Norms for Familiarity, Concreteness, Valence, Arousal, Wordlikeness, and Recall Accuracy for Swahili–Portuguese Word Pairs

Marcos Felipe Rodrigues de Lima¹ and Luciano Grüdtner Buratto¹

Abstract

Normative studies are common in cognitive psychology because they allow us to estimate with more precision the attributes of the stimuli used in empirical studies. The studies reported here had four aims. The first three aims were to obtain estimates for (a) familiarity, concreteness, valence, and arousal for a single set of words in Brazilian Portuguese; (b) wordlikeness (similarity to Portuguese) of a set of foreign words (Swahili); and (c) recall accuracy of Swahili–Portuguese word pairs in a multitrial learning task. The fourth aim was to investigate if any of the assessed measures predicts recall accuracy. One-hundred twenty-eight participants took part in one of the three studies. In Studies 1a and 1b, participants judged 80 Portuguese words for familiarity, concreteness, valence, and arousal and 80 corresponding Swahili words for wordlikeness; in Study 2, participants carried out three study–test cycles of a set of Swahili–Portuguese word pairs. Overall, word-attribute estimates were reliable ($r = .94–.98$) and participants’ responses had high internal consistency (Cronbach’s $\alpha = .84–.98$). Moreover, the relative difficulty of word pairs was retained across trials ($r = .65–.88$). Although different variables correlated with recall accuracy at different time points, multiple regressions indicate that none of the word-attribute variables predicted recall accuracy across trials. These norms may prove fruitful not only for Brazilian human memory researchers but also for international research teams, as it will enable the development of more controlled cross-cultural studies in this field.

Keywords

word pairs, concreteness, emotionality, familiarity, cued-recall

Normative studies are common in cognitive psychology because they allow us to estimate with more precision the attributes of the stimuli used in empirical studies (e.g., Grimaldi et al., 2010; Janczura et al., 2007; Nelson & Dunlosky, 1994). Normed stimuli can, for example, be distributed in different experimental conditions in a balanced manner, increasing internal validity of the experiments (e.g., Pyc & Rawson, 2009, 2010, 2012). Norms also allow us to evaluate how stimulus attributes affect performance (e.g., Witherby & Tauber, 2017). In fact, certain stimulus attributes have a great impact on how likely an item is to be retrieved in a given memory task (Rubin & Friendly, 1986). In free-recall tests, participants recall more concrete words than abstract ones (Witherby & Tauber, 2017), more high-frequency words than low-frequency words (Jia et al., 2016), and more emotional words than neutral words (Johnson & MacKay, 2019).

In recent years, norms for several word attributes, such as emotionality (Kristensen et al., 2011; Oliveira et al., 2013), concreteness (Janczura et al., 2007), and free association (Janczura et al., 2017) have been produced for Brazilian Portuguese. Norms for frequency of occurrence of words in prose texts—an indirect index of objective familiarity—are also available for Brazilian Portuguese (Núcleo Interinstitucional de Linguística Computacional [NILC], 2005). However, norms for familiarity ratings—an indirect index of subjective familiarity—are available only for European Portuguese (Leitão et al., 2010). These two indexes are strongly correlated (Balota et al., 2001), but are based on distinct sources of information (written vs. spoken). Brazilian and European Portuguese have differences at all levels of linguistic structure, including onomasiological variation, which occurs when different terms are used to express the same

¹University of Brasilia, Brazil

Corresponding Author:
Marcos Felipe Rodrigues de Lima, Department of Basic Psychological Processes, Institute of Psychology, University of Brasilia, ICC Sul, sala ASS-012/5, Brasilia DF CEP 70910-900, Brazil.
Email: lima.piraju@gmail.com
referent (Soares da Silva, 2010). Well-known examples of distinct words used to denote the same entities in Brazil and Portugal include telê and ecrã (screen), ônibus and autocarro (bus), and celular and telemóvel (cell phone), respectively. This raises the concern that norms produced from Portuguese samples cannot be used interchangeably with Brazilian samples. In addition, although previous studies have evaluated different word attributes that may be important for memory (e.g., Janczura et al., 2007, 2017; Oliveira et al., 2013), none of them have evaluated familiarity ratings for a set of stimuli in Brazilian Portuguese. As our interest was on several word attributes for a single stimulus database, the first aim of this study was to obtain estimates for familiarity, concreteness, valence, and arousal for a single set of words in Brazilian Portuguese.

