument, which we will adapt in this paper, that has been repeatedly shown to elicit reliable effects at group-level, experimental comparisons (e.g., Branigan et al., 2011, 2016; Hopkins et al., 2017). In this instrument, participants collaborate with a confederate to match and name pictures. The experimental targets are pictures of objects that can be named with both a disfavoured and a favoured name (e.g., *brolly* versus *umbrella*, in British English); these materials have been pre-tested to ensure that participants rarely use the disfavoured name spontaneously, but still

consider it an acceptable name for the object. In the main matching-and-naming instrument, participants always name the experimental targets after the confederate. Lexical entrainment is then measured as the proportion of trials on which participants use the same name used by the confederate. Importantly, Branigan and colleagues have consistently shown that individuals are more likely to use a disfavoured name (e.g., *brolly*) after the partner has used the disfavoured name (e.g. *brolly*) than after the partner has used the favoured name (e.g., *umbrella*) or compared to its baseline frequency of use (Branigan et al., 2011, 2016; Hopkins et al., 2017), and so they have demonstrated that this instrument elicits experimentally reliable entrainment effects for disfavoured names. Moreover, they showed that speakers' propensity to entrain to a partner's use of a disfavoured name was not affected by modality: Individuals were equally likely to use a partner's disfavoured name during a written computerised interactive picture-naming task as during a spoken computerised version of the same task, providing evidence that effects elicited by this instrument generalise to speech (Branigan et al., 2011). But is the instrument also reliable for correlational studies?

Interestingly, this lexical entrainment instrument contains features that could enhance test-retest reliability. Its most critical feature is that it is designed to measure a general tendency to lexically entrain rather than to measure entrainment to a specific lexical item. To wit, each trial offers participants the opportunity to entrain to a different lexical item, which - we assume - means that behavior on that trial is relatively independent of behavior on previous trials (e.g., entraining to call a fridge a *refrigerator* should not in principle influence whether you call an umbrella a *brolly*). By contrast, typical paradigms used for measuring other types of entrainment contain design elements that might reduce test-retest reliability, because they are typically designed to measure entrainment to one feature only. For example, paradigms measuring the reuse of a partner's syntactic choice (syntactic entrainment) tend to assess participants' tendency to entrain to a specific syntactic structure (e.g., passive

structures) and therefore necessitate that participants process the same linguistic structure repeatedly (e.g., Kaschak et al., 2011; Branigan & Messenger, 2016).

Measurements of entrainment to only one linguistic structure are likely to be quite strongly affected by participants' idiosyncratic experience with, or preference for, that specific structure, and thus may not be indicative of a general structure-independent tendency to syntactically entrain. Moreover, if a task repeatedly tests entrainment to a single structure, this would likely increase measurement error and, in principle, could lead to participants showing maximally large effects in the manner described by Hedge et al. (2017), which would leave little room for measuring individual differences. In principle, this issue could be surmounted if syntactic entrainment instruments used different syntactic structures in each critical trial, but such an approach would pose significant practical challenges. By contrast, it is simple to use different lexical items in each trial, such that lexical entrainment instruments measure entrainment anew in each trial, and thus do not fall prey to the criticisms of Hedge and colleagues.

The present studies

To investigate the stability of a lexical entrainment measure, we conducted two internet-based studies in which native speakers of British English engaged in two sessions of an interactive online picture matching-and-naming task. Our task was based on the task used in Branigan et al. (2011), and our materials were normed with a new, representative internet-based sample. In the main task, participants alternated turns with what they believed to be an online partner to either match or name a picture (in reality the 'partner' was always pre-programmed software). Given previous evidence that this task reliably elicits entrainment effects for disfavoured labels when participants have experienced the partner previously using a disfavoured label (Branigan et al., 2011, 2016; Hopkins et al., 2017), we measured entrainment to the use of disfavoured names only: Experimental trials comprised a target that,

in British English, could be named with both a highly favoured name, e.g. *umbrella*, and a disfavoured, but acceptable, name, e.g., *brolly*, and the partner always used the disfavoured name to refer to the targets. Importantly, participants always matched experimental targets (i.e., responded to their partners naming the targets) before themselves naming the targets on a subsequent trial, and we measured entrainment as the proportion of trials on which the participant used the same disfavoured name as they had previously experienced the partner using.

In each study, we sought to establish first whether participants lexically entrained with an unseen partner in an online interactive picture naming-and-matching task, and second whether their propensity to lexically entrain was consistent across time. In our first study, we measured the test-retest reliability of lexical entrainment over a short time period: Participants completed two sessions immediately consecutively. Importantly, in each session, entrainment was measured with different items (e.g., *refrigerator/fridge* would be tested only in Session 1, while *brolly/umbrella* would be tested only in Session 2), meaning that the test-retest reliability should reflect an individuals' general tendency to lexically entrain, rather than their tendency to use particular (low-frequency) terms. In the second study, we measured reliability over a longer time period, with sessions separated by 7-to-8 days.

In both studies, we assessed whether there was a group effect of lexical entrainment by comparing whether the disfavoured name was used more often in the main task than in a spontaneous picture naming-and-matching task that had been used to norm the materials (and that did not offer opportunities for entrainment). We then measured the test-retest reliability of our lexical entrainment measure in two ways. First, we calculated the relative rankings correlation between participants' use of disfavoured names in the first session and in the second session, assessing whether participants' degree of entrainment was ranked the same across sessions. Second, we measured the absolute consistency between participants'

tendency to use disfavoured names in each session, in other words whether the instrument elicited exactly the same result for each participant in each session. In our design, we aimed to minimize situation-specific effects on lexical entrainment by using different stimuli across the two sessions, and by telling participants that they would be playing against different 'partners' in each of the two testing sessions, to avoid any possible partner-specific influences on lexical entrainment. Importantly, since individuals can have encountered disfavoured names in different proportions in previous experience, we aimed to minimize possible effects of past experience of the disfavoured names by using a range of 28 items, so that individual differences in previous experience with particular names could not explain participants' overall tendency to use disfavoured names during this task.

Materials' creation: Norming tasks

We conducted two norming tasks to create our experimental items, which comprised a target picture of an object that could be named with a favoured name in British English (e.g., *umbrella*) and a disfavoured, but acceptable name (e.g., *brolly*). Ethical approval for this norming procedure was obtained from the Psychology Research Ethics Committee of the University of Edinburgh (72-1617/9). In order to create the pairs of favoured and disfavoured names for each experimental target, we conducted an initial pre-test with a different set of participants, drawn from the same population as those in the main studies. 60 native speakers of British English (aged 18-60, M=36, SD=11) answered two questions in an online survey (via Prolific). For each of 120 pictured objects, participants provided a favoured name for the picture (i.e., spontaneous naming, *What is the first word you would use to name this object?*), followed by a less-favoured name (i.e., forced naming, *What other word could you use to name this object?*).

From these ratings, we gathered 50 potential target pictures, for which at least 70% of participants had provided the same favoured name, and at least 15% of participants had

provided the same disfavoured name. Importantly, the disfavoured names did not consistently come from specific registers or dialects of British English. The 50 potential targets were then entered into a second rating task, in which 60 new native speakers of British English (aged 18-60, M=38, SD=10) rated the acceptability of these disfavoured names with respect to the pictures on a scale from 1 to 7, where 1 corresponded to 'Not acceptable at all' and 7 corresponded to 'Highly acceptable'. We used this to create the final set of 28 disfavoured names, each of which had an acceptability rating above 5.3 (M=6.1, SD=.5), and had been used with a frequency below 30% (M=7%, SD=7%).

We then split these items into two sets of 14 (see Tables 1 and 2) that were matched in acceptability (Set 1: M=6.2, SD=.5; Set 2: M=6.1, SD=.4) and frequency of use during spontaneous naming (Set 1: M=7.1%, SD=7.2%; Set 2: M=7.6%, SD=7.1%). Across participants, we counterbalanced which set was presented in the first session, and which in the second session. We also used the first rating task to choose 14 filler pictures, in which at least 80% of participants agreed on the same favoured name.

Table 1. Item Set 1.

Disfavoured name (and favoured name)	Spontaneous naming (%)	Forced naming (%)	Acceptability score (1-7)
pillow (cushion)	12	72	5.3
musical instrument (accordion)	0	18	5.5
picture (painting)	22	43	5.6
make-up (lipstick)	0	33	5.6
silverware (cutlery)	3	10	5.7
flower (rose)	1	91	6.0
rodent (mouse)	0	65	6.3
loo (toilet)	5	63	6.4
mobile (phone)	13	48	6.4
refrigerator (fridge)	2	48	6.5
toad (frog)	8	60	6.7
aeroplane (plane)	20	45	6.8

memory stick (usb)	11	29	6.8
bicycle (bike)	10	67	6.8

Table 2. Item Set 2.

