Are faster participants always faster? Assessing reliability of participants' mean response speed in picture naming

Pamela Fuhrmeister[1], Shereen Elbuy[1], & Audrey Bürki[1]

[1] Universität Potsdam

Author Note

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 317633480 – SFB 1287, Project B05 (Bürki).

Correspondence concerning this article should be addressed to Pamela Fuhrmeister, Karl-Liebknecht-Straße 24-25 14476 Potsdam Germany. E-mail: pamela.fuhrmeister@uni-potsdam.de

Are faster participants always faster? Assessing reliability of participants' mean response speed in picture naming

In psycholinguistic research on language production processes, studies tend to examine behavior at the group level. In the present study, we focus on word production. A measure of choice in this field is the production latency, or time required to prepare a word for production following the presentation of a stimulus, often a picture (e.g., Alario et al., 2004; Costa, Strijkers, Martin, & Thierry, 2009; Fargier & Laganaro, 2019; Fieder, Wartenburger, & Rahman, 2019; Laganaro, Valente, & Perret, 2012; Meyer, 1996; Pinet, Ziegler, & Alario, 2016; Rabovsky, Schad, & Rahman, 2016; Roelofs, 2004). In many of these studies, production latencies from different experimental conditions are compared (e.g., with and without priming, high and low frequency), but variability in effective speed across participants is of little interest: Participants are treated as a random variable in the analysis to ensure that experimental effects generalize to all participants, irrespective of their effective speed. Several effects found in language production studies (e.g., word frequency, age of acquisition, priming effects) have been reported and have been shown to be replicable across groups and studies, suggesting that naming latencies are a good index of language production processes. Yet, individual speakers show a wide range of variability in the time needed to prepare spoken words.

The extent of inter-individual variability in language production has for instance been reported for picture naming tasks (e.g., Bürki, 2017; Laganaro et al., 2012; Valente, Bürki, & Laganaro, 2014). In contrast to the dominant approach, several recent studies have focused on this variability with the idea that inter-individual differences in picture naming speed can inform our understanding of the architecture of the language production system and how it relates to non-linguistic abilities. Some of these studies have for instance taken a correlational approach to investigate relationships between participants' picture-naming speed and performance on cognitive tasks. For example, correlations between measures of sustained attention and picture-naming speed have been reported (Jongman, 2017; Jongman, Meyer, & Roelofs, 2015; Jongman, Roelofs, & Meyer, 2015). Other studies have observed relationships between working memory measures and picture-naming latencies (Piai & Roelofs, 2013; Shao, Roelofs, & Meyer, 2012; but see Klaus & Schriefers, 2018), and a few studies have found that inhibition skills are related to picture-naming speed (Lorenz, Zwitserlood, Regel, & Abdel Rahman, 2019; Shao, Meyer, & Roelofs, 2013; Sikora, Roelofs, Hermans, & Knoors, 2016).

The hypothesis that (picture) naming speed relates to participants' cognitive abilities relies on the assumption that relatively faster speakers are faster each time they are tested. It further implies that differences in naming speed are not specific to the details of the task at hand (e.g., timing of individual trials). The aim of the present study is precisely to assess this reliability for picture naming latencies. We assess the reliability of individual differences in picture-naming speed within the experimental session and over time. A second aim is to test whether participants are still ranked by mean naming speed in a similar way across different manipulations of the same task.

In the remainder of this introduction, we first briefly discuss the concept of reliability in the context of inter-individual differences and how it can be assessed. We then discuss the importance of establishing the reliability (or lack thereof) of participant production speed for language production research. Finally, we introduce the current study in some detail.

Reliability and inter-individual differences

Broadly, the term reliability refers to the consistency or trustworthiness of a measure (Urbina, 2014), although the term has slightly different meanings in different contexts. For example, in studies comparing performance on a task between groups or experimental conditions, the term reliability is sometimes used to describe an experimental effect that is replicable across different samples of participants (Hedge, Powell, & Sumner, 2018). In studies focusing on inter-individual differences, scores on a test or task are said to be reliable if a participant's performance, measured on different occasions, is highly similar. In psycholinguistics, the measure we are often interested in is a participant's mean score on a task. In studies on individual differences, it is important to use measures with low measurement error, or high reliability (Parsons, Kruijt, & Fox, 2019) because such measures will produce consistent inter-individual differences. That is, if the performance of participant A is better than that of participant B on a given day or for a subset of trials, participant A will be better than participant B on a different set of trials or when tested on a different day. In certain subfields of psychology that have traditionally focused on individual differences (e.g., differential psychology), reporting reliability is standard, for example in personality research (Viswesvaran & Ones, 2000) or intelligence testing (Donders, 1997). However, this is not often the case in cognitive psychology (Parsons et al., 2019).

Reliability of participant scores can be measured in different ways. A common way of assessing reliability within an experimental session is to correlate two halves of the data (e.g., even vs. odd trials, first half vs. second half), which is referred to as split-half reliability (Parsons et al., 2019; Urbina, 2014). The term test-retest reliability is often used to refer to reliability over time (e.g., at different sessions, Urbina, 2014). To estimate test-retest reliability of a given measure, the same participants are tested two (or more) times and the correlation between their performance at different time points is computed (Urbina, 2014). If inter-individual differences are reliable, correlations between participants' performance across trials or sessions should be high. To reiterate, saying that a measure is reliable amounts to saying that inter-individual differences are themselves reliable.

An important question in any study of reliability is what evidence is needed to determine whether performance on a task is reliable. For instance, a correlation of .2 could be statistically significant, but a weak correlation that is statistically significant may not be very meaningful when considering the issue of reliability. If a Pearson correlation of .2 was obtained when assessing test-retest reliability for a measure, it means that 96% of the variance remains unexplained. A correlation of .9 would be much more convincing because only 9% of the variance is left unexplained. Though there is not one standard cutoff for an effect size that is considered to be indicative of good reliability, most authors suggest that a measure is reliable enough if the correlation is at least 0.7 or 0.8 (Jhangiani, Chiang, & Price, 2015), though the standard might be higher for clinical situations (Hedge et al., 2018; Nunnally, 1994; Parsons et al., 2019).