The available norms in Brazilian Portuguese allow researchers both to control and to manipulate specific word attributes, in different memory tasks, such as free recall and recognition. Another task commonly used in memory studies is cued recall, in which the experimenter provides cues to the participant at the time of testing (Baddeley et al., 2015). Tasks that use paired associates (e.g., foreign–native word pairs) fall into this category. In a cued-recall task, initially both elements of the pair are presented to the participant (study phase). Next, in the test phase, only the first element of the pair is provided to the participant (i.e., the cue), who must respond with the second (i.e., the target). Some researchers suggest that different encoding strategies may support learning of native–native and foreign–native word pairs (Papagno et al., 1991). When the to-be-learned material is foreign–native word pairs, one factor that can influence their later recall accuracy is wordlikeness, the extent to which a sound sequence is typical in words in the learner’s native language (Gathercole et al., 1991; Papagno et al., 1991). Given this possible influence, the second aim of this study was to estimate the wordlikeness of a set of foreign words.

In addition to estimating different word attributes, researchers have also sought to estimate recall accuracy for word pairs. The most cited of these works, Nelson and Dunlosky (1994) used a multitrial learning task that was also adopted in subsequent normative studies (Bangert & Heydarian, 2017; Cho et al., 2020; Grimaldi et al., 2010). In this kind of task, participants carry out three study–test cycles of a set of word pairs. Each cycle comprised interleaved study and test blocks. In a study trial, participants are asked to learn a series of word pairs. In a test trial, participants are cued with one of the pair elements (typically, the foreign word, but see Bangert & Heydarian, 2017, for an exception), and they are asked to recall the second word element (typically, the native word). In these normative studies, for each word pair, normative recall accuracy is reported as the proportion of participants who correctly recall the target, given the cue, during the test blocks (Trials 1, 2, and 3). Recall accuracy provides a measure of learning difficulty for each pair across study–test cycles.

Normative measures of recall accuracy have been obtained for Swahili–English (Nelson & Dunlosky, 1994), English–Swahili (Bangert & Heydarian, 2017), Lithuanian–English (Grimaldi et al., 2010), and Chinese–English word pairs (Cho et al., 2020). These studies were carried out with native- or proficient-English speakers. However, to date, no study has produced normative measures of recall accuracy for word pairs in which either cue or target or both are in Brazilian Portuguese. Thus, the third aim of this study was to estimate the recall accuracy of Swahili–Portuguese word pairs in a multitrial learning task. The lack of such norms in Brazilian Portuguese reflect the fact that a large amount of human memory research is conducted with participants from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (Henrich et al., 2010). It is thus important to ask which findings generalize to participants from non-WEIRD societies (Roediger & Yamashiro, 2018) and which findings differ across cultures (see, e.g., Papagno et al., 1991). In addition, some scholars in psychology have recently recommended an emphasis on close replications of research findings (e.g., LeBel & Peters, 2011). We believe that a database normed with participants from a non-WEIRD society makes an important contribution to human memory research, because it allows both the replication of previous findings and the introduction of standardized verbal stimuli for cross-cultural memory studies.

Swahili is a language spoken in East African countries, and its use in memory studies is based on a series of arguments posed by Nelson and Dunlosky (1994) that support Swahili’s suitability as a potential source of stimuli. Following Nelson and Dunlosky’s reasoning, we chose Swahili because (a) native speakers of Brazilian Portuguese are unlikely to have been exposed to words in Swahili. This ensures that learners know little about the to-be-learned material (Bjork & Kroll, 2015); (b) like Brazilian Portuguese, Swahili’s writing is based on the Latin alphabet. Thus, in an experimental task, the learner is not burdened with the additional demand of having to learn new symbols from the foreign language; and (c) Swahili words are unlikely to produce floor effects on memory tasks, allowing additional learning in the multitrial learning task (for similar arguments, see Nelson and Dunlosky, 1994).