Disfavoured name (and favoured name)	Spontaneous naming (%)	Forced naming (%)	Acceptability score (1-7)
biro (pen)	3	34	5.3
computer (laptop)	8	83	5.6
rowboat (boat)	3	23	5.7
fag (cigarette)	10	53	6.0
spectacles (glasses)	4	45	6.0
coach (bus)	0	30	6.0
nectarine (peach)	5	15	6.1
hat (cup)	28	52	6.1
hen (chicken)	5	43	6.2
brolly (umbrella)	12	45	6.3
bunny (rabbit)	15	63	6.4
pistol (gun)	0	48	6.4
inflatable ball (ball)	1	26	6.5
bathtub (tub)	3	25	6.7

Study 1: Short-term reliability

Study 1 investigated whether individual levels of lexical entrainment could be reliably measured in two sessions a few minutes apart. Participants completed a picture matching-and-naming-task. On each trial participants were shown two images and, while alternating turns with an alleged partner, they either named or selected one of the pictures. On critical trials, we measured whether participants reused a disfavoured name that their partner had used earlier in the study.

Method

Ethical approval for this study was obtained from the Psychology Research Ethics Committee of the University of Edinburgh (72-1617/9).

Participants. We recruited 60 participants online using the portal Prolific [https://prolific.ac/]. To be included, participants had to be native speakers of British English, born and raised in the United Kingdom, and aged 18-60 (M=36, SD=12). Participants were paid £2.

Procedure. Participants completed two sessions of a matching-and-naming-task, each of which contained 28 matching trials and 28 naming trials. On each trial, participants were shown two pictures (Figure 1), and they then either clicked on the target picture named by their partner (matching trials) or typed the name of the indicated target picture (naming trials). Half of the trials were filler trials, on which the target picture only had a single name (e.g., *onion*). The other half were experimental trials, on which the target picture could either be named with a highly-favoured name (e.g., *umbrella*) or a less-favoured but still acceptable name (e.g., *brolly*). Thus, each session used 14 experimental items and 14 filler items, meaning that participants completed 28 experimental items and 28 filler items in total.

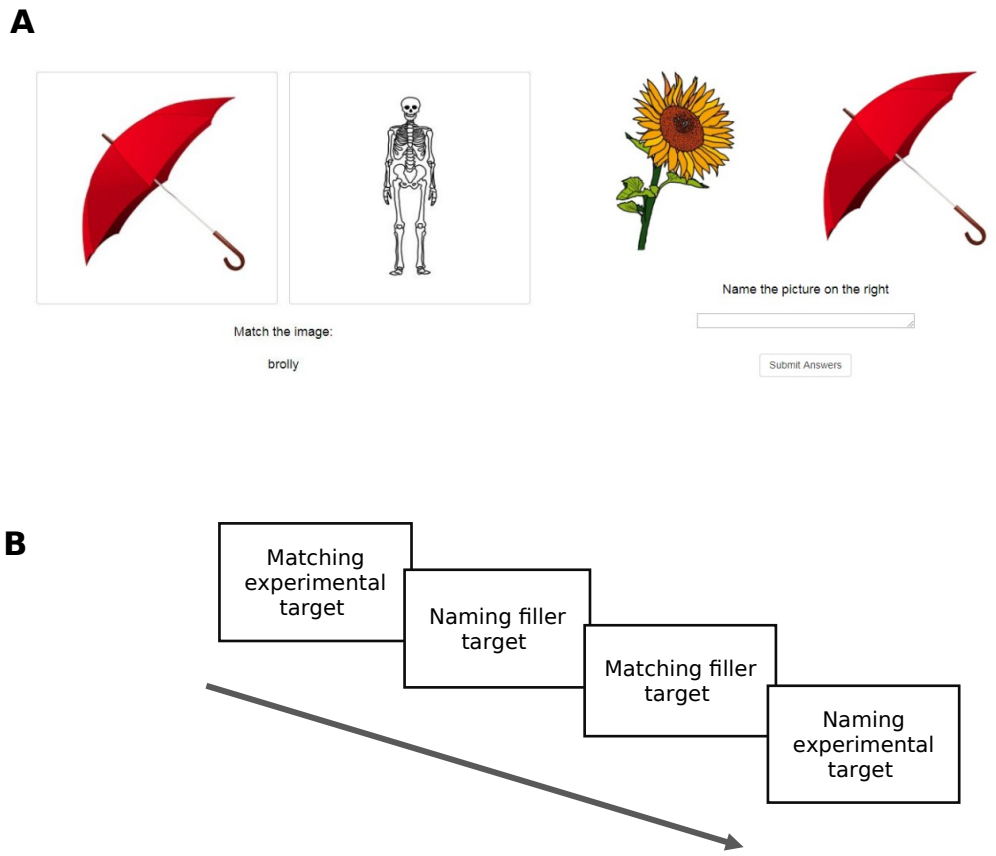

Figure 1. A. Examples of the matching (left) and naming (right) tasks (where the favoured word is *umbrella* and disfavoured is *brolly*). In matching trials, the participant selected the named target picture. In naming trials, they named the target. Targets were presented along with randomly selected distractors. **B. S**equence of experimental item and filler presentation. Participants first matched an experimental target with the corresponding disfavoured name, they subsequently named a filler, matched a filler, and finally named the previously matched experimental target.

The structure of the matching and naming task is illustrated in Figure 1. Participants alternated matching and naming trials with a 'remote player', who was in fact pre-programmed software that provided scripted answers. The trial order was fixed and the latency between matching experimental target and naming experimental target was always 3 trials, as in Figure 1b. Importantly, the trial structure meant that the software 'partner' always

named the experimental targets before the participants, using the disfavoured names exclusively (see Figure 1).

Participants were recruited to take part in this study on Prolific, using an advertisement that was visible only to individuals who met our inclusion criteria (see above). The advertisement stated that participants would play two sessions of a picture matching-and-naming task, and that they would play with a different remote player in each session. Prolific users interested in participating in the study were redirected from Prolific to a Qualtrics survey. After filling in an online consent form, they were told to wait to be matched with a remote player and, after two minutes, they were redirected to the first task (programmed with JSPsych and available at https://github.com/anitatobar/Test-retest-reliability-of-lexical-entrainment-task; de Leeuw, 2015), where they were asked to alternate turns with their partner to match and name one out of two pictures that would appear on the screen.

On each trial, participants saw two pictures and were asked to either wait for their partner's response so that they could select the correct (matching) picture, or to name the picture on the right/left (depending on where the target appeared, which was randomized) (see Figure 1). After matching and naming the 14 experimental items, they were told to wait to be matched to a new remote player. After two minutes, they were told the new partner was waiting for them and were asked to press a key to start the task. During the second session, participants matched and named 14 new experimental items. At the end of the task, participants were redirected to a second Qualtrics survey, where we checked participants' beliefs about the nature of their partner by asking *How many people did you play with during the two naming tasks?*; we coded whether participants reported playing with multiple partners, or explicitly indicated that they suspected they had played with a computer. Finally, participants were redirected to a Prolific website and received a completion code in order for us to confirm their payment.

Results

Data processing and exclusions. We coded all naming trials for whether they showed lexical entrainment (repeating the disfavoured name used by the partner) or not (using any other name). Occasionally, participants named or selected the distractor instead of the target; these trials were coded as NA. We excluded five participants because they reported believing that they had not played with a real person.

Analyses. We conducted two analyses, using the open source R language and environment (R Core Team, 2018) in RStudio (RStudio Team, 2015). All analyses and data can be found at https://github.com/anitatobar/Test-retest-reliability-of-lexical-entrainment-task.

We began by testing for the presence of a lexical entrainment effect, by comparing the percentage of disfavoured names used in our matching-and-naming tasks to the percentage of disfavoured names used in our first norming task. To do so, we used a paired-samples Wilcoxon test over the percentage of use of disfavoured names in each task.

Next, we assessed the test-retest reliability of the task by comparing the proportion of trials on which participants entrained in each session. In all analyses, the variables of interest were logit transformed proportions: We used this transformation over binary proportions to approach normality. In our first analysis, we used a Pearson's correlation to assess the degree to which the instrument could replicate the same ordering between respondents in the two sets of measurements. In our second analysis, we used intra-class correlation coefficients (ICC) to assess whether the instrument could also elicit the same exact result for each individual on each session. The ICC reflects the consistency between two or more raters (in this case, measurement sessions) for the same set of participants (Shrout & Fleiss, 1979; Polit, 2015; Berchotold, 2016), and its values fall between 0 and 1, with an ICC of 1 reflecting perfect consistency.