Reliability of participants' speed in picture naming

Despite the recent interest in individual differences in word production research, we still lack studies testing the reliability of word production measures. In the present study, we focus specifically on reliability of picture-naming measures. The available evidence suggests that picture-naming speed is relatively reliable *within an experimental session*. For example, Shao et al. (2013) reported high split-half reliability (correlation of even and odd trials) for mean picture-naming speed (r = .91) in a picture-naming task. In addition, we find a high correlation between even and odd trials in our own picture-naming data, ρ = 0.97 (95% credible interval [0.93, 1.00], reanalysis of Fuhrmeister, Madec, Lorenz, Elbuy, & Bürki, 2022). Shao et al. (2012) had participants name pictures of objects and actions, and they found a fairly high correlation between participants' mean naming latencies of objects and actions (r = .74) and replicated this high correlation (r = .86) in a later paper (Shao, Roelofs, Acheson, & Meyer, 2014). This suggests participants are also relatively consistent in their speed of naming two different kinds of stimuli.

Importantly, finding a high correlation *within* an experiment is necessary but not sufficient to conclude that picture-naming tasks generate reliable differences across participants. Naming latencies for the different trials of an experiment could be correlated because participants set a goal that is specific to the experimental setting, or feel more or less motivated or alert on a given day due to temporary factors, such as sleep quality and duration the night before, time of day that the experiment took place, or temporary emotional state. Notably, if the consistency of participants' naming times within an experimental session is due to one of the aforementioned factors, correlations may be expected between naming times and performance on non-linguistic tasks because the participants are tested on the language production and non-linguistic tasks on the same day, within the same session. As a result, these correlations may not reflect individual differences in cognitive skills. If a participant's effective speed can partly be explained by their cognitive abilities, we expect naming latencies to also be consistent over time, i.e, when participants are tested on different days.

We further expect that differences across participants will persist when the experimental setting is modified such that it prompts differences in response speed (e.g., shorter inter-stimulus intervals, instructions to respond within a given time window). In picture-naming experiments, it is possible that differences across participants arise because they have room to set different goals for the task. If they have ample time, they can select their own pace. For instance, some participants could decide to respond as quickly as possible, while others might prioritize accuracy over speed, or simply select a comfortable response speed given the allotted time. The hypothesis that differences in naming latencies reflect differences in cognitive abilities implicitly assumes that there is limited room for "strategies" or "decisions."

We mentioned above that several of the studies that examined differences in naming speed across participants used a correlational approach, i.e., correlated naming latencies with performance on a cognitive task. In this context, the reliability of participant naming latencies has crucial methodological implications, in that reliability is directly related to statistical power. When assessing the correlation between two measures, e.g., performance on picture-naming and working memory tasks, the reliability of the individual measures constrains the magnitude of the correlation that can be found *between* these measures (Parsons et al., 2019). The number of participants required to detect a correlation between two measures increases when the reliability of these measures decreases (e.g., Hedge et al., 2018, Table 5). Thus, reliability estimates of measures of interest can be used in power calculations to determine the sample size needed to detect effects of a certain magnitude (Parsons et al., 2019). For example, assuming a true correlation of 0.3 between two measures, 133 participants would be necessary to reach the standard significance threshold when the two tasks have a reliability of 0.8; 239 when the reliability is 0.6. Pilot data from our lab shows a correlation of .28 between a measure of attention and naming latencies. Assuming that this number reflects the true effect size, the number of participants required to detect the correlation given a reliability of 0.8 for each task would be 153. If this estimation is correct, sample sizes such as the one we used (n = 45) are likely to be insufficient to detect such correlations. The reliability of some of the measures of cognitive skills that have been used in correlational studies to explain individual differences in picture naming has been tested independently (see e.g., Borgmann, Risko, Stolz, & Besner, 2007 for reliability of the Simon effect; Congdon et al., 2012 for an example involving the stop-signal reaction time task; and see Conway et al., 2005; Klein & Fiss, 1999; Waters & Caplan, 2003 for working memory measures) or reported in the paper along with correlations with picture-naming measures (Jongman, Roelofs, et al., 2015). If the reliability of one or more measures entered into a correlation is low, power to detect these correlations suffers, and the probability of Type II errors increases. Precise estimates of reliability of both measure entered into a correlation are therefore necessary to make sure that the required sample size is tested. Hedge et al. (2018) even describe reliability of a measure as “a prerequisite to effective correlational research.”

Current study

The current study consists of two experiments. Each experiment tests a group of participants at two different sessions that occur between 7 and 14 days apart. Experiment 1 tests split-half reliability and test-retest reliability of simple picture naming (i.e., naming of bare nouns). The same task will be used for both sessions. The implementation of the task, including timing, mimics that of a standard picture-naming task. Participants are presented with a picture and have 3000ms to provide their response.

Evidence of reliability within and between sessions would be in line with the assumption that cognitive abilities of an individual impact picture naming speed. However, if picture naming speed is not reliable over time, this would suggest that inter-individual differences in naming speed do not index general differences in cognitive abilities. This would not mean that previous studies that have found correlations between cognitive skills and picture-naming speed are not informative. It may simply limit the extent to which we can generalize these findings. For instance, these correlations may not mean that individuals who have better attention or inhibition skills are necessarily faster speakers; rather, it may mean that the amount of attention or inhibition applied within a specific task is a more robust predictor of picture-naming speed.

In Experiment 2, we test the correlation between participants' naming speed on a picture-naming task, in which we manipulate the conditions under which participants name the pictures. One condition is a simple picture-naming task like in Experiment 1, and the other is a speeded naming task, in which participants have a limited amount of time to name the picture (i.e., a response deadline). Previous studies have consistently shown that picture- or word-naming speed is faster with a response deadline, at least at the group level (Damian & Dumay, 2007; Kello, Plaut, & MacWhinney, 2000). With this manipulation, we can test whether participants are ranked similarly in speed with or without a response deadline. Relatedly, we can examine whether participants may be engaging in a speed-accuracy trade-off strategy (Heitz, 2014).