In previous studies, researchers have addressed how word attributes relate to later recall accuracy (Bangert & Heydarian, 2017; Cho et al., 2020; Grimaldi et al., 2010; Nelson & Dunlosky, 1994). For example, Nelson and Dunlosky and Grimaldi et al. found small, but significant correlations between frequency of occurrence of English words and recall accuracy on Trial 1, $r_s = .25$, $p < .05$. In contrast, in Nelson and Dunlosky’s study, neither wordlikeness of Swahili words nor concreteness of English words reliably correlated with recall accuracy on Trial 1. Here, we also addressed possible relationships between word attributes and recall accuracy. Thus, the fourth aim of this study was to investigate if any of the assessed measures predicts recall accuracy.
Table 1. Studies and Samples Characteristics.

| Study (local) | Estimated attributes          | N     | Sample characteristics                  |
|--------------|-------------------------------|-------|-----------------------------------------|
| 1a (Lab)     | Familiarity, concreteness, valence, arousal, and wordlikeness | 18 Females = 8, Males = 10, $M_{age} = 19.06$ years ($SD = 0.94$), age range = 18–21 |
| 1b (Online)  | Familiarity, concreteness, valence, arousal, and wordlikeness | 74 Females = 46, Males = 27, $M_{age} = 24.18$ years ($SD = 5.63$), age range = 18–49 |
| 2 (Lab)      | Recall accuracy               | 36 Females = 20, Males = 16, $M_{age} = 20.58$ years ($SD = 2.78$), age range = 18–28 |

Note. Studies 1a and 2 were run in 2018; Study 1b was run in 2020.
*aOne participant did not inform their sex.

Method

Overview

Table 1 summarizes study and sample characteristics. In three studies, we collected norms for familiarity, concreteness, valence, arousal, wordlikeness (Studies 1a and 1b), and recall accuracy (Study 2). Studies 1a and 2 were run in 2018. To improve our estimates for familiarity, concreteness, valence, arousal, and wordlikeness, we ran an online study with more participants in 2020 (Study 1b).

Participants

One-hundred twenty-eight participants comprised the final sample. Participants from Studies 1a ($n = 18$) and 2 ($n = 36$) were recruited at the University of Brasilia. Seven additional participants were excluded after initial screening analyses: (a) five participants showed signs of depression or anxiety; (b) one participant chose the same response on all judgments, indicating lack of compliance; and (c) one participant in Study 2 had extreme scores in the multitrial learning task, defined as a $z$-score $>|\pm 2.50|$ (see Hair et al., 2014, p. 65).

Data collection from Studies 1a and 2 was conducted simultaneously and each participant took part in only one of these studies. Participants from Study 1b ($n = 74$) were recruited from advertisements posted on social media groups and messaging apps, mainly focused on higher education topics. Participants gave informed consent before starting the tasks. Research was approved by the Research Ethics Committee before data collection.

Instruments

Word pairs. Eighty Swahili–Portuguese word pairs were selected and adapted from Nelson and Dunlosky’s (1994) norms. In addition, six pairs were used both as examples during instructions (Studies 1a and 1b) or as filler items (Study 2). Word pairs were presented in white color centered in a black screen (lowercase, boldface Arial 18 font). When both Swahili and Portuguese words were simultaneously presented (only in Study 2), the Swahili word was always presented on top. In Studies 1a and 2, stimuli were presented on a computer screen and stimulus presentation was controlled with PsychoPy (Peirce, 2007). In Study 1b, we made minor changes in the task, to make it compatible with Pavlovia (https://pavlovia.org/), where the study was hosted. We mention these minor changes in the next section.

Procedure

In both lab studies (Studies 1a and 2), participants were tested individually in a single session. In the online study (Study 1b), the instructions encouraged participants to avoid noisy environments or other sources of distraction (e.g., social networks) while performing tasks. In the lab studies, participants answered the Beck Depression Inventory (Cunha, 2001) and the State–Trace Anxiety Inventory (Biaggio & Natalício, 1979) and then proceeded to the main task. These inventories were used as screening tools to exclude from our final sample participants with serious anxiety and depression symptoms, as they could inadvertently bias the results (e.g., they could bias valence and arousal estimates and memory performance). Out of the five participants excluded via inventories, two were excluded because they showed strong signs of depression (scores 32 [moderate] and 53 [severe] on the Beck Depression Inventory, which ranges from 0 to 63) and three were excluded because they showed strong signs of anxiety (scores 38 and 43 [state anxiety] and 40 [trait anxiety] in the State–Trace Anxiety inventory, which ranges from 20 to 80, with lower scores representing higher levels of state/trait anxiety).