We calculated ICC values adopting two different approaches. First, we used an ANOVA-based approach estimating components of variance (McGraw & Wong, 1996; Shrout & Fleiss, 1979). Following Koo & Li (2016), we used a single-rating, absolute-agreement, two-way random effects model with two raters (testing sessions) across 55 individuals. In the commonly cited Shrout and Fleiss (1979; see also McGraw &Wong, 1996) nomenclature, this corresponds to ICC (2,1), which is sensitive to differences between session means. We used a two-way model because the sets of experimental items were counterbalanced across testing sessions, making participants' scores from Session 1 and Session 2 interchangeable. Moreover, we used a single-rating ICC type to compare each participant's tendency to entrain in Session 1 against their tendency to entrain in Session 2, rather than comparing Session 1 and Session 2 scores as a whole. In addition, we used an absolute agreement ICC definition (instead of a consistency definition) because we were not only interested in measuring ranking consistency across time but, most of all, we wanted to test our measure's ability to provide identical results in each measurement session.

Second, we calculated the ICC using a generalised mixed-effects models approach, which allows the calculation of standard error via bootstrapping (Nakagawa & Schielzeth, 2010). In particular, we built a two-way random effects model using logit transformed proportions as the independent variable. The model included random intercepts for both participants and sessions, and we used 1000 bootstrapping iterations to calculate ICC values and 95% confidence intervals.

In interpreting our reliability results, we adhere to conventional standards for judging test-retest reliability in correlational research: excellent or clinically required (.8), good/substantial (.6), and moderate (.4) (Cicchetti & Sparrow, 1981; Landis & Koch, 1977; see Koo & Li (2016) for discussion). However, it is important to note that setting explicit standards for judging these values is difficult, since the appropriateness of a coefficient

depends on factors such as the purpose for measuring the reliability of the instrument, the time interval between measurements, the types of sample being used, and whether the underlying phenomenon is believed to be volatile (Crocker & Algina, 1986).

Lexical entrainment effect. We found strong evidence for lexical entrainment. On average, participants used the disfavoured names on 36% of critical naming trials (SD=28%) across the two sessions. The percentages of use of disfavoured names during the matching-and-naming task were significantly higher than the percentages of use of these names during the spontaneous naming task used to norm the materials (M=7%, SD=7%, V=1, p<.0001), suggesting the presence of an entrainment effect (see Figure 2).

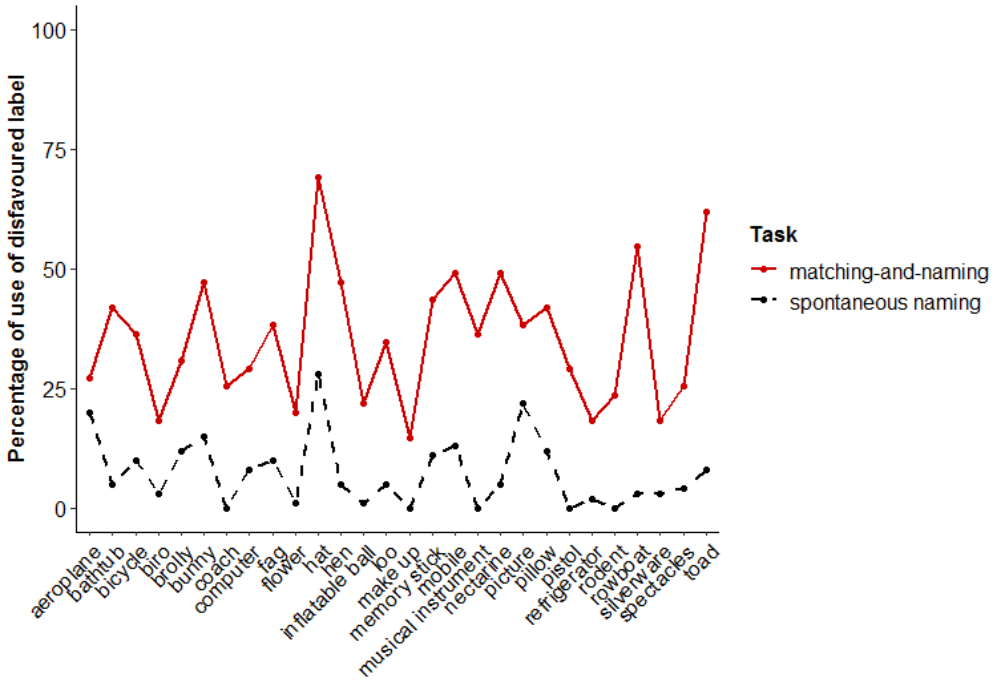

Figure 2. Percentage of use (y-axis) of disfavoured names (x-axis) in Study 1. The black (dashed) line represents the percentage of use of the disfavoured name in the spontaneous naming-task, while the red (solid) line represents the percentage of use of the disfavoured name in the primed matching- and naming-task used to measure lexical entrainment.

Test-retest reliability. Participants used disfavoured names on 34% of critical trials (SD=28%) in Session 1, and on 37% of critical trials (SD=28%) in Session 2. Importantly, our data shows a fairly substantial range of inter-individual variation in the degree of lexical entrainment (see Figure 3), which is an important prerequisite for correlational research. Moreover, we found a significant positive correlation between individuals' rates of lexical entrainment in Session 1 and lexical entrainment in Session 2 (r=.73, p<.0001; 95% CI [.57, .83]). The ANOVA-based approach indicated an ICC value of .73 (p<.0001) with a 95% confidence interval between .57 and .82; this was confirmed by the generalised mixed effects models approach, which showed an ICC of .73 (p<.0001) with a 95% confidence interval between .57 and .83, and a standard error of .065. Importantly, a Levene Test revealed that our model did not show violation of the assumption of homoscedasticity across Session 1 and Session 2 (F=0.07, p=0.8), suggesting that inter-individual variation was similar across measurement sessions. Taken together, these results suggest that the two sets of measurements were not only correlated but also highly consistent at the individual level.

Thus, Study 1 shows that lexical entrainment is an effect that can be reliably elicited at the group level, across a range of items, even in this novel on-line task in which participants believed they were interacting with a remote partner. More importantly, these results also suggest that the test-retest reliability of this lexical entrainment task is quite substantial over a short time window, implying that lexical entrainment shows short-term stability within individuals, and that this task is well-suited for studying individual variation in language processing. Next, we assess the test-retest reliability of our task over a considerably longer time window.

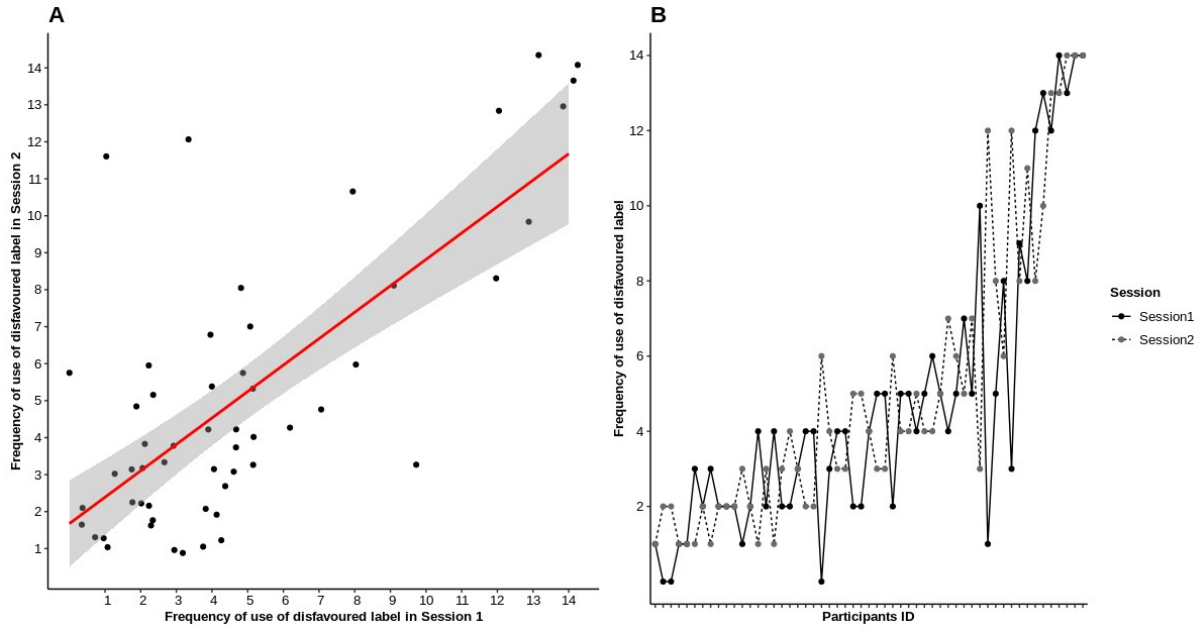

Figure 3. A. Positive correlation (*r*=.72) between the frequency of use of disfavoured names in Session 1 (x-axis) and Session 2 (y-axis). Points are jittered. The red line represents a linear regression between participants' scores in each session, while the grey shadow corresponds to a non-parametric regression smooth. **B**. Individual participants' scores in Session 1 (black, solid line) and Session 2 (grey, dashed line).