In word production studies using a response deadline, the evidence of a speed-accuracy trade-off is mixed. For example, Kello et al. (2000) found similar error rates on speeded and a non-speeded versions of the Stroop task with naming responses. Starreveld and La Heij (1999) and Damian & Dumay (2007, in one out of three experiments) found a speed-accuracy tradeoff in picture-naming experiments, but in both these experiments, participants were specifically instructed to make errors to prioritize speed. Moreover, in these experiments, the speed-accuracy trade-off was examined at the group level. To our knowledge, no studies have looked at individual differences in speed-accuracy trade-offs in picture-naming tasks to determine whether individual participants are engaging in strategies to choose their picture-naming speed. If they do, then the observed inter-individual differences in mean naming speed could in part be due to participant-specific decisions rather than individual differences in cognitive abilities. If participants are engaging in such strategies or picking a specific tempo for the task, we expect that inter-individual differences will be less reliable when measured between picture naming tasks that vary in timing.

Experiment 1

Experiment 1 tested within- and between-session (i.e., split-half and test-retest) reliability of participants' picture-naming speed over two sessions using a simple picture-naming task.

Methods

Participants

Participants were recruited through the online platform Prolific (www.prolific.co), and the study was advertised to native speakers of British English with no history of reading or language disorders. We recruited participants until we had usable data from 50 participants (78 total) because we could reasonably pre-process that amount of data (per experiment) with our current lab resources. Participants were excluded if they did not complete both sessions ($n = 20$), the data were not recorded due to technical errors ($n = 7$), or the recording quality was so low that we were unable to detect the onset of the vocal responses ($n = 1$). We additionally planned to exclude participants if there was an obvious indication they did not follow instructions, for example, if we heard from the recording that they were listening to music, talking to other people, or eating during the experiment. Fairs and Strijkers (2021) did a recent picture-naming study online and found that some participants kept the experiment running in order to get paid for it but did not actually do it. In order to eliminate participants who did not perform the experiment in good faith, we required participants to reach at least 60% accuracy in naming the pictures to be included in the analyses. Those who did not reach this threshold in the first session were not allowed to participate in the second session. This was not necessary in the first experiment because all participants who were not excluded for reasons listed above achieved at least 60% accuracy. Participants were excluded prior to any data analysis, and all excluded participants were replaced so that the final sample size was 50. Participants gave informed consent prior to the experiment and were paid €11 per hour. This study was approved by the ethics board of the University of Potsdam.

Stimuli

We selected 310 pictures from the Multipic database (Duñabeitia et al., 2018) with the highest name agreement ratings that also had corresponding data for frequency and age of acquisition available in relevant databases. The Multipic database provides freely available, colored drawings of 750 words with norms in several languages (Duñabeitia et al., 2018). It has been used in many picture-naming experiments (e.g., Bartolozzi, Jongman, & Meyer, 2021; Borragan, Martin, De Bruin, & Duñabeitia, 2018; Gauvin, Jonen, Choi, McMahon, & Zubicaray, 2018; Zu Wolfsthurn, Robles, & Schiller, 2021), including a recent experiment run online (Fairs & Strijkers, 2021). In cases of duplicate target words for different pictures, one of the pictures was removed and replaced with another. Information on lexical frequency was obtained from the SUBTLEX-UK database (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014), and age of acquisition data was obtained from Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012). The H-index was provided by the database as a measure of name agreement. The H-index takes into consideration how many different names are supplied for the picture as well as the frequency that alternative names are given; a lower H-index indicates higher name agreement. Pictures included in the study had a maximum H-index of 0.52, and at least 88.9% of people gave the modal name when the pictures were normed (Duñabeitia et al., 2018).

We created two different lists from the 310 pictures (155 items each) such that they would be balanced on word frequency, name agreement, and age of acquisition. We chose to equate lists on these three variables because they have been found to be some of the most robust predictors of naming latencies across several studies conducted in several languages (Alario et al., 2004; Cuetos, Ellis, & Alvarez, 1999; Ellis & Morrison, 1998; Snodgrass & Yuditsky, 1996).[1] Lists can be found in Appendix A. The procedure for creating the balanced lists was as follows: We first obtained values of name agreement, frequency, and age of acquisition for all 310 pictures and z-scored these values. These values formed a feature vector for each stimulus item. We then calculated the cosine similarity of the feature vectors for each unique pair of stimuli and sampled one million random sets of 155 pairings of stimulus items and calculated the mean cosine similarity of all the pairings in each set. We chose the set of 155 pairings with the highest mean cosine similarity for the two lists, as these were the most similar in terms of lexical frequency, name agreement, and age of acquisition (cosine similarity = .24). Descriptive statistics on

[1] Imageability is also a robust predictor of naming latencies (Alario et al., 2004); however, we did not have imageability estimates for the pictures used for the present study.

these measurements from each list can be found in Table 1. Reproducible code for list creation can be found at the OSF repository for this project: https://osf.io/v4h2a/. Participants saw one of the two lists of pictures at each session (5 pictures per list for training items; 150 pictures per list for test items), and the order of list presentation was counterbalanced (i.e., half of the participants saw List 1 for Session 1 and half saw List 1 for Session 2). All participants saw the same 5 pictures in each list for training.

Table 1

Mean and standard deviation (SD) of frequency (Zipf scale), age of acquisition (AoA) ratings, and name agreement (H-index) for the words/pictures in each list.

List	Frequency mean	Frequency SD	AoA mean	AoA SD	H-index mean	H-index SD
1	4.17	0.48	5.39	1.43	0.13	0.15
2	4.28	0.64	5.48	1.69	0.13	0.17

Procedure

The experiment consisted of two sessions (see Figure 1). The second session took place between 7 and 14 days after the first session to assess reliability of picture-naming speed over time. All stimuli were presented online using the experiment presentation software PCIbex (Zehr & Schwarz, 2018).

Participants named each picture in a list (5 practice trials, 150 experimental trials) in a simple picture-naming task in each session. At the beginning of each session, participants were familiarized with all of the stimuli by seeing each picture with the printed target word below it on the screen. Participants were asked to study the pictures and were told they will need to recall the name of the pictures for the next part of the experiment.