Studies 1a and 1b. Figure 1 shows a schematic representation of the judgment tasks. The label of the word attribute and the word to be judged were presented at the top and center of the screen, respectively. A scale was presented at the bottom along with labels identifying the extreme values. Studies 1a and 1b were divided into two sets of judgments. At the beginning of each set, participants (a) were provided with the meaning of each one of the to-be-judged attribute (i.e., first set: familiarity, concreteness, valence, and arousal; second set: wordlikeness), (b) practiced the tasks by judging a non-normed word, and (c) had the opportunity to ask any questions about the tasks. Concerning this last point, in Study 1b, the researcher’s (M.F.R.L.) cell phone number and e-mail
address were made available both in the research advertisements and in one of the first instruction screens. Although some participants contacted us for other reasons (e.g., requesting a written statement of participation), none of them contacted us to ask questions about the tasks.

In the first set, participants made judgments of familiarity, concreteness, valence, and arousal for each Brazilian Portuguese word, in this fixed order of judgments, before seeing another word. Presentation order of 80 words was randomized anew for each participant. We chose to begin trials with a familiarity judgment to avoid familiarity overestimation due to prolonged exposure to the word. In the second set, participants made judgments about the wordlikeness of 80 Swahili words. In both sets, there was no time limit on participants’ responses, although they were instructed to work fast. In Study 1a, at the midpoint of each scale, there was a red circle that could be moved either to the left or to the right. Participants should move the red circle to the point in the scale that best represented their response and press the “Enter” key to confirm it. In Study 1b, this red-moving circle was omitted and participants should press the number that best represented their response, without the need to press any additional key. These two minor changes were necessary because, at the time of data collection, the RatingScale Component from PsychoPy was incompatible with Pavlovia’s host service. Thus, we used pictures mimicking the scales, instead of a “real” visual analog scale. In both sets of judgments, no feedback was provided to participants. The estimated time for completion of all tasks was 40 min.

Familiarity ratings ranged from 1 (I never saw/heard that word) to 7 (I see/hear that word almost daily) corresponding to how unfamiliar or familiar they considered each word. Concreteness ratings ranged from 1 (highly abstract) to 7 (highly concrete) corresponding to how abstract or concrete they considered each word. Valence was assessed with the Self-Assessment Manikin scale (SAM; Kristensen et al., 2011) ranging from 1 (negative emotional valence) to 9 (positive emotional valence) corresponding to how unpleasant or pleasant they considered each word. Arousal was also assessed with the SAM scale, ranging from 1 (relaxing) to 9 (exciting) corresponding to how relaxed or aroused they considered each word. Wordlikeness ranged from 1 (Not like a word at all) to 5 (Very like a word) corresponding to how similar a Swahili word is from any Brazilian Portuguese word.

Study 2. Figure 2 shows a schematic representation of one cycle (out of three) of the multitrial learning task. The 80 Swahili–Portuguese word pairs were divided into two lists, each with 40 pairs. The six filler pairs were included in both lists. Thus, each list comprised 46 word pairs, and each participant was exposed to only one list. In each list, the word pairs were divided into three sets, the first with the six filler pairs and the other two with twenty pairs each. Filler items were added to control for possible primacy effects (i.e.,
better recall for items at the beginning of the study list). Word pairs were split in sets of twenty items to control for lag effects (i.e., better recall for items presented closer in time). We made sure that there were at least 26 word pairs between study and test of a given word pair. Within each set, presentation occurred in random order. From the participants’ point of view, only one set was studied, as there was no indication of the end of one set and the beginning of the other set.

Participants carried out three study–test cycles without feedback. Each cycle comprised interleaved study blocks (Figure 2A) and test blocks (Figure 2B), with a brief instruction (e.g., Study Block: Try to learn the association between Swahili and Brazilian Portuguese words) presented at the beginning of each block. In a study trial, participants saw a word pair (e.g., sahani–prato, both words meaning plate) for 10 s and were asked to learn the association between that pair. In a test trial, participants saw only the Swahili word (e.g., sahani) and were asked to recall its meaning in Brazilian Portuguese by typing the corresponding word on a keyboard. A test trial ended either after participants pressed the “Enter” key or after 10 s, regardless of whether a response was given or not (see Figure 2B). On both study and test trials, intertrial interval (ITI) lasted 1 s. The estimated time for completion of all tasks was 60 min.

### Statistical Analyses

#### Studies 1a and 1b

Means and standard deviations were computed for each word and for each word attribute, separately for Studies 1a and 1b. To check the reliability of these judgments, we carried out two analyses. First, we correlated the means for each word attribute and for each word across cohorts (i.e., Study 1a vs. Study 1b). To foreshadow, as we found strong correlations across cohorts in all word attributes, we collapsed data from Studies 1a and 1b in the following analyses. Second, we used the collapsed database and computed Cronbach’s alphas (one for each word attribute), which assess the internal consistency of ratings across participants (see Hair et al., 2014).