Study 2: Over-a-week reliability

Study 1 demonstrated substantial test-retest reliability of our lexical entrainment instrument when the testing sessions occurred with a gap of two minutes between sessions. In Study 2, we investigated whether test-retest reliability remains high when a seven-to-eight day gap is introduced between sessions.

Method

Except where detailed, Study 2 used the same methods as Study 1. Ethical approval for this study was obtained from the Psychology Research Ethics Committee of the University of Edinburgh (72-1617/9).

Participants. Study 2 used 60 further participants, aged 18-60 (M=31, SD=8), who were recruited using the same inclusion criteria as in Study 1.

Materials and procedure. Participants were recruited on Prolific, using an ad only visible to individuals who met our inclusion criteria (see above). The ad stated that participants would play two sessions of a picture matching-and-naming task, making explicit that the first session would take place immediately and the second session would take place in a week's time. The ad also stated that they would play with a different remote player in each session. Prolific users interested in taking part in Session 1 completed the same procedure as Study 1, except that they received a completion code to be paid for their participation in Session 1 immediately. A week later, they were served a new ad, allowing them to participate in Session 2, for which they followed the same procedure as Study 1. The second task was available to be answered for 48 hours. Those participants who wanted to participate in Session 2 were redirected to the second task and followed the same procedure as Study 1. Importantly, participants completed the manipulation check only once, at the end of Session 2, in order not to draw attention towards the nature of the remote partner before Session 2 was completed.

Results

Data processing and exclusions. Coding and exclusions were performed as in Study 1. We excluded ten participants because they did not complete the second session. Another five participants were excluded because they reported believing that they had not played with a real person.

Lexical entrainment effect. Again, we found strong evidence for an entrainment effect. On average, participants used the disfavoured names on 41% of trials (SD=24%) across the two sessions. The percentage of use of disfavoured names during the matching-and-naming task was significantly higher compared to the spontaneous naming task (M=7%, SD=7%, V=1, $p<.0001$), suggesting the presence of an entrainment effect (see Figure 4).

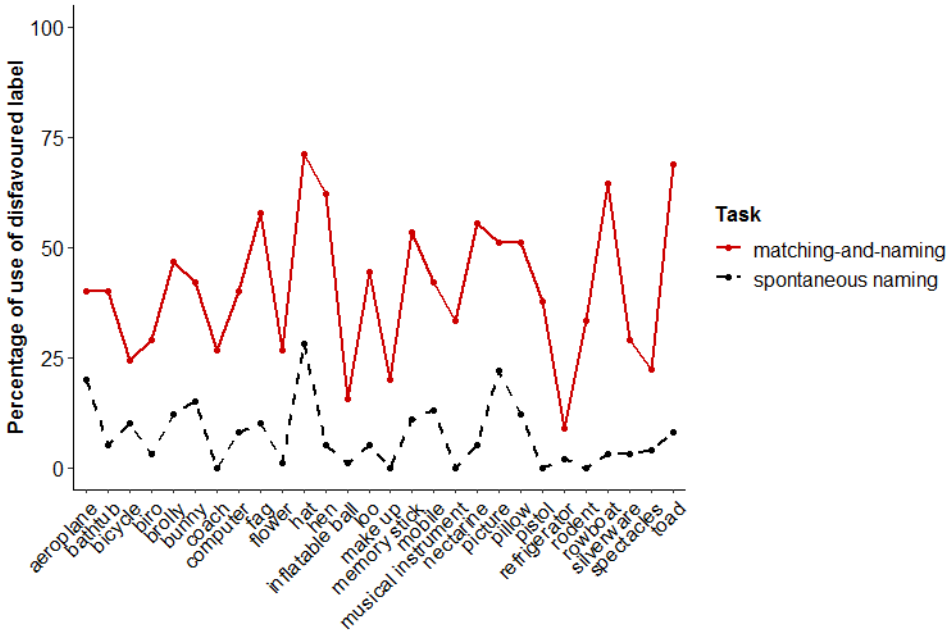

Figure 4. Percentage of use (y-axis) of disfavoured names (x-axis) in Study 2. The black (dashed) line represents the percentage of use of the disfavoured name in the spontaneous naming-task, while the red (solid) line represents the percentage of use of the disfavoured name in the primed matching- and naming-task used to measure lexical entrainment.

Test-retest reliability. Participants used disfavoured names on 40% of naming trials (SD=26%) in Session 1 and on 42% of naming trials (SD=26%) in Session 2. As in Study 1, our data shows a fairly wide range of inter-individual variation in the degree of lexical entrainment (see Figure 5). Moreover, we found a significant positive correlation between individuals' rates of lexical entrainment in Session 1 and lexical entrainment in Session 2 r=.61, p<.0001; 95% CI [.38, .77]). The ANOVA-based approach indicated an ICC value of .61 (p<.0001) with a 95% confidence interval between .4 and .77, and this was confirmed by the generalised mixed effects models approach, which also showed an ICC of .61 (p<.0001) with a 95% confidence interval between .41 and .76, and a standard error of .09. Importantly, the error terms of the generalised model were homogeneously distributed across Session 1 and Session 2 (F=0.17, p=0.7), suggesting that inter-individual variation was similar across measurement sessions. As in Study 1, these results suggest that the two sets of

measurements were not only correlated but that they were also substantially consistent at the individual level. Although the confidence interval width was larger in Study 2 than in Study 1, it is important to note that Study 1's confidence interval is contained within Study 2, which suggests that lexical the test-retest reliability of our lexical entrainment task remains similarly high in the short- and the long-term.

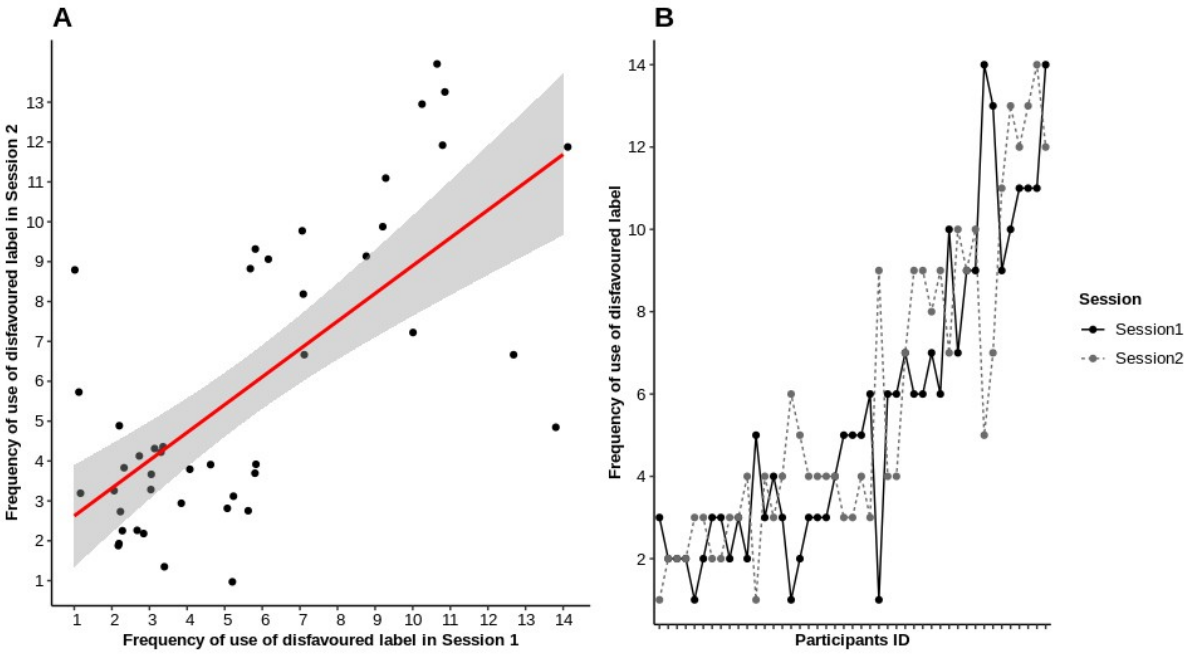

Figure 5. A. Positive correlation (*r*=.61) between the frequency of use of disfavoured names in Session 1 (x-axis) and Session 2 (y-axis). Points are jittered. The red line represents a linear regression between participants' scores in each session, while the grey shadow corresponds to a non-parametric regression smooth. **B**. Individual participants' scores in Session 1 (black, solid line) and Session 2 (grey, dashed line).