The picture-naming task began with a brief practice phase of five trials, followed by the main part of the task with 150 trials. Each trial began with a fixation cross that

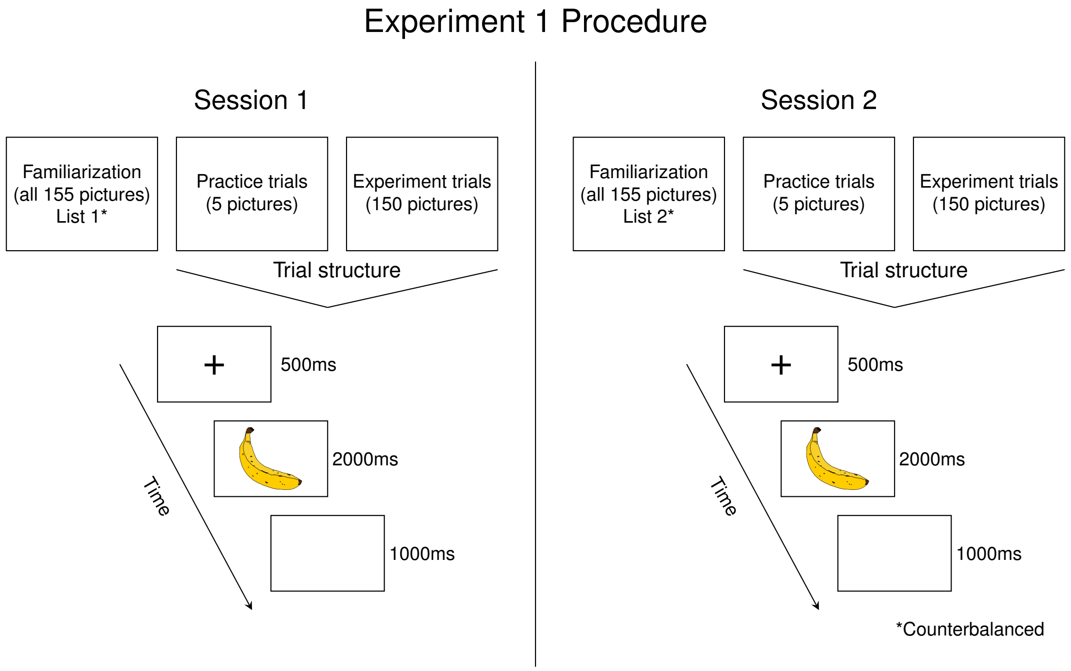
Experiment 1 Procedure
Session 1
Familiarization (all 155 pictures) List 1*
Practice trials (5 pictures)
Experiment trials (150 pictures)
Trial structure
+
500ms
2000ms
1000ms
Time
Session 2
Familiarization (all 155 pictures) List 2*
Practice trials (5 pictures)
Experiment trials (150 pictures)
Trial structure
+
500ms
2000ms
1000ms
Time
*Counterbalanced

Figure 1. Illustration of procedure for Experiment 1. Participants completed the depicted experiment at two different sessions, each with a different stimulus list. The order of the stimulus list presentation was counterbalanced.

appeared in the center of the screen for 500ms, followed by a picture which appeared for 2000ms. Then the picture disappeared and participants saw a blank screen for 1000ms. Vocal responses were recorded from the onset of the picture until the end of the trial. Participants were instructed to name the picture aloud as fast and accurately as they could. All 150 pictures were presented in random order.

Planned analyses

Data preprocessing. Only trials with correct responses were included in the analyses. Incorrect responses included trials for which participants produced the wrong word, exhibited disfluencies (e.g., false starts), and trials on which no response was given within 3000ms (the length of the trial). We did not filter the data for outliers[2]. Picture-naming latencies were calculated for each trial as the time between the picture onset and the onset of the vocal response; the latter was set manually in Praat (Boersma & Weenink, 2021). The preprocessed data set[3] and analysis code for all experiments in the study is publicly available at https://osf.io/v4h2a/.

Split-half (within-session) and test-retest (between-session) reliability. All analyses were done in R (R Core Team, 2021). To estimate split-half and test-retest reliability, we computed the correlation between response times in each half of the data in each session separately (split-half reliability) or between each session (test-retest reliability). Correlations were computed in Bayesian hierarchical models (the correlation of the random effects) using the package brms (Bürkner, 2017). Correlations of random effects estimated from a hierarchical model more accurately reflect participants' "true" effects due to shrinkage from the model and because hierarchical models take trial noise and item variability into account (Chen et al., 2021; Haines et al., 2020; Rouder & Haaf, 2019). We chose this approach as opposed to an intraclass or Pearson's correlation using each participant's mean response speed because this does not take trial or item variability into account (Chen et al., 2021; Rouder & Haaf, 2019). Averaging over trials assumes that all trials and items have the same effect, and we know this is not the case (Alario et al., 2004; Baayen & Milin, 2010). This procedure can therefore underestimate reliability (Chen et al., 2021; Haines et al., 2020; Rouder & Haaf, 2019). We chose to use the Bayesian framework rather than the frequentist framework because Bayesian analyses are better suited to estimating the precision of an effect. We can also obtain correlations of random

[2] As described in the next section, we computed the correlation between the two halves of the data or sessions in a hierarchical model. Due to shrinkage from the model, a few outliers should not influence the correlation very much.

[3] Our ethics board does not allow us to make raw audio recordings of participants publicly available. However, the preprocessed data set includes the all raw output from the experimental software (participant number, trial number, item), as well as the response time and accuracy for each trial.

effects in a frequentist hierarchical model; however, frequentist models only give us a point estimate of the correlation, whereas Bayesian models estimate a distribution of the correlation and a 95% credible interval. This allows us to better characterize the uncertainty of the estimate, which, as we explain in more detail in the next section, is crucial for making decisions about whether a measure is reliable enough for a given purpose. For example, a correlation of .7 would be indicative of good reliability by some standards; however, if the credible interval obtained for that estimate is wide (e.g., [.4,1]), that suggests that the true correlation could potentially be much lower and would no longer be considered to show good reliability.

We followed the procedure detailed in Chen et al. (2021) to fit the following models: To estimate split-half reliability, we fit a no-intercept model that predicts response times with a fixed effect of trial type (even or odd). Instead of estimating an intercept and slope for trial type, this model estimates intercepts for each level of trial type separately. The same structure was reflected in the by-participant random effects: we estimated by-participant intercepts for each level of trial type and random intercepts for item. This means that the correlation of the random by-participant adjustments indexes the correlation between the even and odd trials (i.e., split-half reliability). We repeated this process for the second session in a second model. To estimate test-retest reliability, we fit a third model that is identical to the one described here, except the fixed effect was session (first or second).