#### Study 2

Two judges independently rated participants’ answers. As the focus of these norms was on recall accuracy (i.e., vocabulary learning across cycles), typing and spelling errors were not counted as errors. Answers of eight participants (22.2% of sample) were rated by both judges. We computed Cohen’s kappa, a chance-corrected measure of agreement between two judges for categorical data, which provides an index of how reliable the scores made by the judges were (Howell, 2013). Judges showed a high level of agreement, $\kappa = .98, p < .001$, which can be considered “almost perfect,” according to Landis and Koch’s (1977) benchmarks. The following analyses were carried out on the recall accuracy (i.e., the proportion of participants who correctly recalled a given Portuguese word) on Trials 1, 2, and 3.

### Studies 1a, 1b, and 2

We ran a series of Pearson’s correlations to assess relationships among word-attribute variables and recall accuracy across trials. Next, we ran three multiple linear regression models to investigate if any measure predicts recall accuracy on Trials 1, 2, and 3. Familiarity, concreteness, valence, arousal, wordlikeness, log frequency of occurrence (from NILC, 2005), and word length (Swahili
and Portuguese) were entered into each model as predictors. Although several word attributes have been shown to affect free recall (Paivio, 1968; Rubin & Friendly, 1986), less is known about their contributions in cued-recall tasks. As we did not have strong a priori hypothesis regarding which word attributes play an important role in predicting (cued-)recall accuracy, we entered all predictors simultaneously into the regression models.

Results and Discussion

To remember, our three studies had four aims: (a) to obtain estimates for familiarity, concreteness, valence, and arousal for a single set of words in Brazilian Portuguese; (b) to estimate wordlikeness of a set of foreign (Swahili) words; (c) to estimate the recall accuracy of Swahili–Portuguese word pairs in a multitrial learning task; and (d) to investigate if any measure predicts the recall accuracy on Trials 1, 2, and 3. In the following four sections, we present our results. First, we present word-attribute estimates, as well as the relationships among them (Study 1a and 1b; aims 1 and 2). Second, we present recall-accuracy estimates in the multitrial learning task (Study 2; aim 3). Third, we present results of cross-studies regression analyses, assessing if any of the word attributes predict recall accuracy (aim 4). Finally, to provide converging evidence of the reliability of the normative database presented here, we briefly describe three retrieval practice experiments reported elsewhere (Lage, 2019; Lima et al., 2020), in which a sample of 40 Swahili–Portuguese word pairs from the present norms were used.

The normative data for all 80 Swahili–Portuguese word pairs can be found on the Open Science Framework website (https://osf.io/ucx7h/).

Studies 1a and 1b

Figure 3 depicts scatterplots showing the relationship between the word-attribute estimates for Studies 1a and 1b. All panels of Figure 3 show strong positive correlations across studies ($r = .94–.98$), which suggests that the norms for word attributes are reliable. Thus, in the following analyses, we collapsed data from Studies 1a and 1b.

Table 2 summarizes descriptive statistics and internal consistencies for all word attributes. We highlight some results showed in Table 2. First, as can be seen in the last column of Table 2, we found acceptable values for internal consistency in participants’ judgments of word attributes (Cronbach’s $\alpha = .84–.98$; see Hair et al., 2014). Second, the word attribute familiarity had the lowest variability and the greater minimum estimate, which suggests that the 80 normed, Brazilian Portuguese words tend to be judged as having medium-to-high familiarity—a pattern also depicted in Figure 3, panel A. As we chose to use words normed for recall accuracy by Nelson and Dunlosky (1994), an inevitable risk of this methodological decision is the lack of guarantee that the normed words would reflect the full range of familiarity—as well as of other word attributes—scale, which seems to have been the case. Nevertheless, some interesting results were found using familiarity estimates (described in the next sections). Unlike familiarity estimates, the other word attributes seemed to vary over a greater range of scales used (see also Figure 3, panels B–E).
Third, like familiarity, concreteness judgments tended to have values above the average (see Table 2); however, unlike familiarity, some words represented the lower limit of the concreteness scale (e.g., alma [soul] and mistério [mystery]; see also Figure 3B). Concreteness is a word attribute which is related to the extent that a word brings to one’s mind a sensory experience. Whereas alma and mistério are abstract ideas, without a specific or immediate sensory match in the world, words like tomate (tomato) and prato (plate) have referents in the world that can be experienced from the senses. Thus, for concreteness estimates to be deemed reliable, they would need to reflect this attribute. Average concreteness for tomate, prato, mistério, and alma are 6.92, 6.85, 2.89, and 2.52, respectively. Thus, participants’ responses seem to reflect this word attribute. Fourth, although valence and arousal have approximately similar averages, their estimates appear to be skewed in opposite directions (see “Skew,” in Table 2), with more word positive than negative valence words and more relaxing than exciting words, respectively (see also Figure 3, panels C–D).