The results from Study 2 replicate the finding that lexical entrainment is an effect that can be reliably elicited at the group level, across a range of items, when participants believed they were interacting online with a remote partner. It also shows that our lexical entrainment task reaches a substantial level of test-retest reliability over a seven-to-eight day period. However, although the reliability of our task is still substantial over a week, it is lower than in

the first study, which suggests that lexical entrainment may be influenced by situational factors.

Follow-up Analyses

The analyses reported above demonstrate that our lexical entrainment task elicits robust lexical entrainment effects at the group level, and reaches a substantial level of test-retest reliability not only across sessions separated by only two minutes, but also across sessions separated by seven-to-eight days. However, the ICC coefficient was slightly higher in the short-term (Study 1) than in the long-term (Study 2). Although Study 1's confidence interval is contained within Study 2, which suggests that the test-retest reliability of our lexical entrainment task is substantial in the short- and the long-term, in this section we report an additional analysis assessing the overall within-participants reliability of our task across results from both Study 1 and Study 2 together.

To do this, we used a Pearson correlation and both an ANOVA-based and a mixed effects model approach. To account for possible variances in the data explained by reliability differences between Study 1 and Study 2, we assessed the overall reliability of our task using an adjusted mixed-effects model, which included Study as fixed effect (sum contrast coded, i.e., -1/1). Across the two studies, participants used disfavoured names on 37% of critical trials (SD=27%) in Session 1, and on 39% of critical trials (SD=27%) in Session 2. We found a significant positive correlation between individuals' rates of lexical entrainment in Session 1 and in Session 2 ($r=.68$, $p<.0001$; 95% CI [0.56, 0.78]). The ANOVA-based approached indicated an obtained ICC value of .68 ($p<.0001$) with a 95% confidence interval between .56 and .78, and this was confirmed by a generalised mixed effects model approach, which also showed an ICC value of .68 ($SE=.06$, $p<.0001$) with a 95% interval between .54 and .77. As expected, the error terms of the generalised model were homogeneously distributed across Session 1 and Session 2 ($F=.01$, $p>.05$), suggesting that inter-individual variation was similar

across measurement sessions. Critically, the error terms of our model were also homogeneously distributed by Study ($F=.14$, $p>.05$), suggesting that inter-individual variation was similar in Study 1 and Study 2. These results, taken together with the finding that Study 1's CI is contained within Study 2's CI, support that our lexical entrainment task reaches a substantial level of test-retest reliability both in the short-term and in the long-term.

In further follow-up analyses, we addressed three critical features that could have affected the precision of measurement of our task: 1) sample size used in each study, 2) number of critical trials in our task, and 3) measurement independence between trials.

Although our samples are larger than many previous test-retest reliability studies in psychology and psycholinguistics (e.g., Larson, Baldwin, Good, & Fair, 2010; Strauss, Allen, Jorgensen, & Cramer, 2005; Arnon, 2019), it nevertheless is important to understand whether the precision of our estimates could be substantially increased if our sample size had been much larger, at least within the realms of practical possibility (cf. Schönbrodt & Perugini, 2013). For instance, if we had recruited twice the number of participants we used, would the precision of our measurement be substantially improved? Or, if we had used only half of the participants we recruited, how precise would our measurement be? To interrogate how sample size affected the precision of our estimates, we conducted a resampling analysis examining how test-retest reliability between Session 1 and Session 2 changed, as we varied the number of participants in each study. This thus provided us with a window into the effect of varying sample size, up to the limit of our instrument's measurement error.

To do this, we repeatedly compared the by-participant correlation across Session 1 and Session 2 over subsets of between 15 participants and the total number of participants in each study (Study 1 = 55, Study 2 = 45). We did this 1000 times for each sample size, drawing participants with replacement, and we used the resulting data to calculate mean correlations across sessions, and to estimate 80% confidence intervals. The resulting

Pearson’s correlations and their 80% confidence intervals are illustrated in Figure 6. By analysing and extrapolating from these data, we can assess the effects of sample size. As expected, increased sample size leads to narrower confidence intervals in both studies. However, the benefit of increasing the sample size seems to level off almost completely within the scope of our resampling analyses. Although this level off may result in part from the limits imposed by the measurement error of our task, they also suggest that increasing our sample size within the limits of what is practically possible for a psychology study would not change the overall interpretation of our retest reliability results. In particular, in Study 1, by about 35 participants the confidence interval's lower bound stabilises around .55 and the higher bound stabilises around .8, suggesting that with only 35 participants we would have found very similar results to the ones we found using 55 participants, which in turn suggests that increasing the number of participants would not make a difference in the interpretation of our short-term reliability results. In Study 2, by about 35 participants the confidence interval already ranges between around .4 and .77, suggesting that a sample size of 35 participants would provide a similar long-term reliability estimate to the one we obtained with an n of 45, which in turn suggests that increasing the number of participants would not affect the interpretation of our long-term reliability results either. Taken together, these results suggest that increasing our sample size would not have a substantial impact on the interpretation of our results: Test-retest reliability is substantially stable both in the short- and the long-term. However, our resampling analyses show that, as expected, a larger sample size would provide a more precise measurement of lexical entrainment, which would of course be helpful for future work.

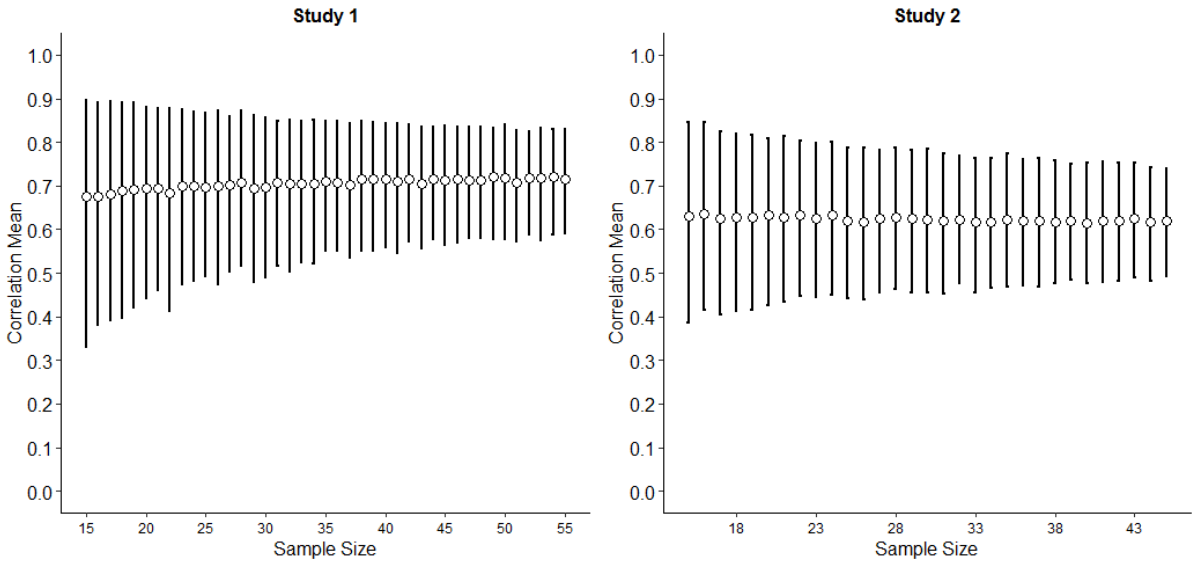

Figure 6. Mean correlations and 95% confidence intervals between participants' entrainment score in Session 1 and Session 2 (y-axis) for increasing number of sample size (x-axis).