We used the following regularizing priors to constrain the model estimates so that extreme values will be unlikely (Schad, Betancourt, & Vasishth, 2021). For the intercepts, we assumed a normal distribution with a mean of 6.75 and a standard deviation of 1.5 on the log scale, which corresponds to a mean of 854 on the millisecond scale. One standard deviation below the mean would be 191ms (exp(6.75)/exp(1.5)), and one standard deviation above the mean would be 3828ms (exp(6.75)*exp(1.5)). For the residual error and the by-subject standard deviation, we assumed a truncated normal distribution with a mean of 0 and standard deviation of 1, and for the correlation between random effects, we used the LKJ prior with parameter $\eta = 2$. Below we report means and 95% credible intervals of the posterior distribution of the correlation between the by-participant random effects.

These analyses will serve as a conceptual replication of previous work that has shown that picture-naming speed is reliable within an experimental session (e.g., Shao et al., 2012 and our pilot data mentioned above), and we expect to replicate this finding. Calculating split-half reliability will help validate the current stimulus set and procedure in order to calculate test-retest reliability. Split-half reliability can additionally be useful in interpreting test-retest reliability because estimates of split-half reliability serve as upper limits to the estimates we can expect to see for test-retest reliability.

As discussed in the introduction, there is no standard threshold that is used to determine whether a measure is reliable or not (e.g., Parsons et al., 2019), likely because what is considered "reliable enough" will depend on the purpose of the measure. It is of theoretical interest to know how reliable word production measures are over time, so to assess this (both for split-half and test-retest reliability), we will use a graded approach to interpreting correlation coefficients. The ranges of correlation coefficients and typical interpretations (Hedge et al., 2018; Landis & Koch, 1977) can be found in Table 2. To determine whether our estimates fall within or above the pre-defined ranges in Table 2, we will use the region of practical equivalence (ROPE) procedure, as explained in Kruschke (2018). The ROPE procedure is a decision-making procedure, in which the researcher defines a range of values (the ROPE) that are "practically equivalent" to a value, such as zero (e.g., in a null-hypothesis significance test). The mean of the posterior distribution and 95% credible interval are computed, and if the credible interval falls completely within the ROPE, we accept that the data are "practically equivalent" to the target value (or range of values); if the credible interval falls completely outside the ROPE, we reject it. For the present purposes, we have defined a ROPE for each of the ranges in Table 2. If the credible interval falls completely within a certain ROPE, we will accept that range of values and interpretation; however, if it spans more than one ROPE, we will only accept that the measure has at least the reliability of the lowest ROPE that the credible interval spans. For example, if our credible interval falls completely within the ROPE for "excellent" reliability, we will accept that the reliability of the measure is excellent. If, however, the posterior mean is .82 but the credible interval is [.75,.89], we would only consider the measure to have "good" reliability because the credible interval overlaps with that ROPE.

Table 2

Ranges of correlation coefficients and their typical interpretations for reliability (e.g., Hedge et al., 2018; Landis & Koch, 1977).

Correlation coefficient	Interpretation
.81-1	Excellent
.61-.8	Good
.41-.6	Moderate
<.4	Poor

We acknowledge that these are arbitrary ranges; however, they can be useful in deciding whether a measure is reliable enough for various purposes. In any case, we encourage readers to examine the posterior distribution means and credible intervals and decide for themselves whether these measures are sufficiently reliable for their purposes.

Results

Split-half reliability

The split-half reliability (i.e., correlation of response times in even and odd trials) in Session 1 was ρ = XX [XX, XX], and we replicate this high correlation in Session 2, ρ =

XX [XX, XX] (see Figure X).

Test-retest reliability

The test-retest reliability for Experiment 1 (i.e., the correlation of response times in Sessions 1 and 2) was ρ = XX [XX, XX] (see Figure X).

Discussion

In Experiment 1, our goal was to replicate findings previously reported in the literature on split-half reliability measures of picture naming, as well as to extend these findings and report test-retest reliability for these measures. Our results first suggest that the reliability of participants' picture naming speed within a session (split-half reliability) is excellent: we got a near perfect correlation between the even and odd trials that was replicated in the second session.

The test-retest reliability of picture naming speed was not quite as high as split-half reliability, but the correlation and credible interval still fell within the good range in our pre-defined ranges. This means that participants are fairly consistent in their picture naming speed even up to two weeks later, at least when performing the same task. This is cause for optimism: Previous studies have correlated measures of cognitive skills with picture naming speed, and the reliability of many of these measures of cognitive skills has already been shown to be relatively high. However, until now we did not know whether picture naming was also a reliable measure (over time) and therefore it was uncertain whether the correlations that had previously been found between cognitive skills and picture naming speed reflected something about individuals or rather just the amount of cognitive effort that was applied to the task in the moment. The fact that we do see fairly good reliability of picture naming speed over time suggests that we may be able to interpret correlations in a more specific way, i.e., that picture naming speed does reflect cognitive abilities of the individual.

However, it is possible that the results of this experiment are limited because the task was identical (except with different items to name) between the two sessions. The participants also had plenty of time (3 seconds total) to name the pictures on each trial. This does not tell us if participants who were slower to name the pictures were slower because they are incapable of naming pictures faster or because they chose a strategy (perhaps to prioritize accuracy) that led them to name the pictures slower. We address this possibility in Experiment 2.

Something about power and how we still might need large sample sizes…

Experiment 2

In Experiment 2, we examine the possibility that participants are engaging in strategies to determine their speed of picture naming. To this end, participants completed a picture-naming task with different instructions. We manipulated task instructions by prompting participants to respond under time pressure in one condition (a speeded condition), and in a non-speeded condition, participants have the same amount of time to respond as in Experiment 1.

Split-half reliability in the non-speeded condition will serve as a replication of Experiment 1, and split-half reliability in the speeded condition can inform the interpretation of the correlation between task manipulations. For instance, if the correlation is low or lower than the correlation between sessions in Experiment 1, split-half reliability estimates of *each* condition can suggest whether the correlation *between* conditions is low due to measurement error (i.e., low split-half reliability in one or both task manipulations) or because participants are not consistent across different manipulations of the task. The correlation of participant speed *between* task conditions can shed light on whether picture-naming speed may be an intrinsic property of participants or whether it is due to participants' use of timing strategies in a given experimental context.

For example, if we see a strong positive correlation between conditions, the strategy explanation would be less plausible because participants would still be ranked similarly by speed even when they do not have enough time to choose their pace.