Finally, Swahili words are mainly assessed as having low wordlikeness (i.e., below the midpoint of scale, 3), although words ranged across the whole scale. An interesting result, not shown in Table 2, wordlikeness estimates appear to have face validity, as the Swahili words chama and pipa, two Swahili words that are homonymous to words in Portuguese, were the only ones which approached ceiling (Ms = 4.86 and 4.91, respectively). Again, the inclusion of these words in the present norms is based on a selection of words from Nelson and Dunlosky’s (1994) norms, and it was not intended to produce validity evidence of wordlikeness estimates—but, in hindsight, it has served this purpose. In contrast, Swahili words ikitisadi and mbwa had the lowest wordlikeness estimates (Ms = 1.26 and 1.30, respectively). These words have consonant clusters that do not occur in Brazilian Portuguese. Thus, the wordlikeness estimates across words seem to reflect the extent to which Swahili words sound like Brazilian Portuguese words.

The first eight rows of Table 3 shows correlations among different word attributes. A series of correlations provide further support for the reliability of word-attribute estimates. First, familiarity estimates correlated positively with log frequency of occurrence per million words (NILC, 2005), a related construct. This result replicates Balota et al. (2001), who also found significant correlations between subjective and objective English word-frequency indices. Second, familiarity estimates significantly correlated with valence and arousal, but not with concreteness. A similar pattern was reported by Paivio (1968), although he measured only emotionality rather than valence and arousal (familiarity and emotionality, \( r = -.25 \); familiarity and concreteness, \( r = .01 \); see Paivio’s Table 1). Third, concreteness estimates significantly correlated with arousal, but not with valence. Paivio also found a significant correlation between concreteness and emotionality (\( r = -.54 \); see Paivio’s Table 1). These results suggest that arousal estimates are more similar to the emotionality estimates measured in Paivio’s study. Fourth, the skewness of the valence and arousal estimates, which ran in opposite directions, was translated into a strong negative correlation between these variables, a result similar to the visual pattern depicted by Oliveira et al. (2013). This result indicates that words that have a higher valence (i.e., positive) tend to be more relaxing (e.g., colchão [mattress]), whereas words with a lower valence (i.e., negative) tend to be more arousing (e.g., cadáver [corps]). Nonetheless, there were exceptions, such as ciência (science) that despite having a positive valence is also considered an arousing word (participants were sampled from a university setting, so this effect can be population-specific.) There were no cases showing the opposite pattern (i.e., negatively valenced, but relaxing, words).

In sum, a series of analysis suggest that the word attributes collapsed across Studies 1a and 1b are reliable.

### Study 2

A repeated-measures analysis of variance (ANOVA) was carried out on participants’ average performance. Because the sphericity assumption was violated, \( W = .55, \chi^2 (2) = 20.12, p < .001 \), we applied the Greenhouse–Geisser correction to adjust the degrees of freedom of the ANOVA. The ANOVA showed that participants’ average performance significantly increased across test trials, \( M_{\text{Trial}1} = .16 (SD = .13) \), \( M_{\text{Trial}2} = .43 (SD = .26) \), \( M_{\text{Trial}3} = .62 (SD = .27) \),
The significant correlations with familiarity and recall accuracy on Trials 2 and 3—but not on Trial 1—leave open the possibility that different word attributes may play a role in recall accuracy at different time points.