Second, and relatedly, we addressed how many binary responses (or trials) are actually necessary to adequately capture the degree to which an individual tends to entrain. For example, if our fourteen-trial instrument were instead twice as long, would the benefit of enhanced measurement outweigh the cost of the additional trials? Or if our instrument were only half as long, how accurate would our measurements be? To provide initial answers to these questions, we conducted a second resampling analysis, now examining how test-retest reliability between Session 1 and Session 2 changed as we varied the number of trials in Session 1. This provided us with a window into the effect of varying trial number, up to the limit of Session 2's measurement error.

To do this, we repeatedly compared the by-participant correlation across Session 1 and Session 2, but where Session 1 scores were now calculated for randomly sampled subsets of between 1 and 14 trials. We did this 1000 times for each number of trials, drawing trials by participant without replacement, and we used the resulting data to calculate mean correlation across sessions, and to estimate 95% confidence intervals. The resulting Pearson's

correlations are illustrated in Figure 7, and show that, for both studies, reliability increased as the number of trials increased. Importantly, this figure suggests that different numbers of trials are needed to reach the same levels of short-term vs. long-term reliability. In particular, a substantial level of short-term reliability (a correlation greater than .6) can be reached with a relatively small number of trials: By about 8-to-10 trials, the confidence intervals in Study 1 no longer range less than .6, and thus a short 8-to-10 trial instrument may be appropriate for studies that require only somewhat precise measurement. However, Figure 7 also shows that by about the same 8-to-10 trials, long-term reliability reaches only a moderate level in Study 2, i.e., confidence intervals no longer range below .4. In addition, Figure 7 shows that the benefits of increasing the number of trials seem to decrease as the number of trials gets higher in each study, and they start to level off by about 10-to-11 trials in both studies, but there is no absolute level off within our sampling scheme. Although this suggests that the reliability of our instrument might in principle keep increasing if the number of trials reaches over 14, it also suggests that the benefits of adding more trials are not likely to change the interpretation of our results, i.e., that lexical entrainment is a substantially stable behavior both in the short-term and in the long-term, and that the long-term reliability of lexical entrainment is only slightly lower than its short-term reliability. However, these results highlight that although a short 8-10 trial instrument may be appropriate for studies that require only somewhat precise measurement of lexical entrainment, longer instruments might be better-suited for correlational studies in general, and individual differences in particular.

Moreover, these results suggest that an instrument with more trials would be useful to distinguish particular individuals who scored close to each other, providing a more precise measurement of their tendency to lexically entrain. For instance, inspection of Figures 3b and 5b suggests that a fairly large proportion of participants entrained only a small number of times, and thus an instrument with a greater number of trials could be useful to distinguish

these particular individuals, providing a more precise measurement of their tendency to lexically entrain, both in the short- and the long-term.

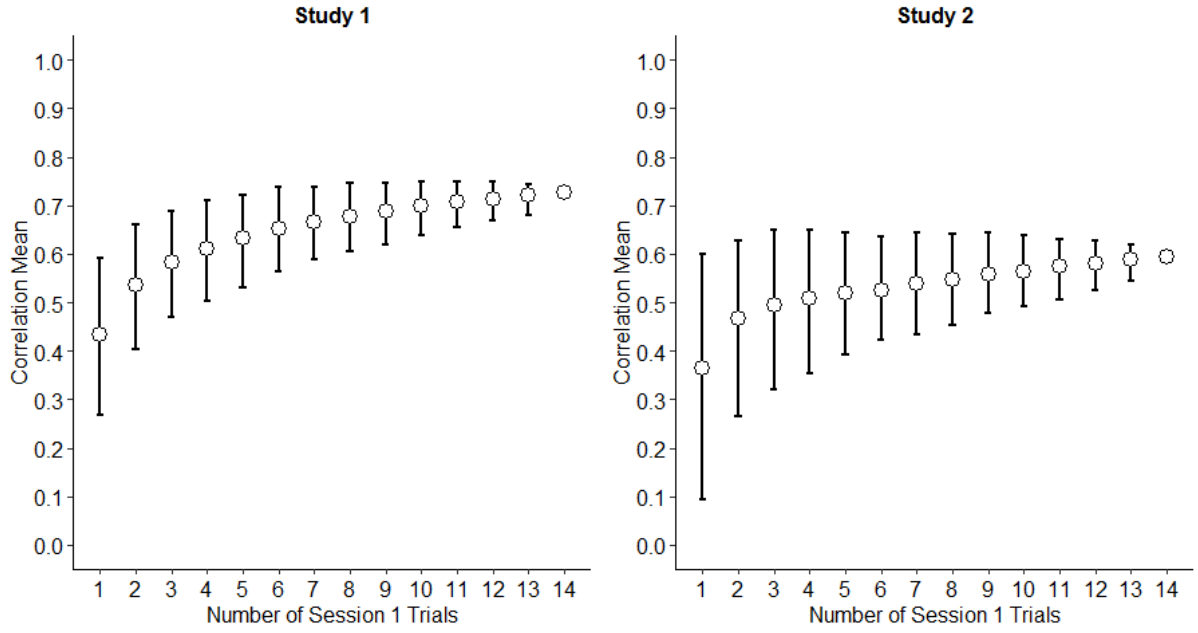

Figure 7. Mean correlations and 95% confidence intervals between participants' entrainment score in Session 2 and participants' entrainment by Session 1 trials (y-axis) for increasing number of Session 1 trials (x-axis).

Third, we additionally investigated the measurement independence between trials. We have argued that a critical methodological feature of this instrument is that each trial provides a relatively independent measure of participants' tendency to lexically entrain, such that whether participants entrained on one trial should not have a direct causal effect on the likelihood that they would entrain on the next trial. This is because we used different items on each trial (so that participants never entrained to the same name twice), and because the low-frequency names that we used were not drawn from any particular dialect or register (e.g., there is no British English dialect or register that standardly uses *brolly* for umbrella and *pillow* for cushion), so that there was no higher-order reason for using a low-frequency name.

To evaluate this argument we regressed entrainment against trial number for each study. Crucially, if trials are indeed independent of one another, then the degree of lexical entrainment should not vary strongly over the course of the experiment. Trial number was entered as a fixed effect (values were centred and standardized), and participant and items were treated as random effects. We included by-participants random slopes for trial number. As Figure 8 shows, there was a slight numerical tendency for reduced entrainment across each Study, but the effect of trial number on participants' tendency to use disfavoured names was not significant in either study (Study 1: β=-0.27, p=.22, $\chi2(1)$=1.47, p=.22; Study 2: β=-0.35, p=.055, $\chi2(1)$=3.37, p=.06)[ii]. These results are thus consistent with behavior on each trial being relatively independent of behavior on previous trials, and thus suggest that individual-level measurement error did not increase throughout the task as a function of trial dependence. This finding is thus consistent with the claim that our instrument succeeded in measuring individual-level behavior, because each trial provided a relatively independent instance of measurement, and in turn supports that our study design is well-suited for correlational research.

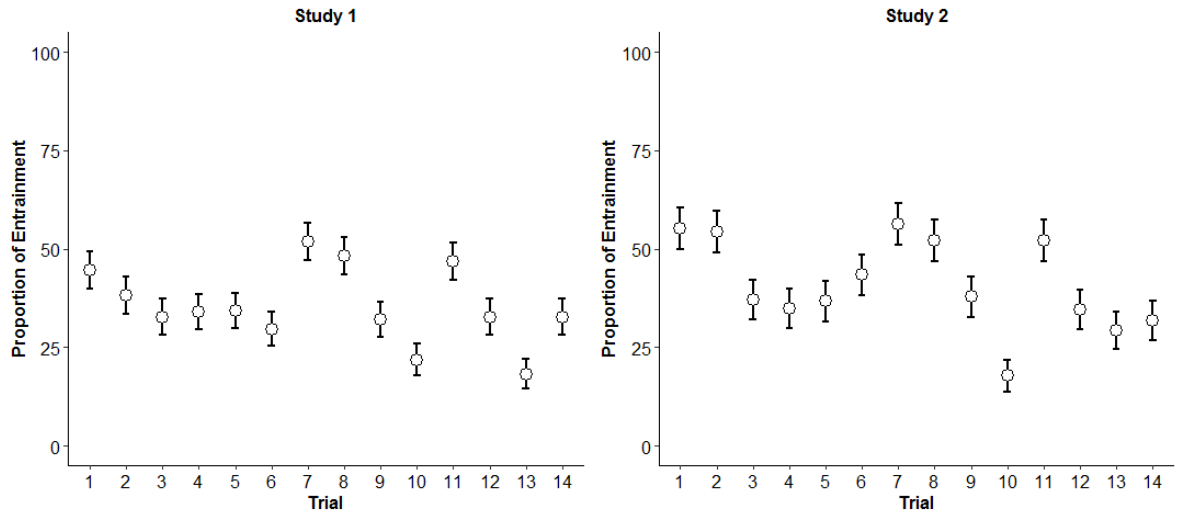

Figure 8. Percentage of lexical entrainment and standard error (y-axis) by trial number (x-axis). The figure includes trials from Session 1 and Session 2 for each study.