One obvious strategy that participants could engage in is a speed-accuracy trade-off. To assess this specific possibility, we additionally correlated participants' error rates with speed in each task condition separately to assess (in either condition) whether participants who respond faster are sacrificing accuracy to accomplish this.

Methods

Participants

Fifty participants were recruited from Prolific with the same exclusionary and replacement criteria as in Experiment 1. Participants who participated in Experiment 1 were excluded from participating in Experiment 2.

Stimuli

The same stimuli from Experiment 1 were used for Experiment 2.

Procedure

The experiment was conducted in two separate sessions that took place between 7 and 14 days apart (see Figure 2). Participants completed a simple picture-naming task in both sessions. Participants completed the picture-naming task under two different conditions: speeded and non-speeded. Pilot data from our lab from a similar task suggested that the order of the speeded conditions within a session influences naming speed, in that participants who had the speeded block first were also faster to respond in subsequent blocks even when it was not necessary to. Therefore, participants will receive only one condition (speeded or non-speeded) in a session. The order of presentation of the conditions and the list of stimuli that participants name in a session/condition were be counterbalanced.

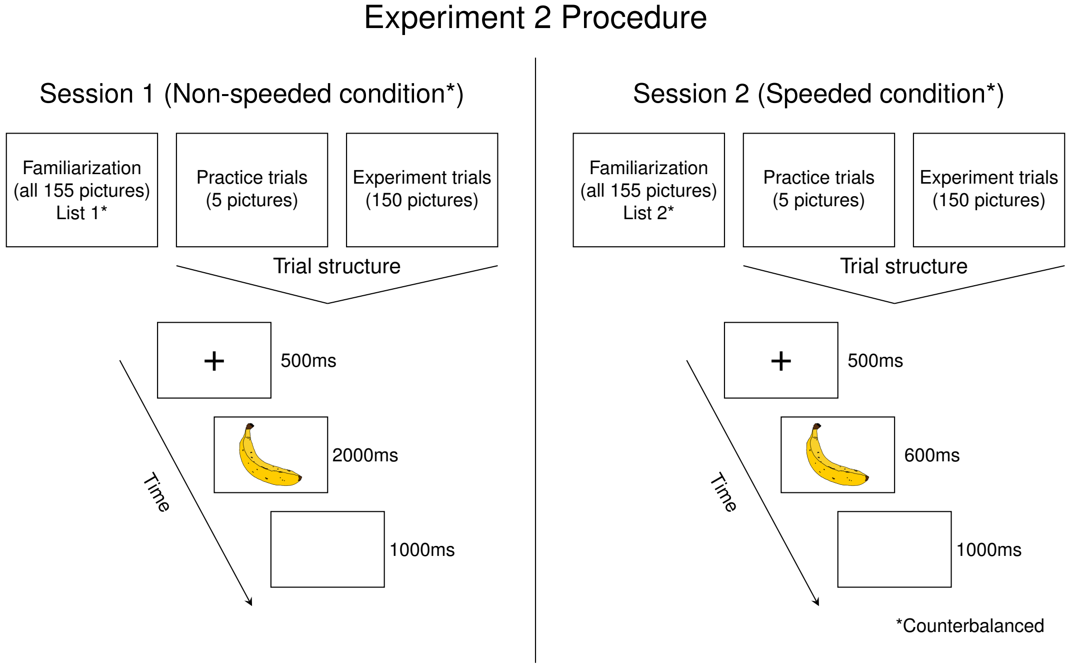
Experiment 2 Procedure
Session 1 (Non-speeded condition*)
Familiarization (all 155 pictures) List 1*
Practice trials (5 pictures)
Experiment trials (150 pictures)
Trial structure
+
500ms
2000ms
1000ms
Time
Session 2 (Speeded condition*)
Familiarization (all 155 pictures) List 2*
Practice trials (5 pictures)
Experiment trials (150 pictures)
Trial structure
+
500ms
600ms
1000ms
Time
*Counterbalanced

Figure 2. Illustration of procedure for Experiment 2. Participants completed the depicted experiment at two different sessions, each with a different stimulus list and condition (speeded or non-speeded). The order of the stimulus list presentation and the session at which participants receive the speeded or non-speeded condition was counterbalanced.

Picture-naming task. The non-speeded condition was identical to the task described in Experiment 1 with the exact same trial structure. In the speeded condition, participants completed a speeded deadline task (e.g., Damian & Dumay, 2007; Gerhand & Barry, 1999; Kello et al., 2000). This task was similar to the non-speeded task, but the duration of the picture presentation was shortened: Each trial began with a fixation cross that appeared in the center of the screen for 500ms, followed by a picture which appeared for 600ms. The picture then disappeared and participants saw a blank screen for 1000ms. Participants were asked to respond before the picture disappeared. As in Experiment 1 and the non-speeded condition, participants first completed a familiarization phase, in which they saw all the pictures they would name for that session presented with their corresponding name. They then completed five practice trials to practice the process, and they named all 150 pictures presented in random order. Vocal responses were recorded from the picture onset until the end of the trial.

Planned analyses

Response time analyses. The data were preprocessed as described in Experiment 1. Incorrect responses were excluded, and response speed was calculated as the time from the stimulus onset to the vocal onset. No outlier trials were removed. For computing split-half reliability and the correlation between the two task conditions, we used the same procedures described above in Experiment 1: The correlation between by-participant random effects was computed in a Bayesian hierarchical model. We used the same priors as in Experiment 1.

These results will first provide a replication of the split-half reliability of Experiment 1, and they will additionally tell us whether participants' word production speed is reliable within a session when participants are under time pressure. The strength of the correlation between the two versions of the task will inform us on the degree to which participants are consistent in their speed across speed conditions.

Speed-accuracy tradeoff. The correlation between participants' speed on each version of the task alone will not be sufficient to tell us whether participants are or are not engaging in strategies or picking a certain speed to name the pictures. To this end, we will correlate participants' speed and accuracy in each version of the task (i.e., we will have one correlation for the speeded version and one correlation for the non-speeded version).