Discussion

Language use is variable not only within individuals but also between individuals. Studying individual variation in language processing is critical for understanding language use, but presents important methodological challenges that are quite distinct from the challenges posed when conducting psycholinguistic experiments. In particular, one important concern is that the test-retest reliability of some experimental psycholinguistic tasks may not be sufficient to allow the study of individual variation. It is therefore necessary to establish their level of consistency in advance.

In this paper, we have established the consistency of an instrument for studying lexical entrainment, which is the tendency to reuse a partner's lexical choice. Participants completed two sessions of a task in which they alternated turns with a partner to match and name pictures with two possible names. Participants always named the pictures after the partner, who used only disfavoured names, and we measured lexical entrainment as the proportion of trials on which participants repeated the disfavoured names. In Study 1, the two measurement sessions took place two minutes apart; in Study 2, they took place seven-to-eight days apart. We found robust evidence for strong lexical entrainment in both studies. Although the disfavoured names had a frequency of use of 7% in a spontaneous naming-task, participants in our main experiment used disfavoured names 36% of the time in Study 1 and 41% of the time in Study 2. But more importantly, the studies reported here demonstrated that lexical entrainment was relatively consistent across sessions, both in terms of participants' relative ranking and in terms of their absolute tendency to entrain. In particular, individuals' tendency to lexically entrain reached a substantial level of reliability both in the short-term (i.e., across sessions separated by two minutes) and in the long-term (i.e., across sessions that were separated by seven-to-eight days). This pattern suggests that lexical

entrainment has promise as an instrument for carrying out not only group-level experimental research but also individual-level correlational research.

What can these key findings – that lexical entrainment is robust, and that this effect is substantially stable within individuals – tell us about entrainment as a general phenomenon? Although our current results are not informative about exactly what mechanisms underlie lexical entrainment in this specific context (e.g., audience design, priming, pro-sociality), they provide evidence that these mechanisms contribute to making lexical entrainment a substantially stable behavior when the communicative situation is held roughly constant, as in our studies. Critically, the finding that the reliability of lexical entrainment reaches a substantial level even across 7-to-8 days suggests that inter-individual variability in the degree of lexical entrainment is underlain by stable individual differences, although what those traits are remain to be seen. For example, under the assumption that lexical entrainment interacts with social affect and pro-sociality (e.g., Van Baaren, 2003; Palomares et al., 2016), a person with a high propensity towards pro-sociality (agreeableness) may in general be more likely to lexically entrain than a person with a high propensity to feel very anxious in social situations (neuroticism) (e.g., see Gill, Harrison, & Oberlander, 2004, for individual differences in interpersonal syntactic priming). Alternately, assuming that lexical priming effects underlie lexical entrainment (e.g., Pickering & Garrod, 2004), individuals who are more susceptible to lexical priming may exhibit a greater basal tendency to reuse a partner's lexical label (i.e., a recently processed lexical label) than individuals who are less susceptible to lexical priming.

Importantly, previous findings showing that lexical entrainment is influenced by situational factors (e.g., Brennan & Clark, 1996; Branigan et al., 2011, Van Baaren et al., 2003), taken together with our finding that lexical entrainment is stable within individuals, suggest that research on lexical entrainment could also be informative about how individual

differences and situational factors interact with each other during language use. The degree of lexical entrainment may vary between individuals not only depending on their basal propensity towards entrainment and/or on features of the communicative situation, but also as a result of the interaction between each individual's disposition towards lexical entrainment and relevant situational factors. For example, lexical entrainment research can be informative about whether personality traits interact with characteristics of our conversational partner (e.g., are individuals who are more agreeable more likely to entrain to a non-prototypical community member than individuals who are less agreeable?).

In addition, the attested reliability of our instrument shows that, in principle, null findings in previous correlational studies on lexical entrainment may be indicative of true null associations, such as the null association reported by Hopkins et al. (2017) between individuals' propensity towards lexical entrainment and their perspective-taking skills. That said, there are still reasons for caution in interpreting such data. For one, while we have shown that lexical entrainment has strong test-retest reliability, the reliability of other instruments used in such studies (e.g., the mental attribution instrument used in Hopkins et al.) is not always known. In addition, it is also possible that test-retest reliability may vary across populations. We demonstrated substantial reliability in British English adults recruited and tested online, but reliability may differ for, e.g., the typically developing and autistic children studied by Hopkins and colleagues.

Going forward, we recommend that any instrument for studying lexical entrainment should be appropriately validated with the population that it will be applied to, and this applies not only for measuring reliability but also for appropriately norming the low-frequency names that are used as stimuli. For instance, both studies reported in this paper used online participants, native speakers of British English who were users of Prolific

[https://prolific.ac/]; we therefore normed our materials in a different sample drawn from the exact same population.

Although the attested substantial reliability of our task suggests that it is well-suited for correlational research, we reported additional analyses that interrogated which methodological features of our task contribute to this, and which features could be improved (see Follow-up Analyses section). In particular, we recommend the use of a range of different items to ensure independence of measurement in each trial and thus prevent measurement error from increasing throughout the task; accordingly, we encourage the incorporation of analyses to rule out trial order effects on entrainment. Moreover, we found that increasing the number of trials had a positive impact on the level of reliability of our task, although the benefits of increasing the number of trials decreased as the sample size increased. However, it is still possible that a task with a greater number of trials could provide a more precise measurement of individuals' basal tendency to lexically entrain. Given that a high number of participants in our studies scored close to zero, using a task with a greater number of trials would also allow distinguishing these participants, providing a more accurate measurement of each individual's tendency to entrain.

Strikingly, the fact that lexical entrainment is stable within individuals not only suggests that the mechanisms supporting lexical entrainment are stable when individuals believe themselves to be interacting with a remote player, but also that the way in which we make lexical choices can potentially reflect stable individual differences in how we process language. Lexical entrainment research has already suggested that language production is affected by memory, perspective-taking, and pro-sociality, among other factors (Branigan et al., 2011; Brennan & Clark, 1996; van Baaren et al., 2003). Future studies using an individual differences approach can further develop accounts of language processing, by addressing questions such as whether the degree to which individuals display lexical entrainment might

be predicted by social psychological factors (e.g., perspective-taking skills), personality traits (e.g., degree of pro-sociality or agreeableness), cognitive effects (e.g., ease of lexical access) or even by demographic variables (e.g., gender and age). Importantly, the fact that this lexical entrainment task is well-suited to both experimental and correlational research is promising for understanding how situational factors and individual differences interact during language processing, which has hitherto attracted little attention in psycholinguistic research.

Additionally, the finding that lexical entrainment is fairly stable across measurement sessions opens up the question of whether other types of entrainment, e.g., syntactic or phonetic, might be similarly stable. This is theoretically important because it is currently unclear whether entrainment is underpinned by domain-general mechanisms that might cause a person to entrain at similar rates for both lexical and grammatical stimuli, or whether different types of entrainment rely on importantly different processes; for example, lexical entrainment may be more sensitive to perspective-taking abilities than syntactic entrainment (e.g., Branigan et al., 2011). To test these issues, we need instruments that can reliably measure entrainment at various levels of linguistic structure. We have shown how this can be done for lexical entrainment; future studies should similarly focus on validating the test-retest reliability of instruments measuring linguistic entrainment at other levels of structure. To this end, and in light of the results reported in the Follow-up Analyses section, we suggest that future linguistic entrainment instruments aimed at studying individual differences should be designed to measure behavior independently in each trial (i.e., testing entrainment to different linguistic structures) and should include a number of trials as large as practically possible.