For an estimate of participant speed to enter into these correlations, we extracted by-participant intercepts from the model that estimated the correlation between the two sessions (i.e., each participant had two intercepts). For an estimate of participants' accuracy, we fit a Bayesian hierarchical model with a binomial link that predicts accuracy (0 or 1) and extracted the by-participant random intercepts. Like the response time model, this model included a fixed effect for task condition (speeded or non-speeded) but estimated separate intercepts for the two task conditions (as in the previous models described). We again used regularizing priors. For the intercepts, we assumed a normal prior with mean 0 and standard deviation of 1.5 (on the log odds scale). For the by-subject standard deviation, we assumed a truncated normal distribution with mean 0 and standard deviation of 1, and for the correlation between random effects, we used the LKJ prior with parameter $\eta = 2$.

Correlations between accuracy and speed for each version of the task were computed using the BayesFactor package (Morey & Rouder, 2018). For the prior distribution of the correlation, ρ, we used regularizing priors with a shifted and scaled beta (3,3) distribution (to center the distribution around zero instead of .5, Ly, Verhagen, & Wagenmakers, 2016). This distribution gives more weight to values around zero and downweights extreme values (i.e., -1 or 1). We report the mean of the posterior distribution of ρ and 95% credible intervals.

References

Alario, F.-X., Ferrand, L., Laganaro, M., New, B., Frauenfelder, U. H., & Segui, J. (2004). Predictors of picture naming speed. *Behavior Research Methods, Instruments, & Computers*, *36*(1), 140–155.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Bartolozzi, F., Jongman, S. R., & Meyer, A. S. (2021). Concurrent speech planning does not eliminate repetition priming from spoken words: Evidence from linguistic dual-tasking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(3), 466.

Boersma, P., & Weenink, D. (2021). *Praat: Doing phonetics by computer [computer program]. Version 6.1. 44.*

Borgmann, K. W., Risko, E. F., Stolz, J. A., & Besner, D. (2007). Simon says: Reliability and the role of working memory and attentional control in the simon task. *Psychonomic Bulletin & Review*, *14*(2), 313–319.

Borragan, M., Martin, C. D., De Bruin, A., & Duñabeitia, J. A. (2018). Exploring different types of inhibition during bilingual language production. *Frontiers in Psychology*, *9*, 2256.

Bürki, A. (2017). Electrophysiological characterization of facilitation and interference in the picture-word interference paradigm. *Psychophysiology*, *54*(9), 1370–1392.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Chen, G., Pine, D. S., Brotman, M. A., Smith, A. R., Cox, R. W., & Haller, S. P. (2021). Trial and error: A hierarchical modeling approach to test-retest reliability. *NeuroImage*, *245*, 118647.

Congdon, E., Mumford, J. A., Cohen, J. R., Galvan, A., Canli, T., & Poldrack, R. A. (2012). Measurement and reliability of response inhibition. *Frontiers in Psychology*, *3*, 37.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.

Costa, A., Strijkers, K., Martin, C., & Thierry, G. (2009). The time course of word retrieval revealed by event-related brain potentials during overt speech. *Proceedings of the National Academy of Sciences*, *106*(50), 21442–21446.

Cuetos, F., Ellis, A. W., & Alvarez, B. (1999). Naming times for the snodgrass and vanderwart pictures in spanish. *Behavior Research Methods, Instruments, & Computers*, *31*(4), 650–658.

Damian, M. F., & Dumay, N. (2007). Time pressure and phonological advance planning in spoken production. *Journal of Memory and Language*, *57*(2), 195–209.

Donders, J. (1997). A short form of the WISC–III for clinical use. *Psychological Assessment*, *9*(1), 15.

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six european languages. *Quarterly Journal of Experimental Psychology*, *71*(4), 808–816.

Ellis, A. W., & Morrison, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(2), 515.

Fairs, A., & Strijkers, K. (2021). Can we use the internet to study speech production? Yes we can! Evidence contrasting online versus laboratory naming latencies and errors. *PLoS One*, *16*(10).

Fargier, R., & Laganaro, M. (2019). Interference in speaking while hearing and vice versa. *Scientific Reports*, *9*(1), 1–13.

Fieder, N., Wartenburger, I., & Rahman, R. A. (2019). A close call: Interference from semantic neighbourhood density and similarity in language production. *Memory & Cognition*, *47*(1), 145–168.

Fuhrmeister, P., Madec, S., Lorenz, A., Elbuy, S., & Bürki, A. (2022). Behavioral and EEG evidence for inter-individual variability in late encoding stages of word production. *Language, Cognition, and Neuroscience.*

Gauvin, H. S., Jonen, M. K., Choi, J., McMahon, K., & Zubicaray, G. I. de. (2018). No lexical competition without priming: Evidence from the picture–word interference paradigm. *Quarterly Journal of Experimental Psychology*, *71*(12), 2562–2570.

Gerhand, S., & Barry, C. (1999). Age-of-acquisition and frequency effects in speeded word naming. *Cognition*, *73*(2), B27–B36.

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., … Turner, B. (2020). *Learning from the reliability paradox: How theoretically informed generative models can advance the social, behavioral, and brain sciences.*

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.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 150.

Jhangiani, R. S., Chiang, I., & Price, P. C. (2015). *Research methods in psychology-2nd canadian edition.* BC Campus.

Jongman, S. R. (2017). Sustained attention ability affects simple picture naming. *Collabra: Psychology*, *3*(1).

Jongman, S. R., Meyer, A. S., & Roelofs, A. (2015). The role of sustained attention in the production of conjoined noun phrases: An individual differences study. *PloS One*, *10*(9), e0137557.

Jongman, S. R., Roelofs, A., & Meyer, A. S. (2015). Sustained attention in language production: An individual differences investigation. *Quarterly Journal of Experimental Psychology*, *68*(4), 710–730.

Kello, C. T., Plaut, D. C., & MacWhinney, B. (2000). The task dependence of staged versus cascaded processing: An empirical and computational study of stroop interference in speech perception. *Journal of Experimental Psychology: General*, *129*(3), 340.

Klaus, J., & Schriefers, H. (2018). An investigation of the role of working memory capacity and naming speed in phonological advance planning in language production. *The Mental Lexicon*, *13*(2), 159–185.

Klein, K., & Fiss, W. H. (1999). The reliability and stability of the turner and engle working memory task. *Behavior Research Methods, Instruments, & Computers*, *31*(3), 429–432.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, *44*(4), 978–990.

Laganaro, M., Valente, A., & Perret, C. (2012). Time course of word production in fast and slow speakers: A high density ERP topographic study. *NeuroImage*, *59*(4), 3881–3888.

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

Lorenz, A., Zwitserlood, P., Regel, S., & Abdel Rahman, R. (2019). Age-related effects in compound production: Evidence from a double-object picture naming task. *Quarterly Journal of Experimental Psychology*, *72*(7), 1667–1681.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold jeffreys's default bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32.

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.

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs* (R package version 0.9.12-4.2). https://CRAN.R-project.org/package=BayesFactor.

Nunnally, J. C. (1994). *Psychometric theory 3E*. Tata McGraw-hill education.

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.

Piai, V., & Roelofs, A. (2013). Working memory capacity and dual-task interference in picture naming. *Acta Psychologica*, *142*(3), 332–342.

Pinet, S., Ziegler, J. C., & Alario, F.-X. (2016). Typing is writing: Linguistic properties modulate typing execution. *Psychonomic Bulletin & Review*, *23*(6), 1898–1906.

R Core Team. (2021). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rabovsky, M., Schad, D. J., & Rahman, R. A. (2016). Language production is facilitated by semantic richness but inhibited by semantic density: Evidence from picture naming. *Cognition*, *146*, 240–244.

Roelofs, A. (2004). Seriality of phonological encoding in naming objects and reading their names. *Memory & Cognition*, *32*(2), 212–222.

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

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103.

Shao, Z., Meyer, A. S., & Roelofs, A. (2013). Selective and nonselective inhibition of competitors in picture naming. *Memory & Cognition*, *41*(8), 1200–1211.

Shao, Z., Roelofs, A., Acheson, D. J., & Meyer, A. S. (2014). Electrophysiological evidence that inhibition supports lexical selection in picture naming. *Brain Research*, *1586*, 130–142.

Shao, Z., Roelofs, A., & Meyer, A. S. (2012). Sources of individual differences in the speed of naming objects and actions: The contribution of executive control. *Quarterly Journal of Experimental Psychology*, *65*(10), 1927–1944.

Sikora, K., Roelofs, A., Hermans, D., & Knoors, H. (2016). Executive control in spoken noun-phrase production: Contributions of updating, inhibiting, and shifting. *Quarterly Journal of Experimental Psychology*, *69*(9), 1719–1740.

Snodgrass, J. G., & Yuditsky, T. (1996). Naming times for the snodgrass and vanderwart pictures. *Behavior Research Methods, Instruments, & Computers*, *28*(4), 516–536.

Starreveld, P. A., & La Heij, W. (1999). Word substitution errors in a speeded picture-word task. *The American Journal of Psychology*, *112*(4), 521.

Urbina, S. (2014). *Essentials of psychological testing.* John Wiley & Sons.

Valente, A., Bürki, A., & Laganaro, M. (2014). ERP correlates of word production predictors in picture naming: A trial by trial multiple regression analysis from stimulus onset to response. *Frontiers in Neuroscience*, *8*, 390.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014).

SUBTLEX-UK: A new and improved word frequency database for british english. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Viswesvaran, C., & Ones, D. S. (2000). Measurement error in "big five factors" personality assessment: Reliability generalization across studies and measures. *Educational and Psychological Measurement*, *60*(2), 224–235.

Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 550–564.

Zehr, J., & Schwarz, F. (2018). Penncontroller for internet based experiments (ibex). *URL Https://Doi. Org/10*, *17605*.

Zu Wolfsthurn, S. von G., Robles, L. P., & Schiller, N. O. (2021). Cross-linguistic interference in late language learners: An ERP study. *Brain and Language*, *221*, 104993.

Appendix A

List 1	List 2
anchor	airport
apple	ambulance
arm	ankle
arrow	apron
avocado	asparagus
basket	baby
battery	back
beak	balloon
bear	banana
bell	bat
bench	beard
bib	bed
bike	bedroom
bone	belt
book	bible
bottle	bin
boxer	bomb
bracelet	boomerang
bull	box
bullet	bra
bus	brain
butter	broccoli
butterfly	broom
candle	burger

carrot
cherry
clown
coat
comb
compass
computer
cork
crab
dice
doctor
donkey
doughnut
drawer
dress
drum
ear
elbow
fairy
fireplace
flag
flower
fox
fridge
ghost
glove

butcher
button
cactus
cage
calculator
calendar
camera
cannon
car
caravan
chain
chair
chocolate
choir
cigarette
cloud
coconut
coffee
coffin
condom
cone
cowboy
crown
dentist
desert
devil

heel	diamond
helmet	dinosaur
island	dog
jellyfish	dolphin
kangaroo	dominoes
kite	door
kiwi	dragon
lawnmower	duck
lighter	earring
limo	egg
magnet	elephant
map	eye
medal	face
mermaid	fan
net	farm
peg	feather
plumber	finger
pocket	fire
rabbit	fist
rhino	foot
rocket	football
shield	frog
shower	fruit
tomato	giraffe
tray	girl
lung	glass

castle
hippo
tractor
circle

skirt
mouth
dummy
camel
olive

robot
tie
knife
snake
onion

thumb
cow
squirrel
piano
sun

suit
goat
pipe
lamp
caterpillar

key
suitcase

glasses
greenhouse
guitar
gun

hair
hand
handle
harp
heart

hedgehog
honey
house
iron
jar

keyboard
king
knee
lab
leaf

leg
lemon
lion
lizard
man

mask
maze

owl	microphone
curtain	mirror
saw	moustache
kitchen	needle
flipper	nose
rope	nun
koala	nurse
lighthouse	parachute
wall	parrot
fence	penguin
shoe	pilot
potato	pirate
teacher	pizza
fork	pool
gym	printer
helicopter	pumpkin
hat	ring
wing	road
whale	roof
pear	scarecrow
soap	scarf
mountain	scorpion
zebra	screwdriver
sweet	shadow
sausage	shark
wave	sheep

tattoo	shirt
submarine	sink
triangle	skateboard
tunnel	skeleton
star	skull
tongue	snail
sword	sock
judge	spider
pencil	spoon
orange	stapler
rose	strawberry
nest	swimming
puppet	tank
train	telephone
queen	television
pepper	thermometer
pineapple	tooth
tiger	torch
scissors	trumpet
ruler	umbrella
windmill	vein
wheelbarrow	volcano
tap	waiter
swan	wallet
tambourine	watch
zip	well

tree	witch