Likewise, the attested reliabilities of our instrument can have important consequences for theories of non-linguistic behavioral mimicry. During social interaction, people not only entrain to their interlocutor’s language use, but also to other non-linguistic behaviors, including body postures, gestures, facial expressions, and emotional reactions (see Chartrand

& Lakin, 2013, for a review). And, although there is a general tendency to conceptualise linguistic entrainment as a kind of behavioral mimicry (e.g., van Baaren et al., 2003; Chartrand et al., 2005), it remains unclear whether linguistic and non-linguistic imitative behaviors are supported by the same constructs. Future studies on lexical entrainment could illuminate this debate by interrogating whether the degree of lexical entrainment might be predicted by degrees of non-linguistic behavioral mimicry (e.g., mimicry of facial expressions of emotions) and their underlying constructs (e.g., measures of social competence, e.g., Mauersberger et al., 2015). For instance, are individuals who are more likely to entrain to a partner's lexical choice also more likely to mimic a partner's emotional facial expressions? If so, are both their tendencies to lexically entrain and to mimic their partner's emotional facial expression correlated with the same potential underlying mechanism (e.g., social competence)?

One potential concern that could be raised about this instrument is the degree to which it solely measures entrainment. For example, in an alternative formulation of this instrument, we might have measured entrainment as the difference between each participants' tendency to use low-frequency names when primed, and when not primed. However, although such a design might be ideal in theory, we suggest that its benefits in practice are small, and its disadvantages are serious. Its key advantage would be to account for participants' baseline tendencies to use the low-frequency labels, which could partially explain the correlations across sessions. However, our design accounted for such a tendency by using different labels across the two sessions, and by using low-frequency labels that were not systematically drawn from a particular dialect or register.

Moreover, the alternative subtractive design would have significant difficulty providing precise measures of entrainment. In particular, taking a baseline measure would a) prime participants to use higher-frequency labels for the relevant objects, thus minimizing

subsequent entrainment (Branigan et al., 2011); b) reduce the number of trials available to measure entrainment, compromising statistical power; and c) could potentially increase measurement error, as each subject's score would now have two sources of error: one for measuring the baseline and one for measuring entrainment. These considerations, and the fact that entrainment is reliably seen at the group level, suggest that our single-measure instrument provides the most efficacious way of capturing this phenomenon.

In sum, we have argued that the study of individual variation in language processing requires instruments that elicit a wide range of effects between individuals and that have high test-retest reliability. It is therefore necessary to develop tasks that meet these two criteria as a necessary precursor to testing theoretical accounts of language processing. In this paper, we have shown how this can be done for the case of lexical entrainment, a phenomenon that is informative of individual variability in how we make lexical choices in particular, and in how we process language more generally. In particular, we have shown that online naming-tasks measuring lexical entrainment can in principle be informative about factors affecting language processing. We therefore encourage the use of this instrument – adapted appropriately to the population of interest - for the study of individual differences.

References

Arnon, I. (2019). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01205-5

Berchtold, A. (2016). Test–retest: Agreement or reliability? *Methodological Innovations, 9*, pp. 1-7, https://doi.org/10.1177/205979911667287

Branigan, H. P., & Messenger, K. (2016). Consistent and cumulative effects of syntactic experience in children's sentence production: Evidence for error-based implicit learning. *Cognition*, *157*, 250-256.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41-57.

Branigan, H. P., Tosi, A., & Gillespie-Smith, K. (2016). Spontaneous lexical alignment in children with an autistic spectrum disorder and their typically developing peers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(11), 1821-1831.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482-1493.

Chartrand, T. L., Maddux, W. W., & Lakin, J. L. (2005). Beyond the perception-behavior link: The ubiquitous utility and motivational moderators of nonconscious mimicry. *The new unconscious*, 334-361.

Chartrand, T. L., & Lakin, J. L. (2013). The antecedents and consequences of human behavioral mimicry. *Annual review of psychology*, *64*, 285-308.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American journal of mental deficiency*, *86*(2), 127-137

Clark, H. H. (1996). *Using language*. Cambridge university press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1-39.

Eckert, P., & Rickford, J. R. (Eds.). (2001). *Style and sociolinguistic variation*. Cambridge University Press.

Gallois, C., & Callan, V. J. (1991). Interethnic accommodation: The role of norms. *Contexts of accommodation: Developments in applied sociolinguistics. Edited by Giles, H., Coupland, J. and Coupland, N., New York, NY: Cambridge University Press. 245-269.*

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181-218.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in cognitive sciences*, *8*(1), 8-11.

Gregory, M., & Carroll, S. (2018). *Language and situation: Language varieties and their social contexts*. Routledge.

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.

Hopkins, Z., Yuill, N., & Branigan, H. P. (2017). Inhibitory control and lexical alignment in children with an autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, *58*(10), 1155-1165.

Horton, W. S. (2014). Individual differences in perspective taking and field-independence mediate structural persistence in dialog. *Acta psychologica*, *150*, 41-48.

Horton, W. S., & Gerrig, R. J. (2002). Speakers' experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language*, *47*(4), 589-606.

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127-142.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *L. Pervin and O.P. John (Eds.), Handbook of personality: Theory and research (2nd ed.). New York: Guilford*, 102-138.

Juncos-Rabadán, O., Facal, D., Rodríguez, M. S., & Pereiro, A. X. (2010). Lexical knowledge and lexical retrieval in ageing: Insights from a tip-of-the-tongue (TOT) study. *Language and Cognitive Processes*, *25*(10), 1301-1334.

Kaschak, M. P., Kutta, T. J., & Jones, J. L. (2011). Structural priming as implicit learning: Cumulative priming effects and individual differences. *Psychonomic bulletin & review*, *18*(6), 1133-1139.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in cognitive sciences*, *22*(2), 154-169.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Larson, M. J., Baldwin, S. A., Good, D. A., & Fair, J. E. (2010). Temporal stability of the error-related negativity (ERN) and post-error positivity (Pe): The role of number of trials. *Psychophysiology*, *47*(6), 1167-1171.

Leaper, C., & Ayres, M. M. (2007). A meta-analytic review of gender variations in adults' language use: Talkativeness, affiliative speech, and assertive speech. *Personality and Social Psychology Review*, *11*(4), 328-363.

March, E., Wales, R., & Pattison, P. (2003). Language use in normal ageing and dementia of the Alzheimer type. *Clinical Psychologist*, 7(1), 44-49.

Mauersberger, H., Blaison, C., Kafetsios, K., Kessler, C. L., & Hess, U. (2015). Individual differences in emotional mimicry: Underlying traits and social consequences. *European Journal of Personality*, *29*(5), 512-529.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, *1*(1), 30.

Meyer, A. S. (1996). Lexical access in phrase and sentence production: Results from picture–word interference experiments. *Journal of memory and Language*, *35*(4), 477-496.

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews*, *85*(4), 935-956.

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, *4*(5), 648-654.

Palomares, N. A., Giles, H., Soliz, J., & Gallois, C. (2016). Intergroup accommodation, social categories, and identities. *Giles, H. (Ed.), Communication accommodation theory: Negotiating personal relationships and social identities across contexts*, *Cambridge University Press*, 123-151.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics, 79*(2), 637-659.

Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and language*, *39*(4), 633-651.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(2), 169-190.

Polit, D. F. (2015). Assessing measurement in health: beyond reliability and validity. *International journal of nursing studies*, *52*(11), 1746-1753.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609-612.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of memory and language*, *81*, 105-120.

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. Behavior research methods, 49(2), 418-432.

Strauss, G. P., Allen, D. N., Jorgensen, M. L., & Cramer, S. L. (2005). Test-retest reliability of standard and emotional stroop tasks: an investigation of color-word and picture-word versions. *Assessment*, *12*(3), 330-337.

Van Baaren, R. B., Holland, R. W., Steenaert, B., & van Knippenberg, A. (2003). Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, *39*(4), 393-398.

[i] This phenomenon is also sometimes referred as lexical alignment, to imply that the phenomenon of lexical entrainment is driven by two interlocutors' simultaneous activation of the same lexical representation, i.e., corresponding to a shared conceptualisation of a specific entity (Clark & Wilkes-Gibbs, 1986; Pickering and Garrod, 2004). We use the term lexical entrainment as we focus on speakers' overt reuse of their partner's lexical choices, and thus we do not consider the underlying mechanisms (e.g., lexical priming, Pickering & Garrod, 2004; audience design, Clark, 1996) that might give rise to this observable behavior.

[ii] Importantly, note that these results also rule out any possibility that participants were primed to use low-frequency words more generally by their 'partner's' use of a high proportion of low frequency words (i.e., the disfavored names for experimental items). If participants were learning to use low frequency words over the course of the study, then we would expect them to show increasing entrainment as the study progressed. But instead, we found that entrainment rates gradually declined over the course of the study.