Second, unlike Nelson and Dunlosky (1994), we found a significant correlation between recall accuracy on Trial 1 and wordlikeness (see Table 3). We noticed, however, that our mean wordlikeness estimate had greater variability than that by Nelson and Dunlosky’s median wordlikeness, which may partially explain their null results. We found a similar result when we correlated our median wordlikeness and recall accuracy on Trial 1, $r = .31$, $p = .006$. However, our median wordlikeness estimate had an even greater variability than our mean wordlikeness estimate ($SD_s = 1.39$ vs. 0.95), contrasting with a lower variability from median wordlikeness estimates from Nelson and Dunlosky (1994; SD = 0.74 and 0.72, based on 100 and 80 words, respectively). Alternatively, mediator-based accounts of retrieval practice effects (Pyc & Rawson, 2010) posit that participant-generated keywords that look or sound similar to the foreign-word cue and that are semantically related to the native-word target can enhance later memory for that foreign–native word pair. In some textbooks (e.g., Baddeley et al., 2015, p. 485), the word pair *wingu-cloud* is typically used to highlight the role of mediators: The word *wingu* look like to *wingu* and is semantically related to *cloud*, thus possibly linking *wingu to cloud*. However, in Brazilian Portuguese, this same associative chain (*wingu* → *asa* → *nuvem*) is meaningless, because the word *asa* is semantically related to *nuvem*, but it is neither orthographically nor phonologically related to *wingu*. A better example in Brazilian Portuguese could be *sahani* → *salame* → *prato*, in which the word *salame* (*salami*) sounds like *sahani* and can be semantically related to *prato* (*plate*). A decade ago, Pyc and Rawson (2010) claimed that one of the key factors for mediator effectiveness is mediator retrievability, defined as how likely a word is to be recalled after prompted with the cue. Indeed, a cue’s wordlikeness may support phonological associations

Table 3. Correlation Matrix for All Measures.

| Measure | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 |
|---------|----|----|----|----|----|----|----|----|----|----|----|
| 1. Familiarity | 1  |    |    |    |    |    |    |    |    |    |    |
| 2. Concreteness | .05 | 1  |    |    |    |    |    |    |    |    |    |
| 3. Valence | .42** | .16 | 1  |    |    |    |    |    |    |    |    |
| 4. Arousal | −.26* | −.35** | −.86** | 1  |    |    |    |    |    |    |    |
| 5. Wordlikeness | −.006 | .04 | −.05 | −.004 | 1  |    |    |    |    |    |    |
| 6. Frequency of occurrence | .60** | −.29*** | .27* | −.008 | −.10 | 1  |    |    |    |    |    |
| 7. Swahili word length | −.02 | −.13 | .07 | −.04 | −.33*** | .17 | 1  |    |    |    |    |
| 8. Portuguese word length | −.30*** | −.16 | −.27* | .35*** | .002 | −.25* | .14 | 1  |    |    |    |
| 9. Recall accuracy, Trial 1 | .11 | .06 | .22 | −.14 | .28* | −.009 | −.16 | −.09 | 1  |    |    |
| 10. Recall accuracy, Trial 2 | .25* | .08 | .25* | −.16 | .15 | .12 | −.22* | −.19 | .78** | 1  |    |
| 11. Recall accuracy, Trial 3 | .28* | .13 | .23* | −.15 | .13 | .16 | −.26* | −.18 | .65** | .88** | 1  |

Note. Measure 6 is based on log frequency of occurrence per million words (NILC, 2005). Word length refers to number of letters for each word. *p < .05. **p < .01. $F(1.38, 48.39) = 199.85, p < .001, \eta^2 = .85$. Recall accuracy across word pairs and across trials had a wide range of difficulty, $range_{trial1} = .00–.71, range_{trial2} = .05–.95, range_{trial3} = .26–1$, allowing researchers to choose word pairs that lead to different learning rates. These analyses support Swahili’s suitability as a potential source of stimuli for human memory research (see Nelson & Dunlosky, 1994). Finally, mirroring previous normative studies, we correlated recall accuracy for each word pair across trials. As Table 3 shows, recall accuracy on Trials 1, 2, and 3 had strongly significant correlations ($r_s = .65–.88$), thus replicating previous normative studies which demonstrates that the relative difficulty of word pairs was retained across trials (Bangert & Heydarian, 2017; Cho et al., 2020; Grimaldi et al., 2010; Nelson & Dunlosky, 1994).
between cue and mediators, which may increase the likelihood of target retrieval. This alternative hypothesis for the different results obtained here and by Nelson and Dunlosky states that Swahili words, at least those ones used in both normative studies, are more likely to look or sound like Brazilian Portuguese words (vs. English words) that can be easily linked to Swahili translations (in Brazilian Portuguese or in English). Although somewhat speculative, this hypothesis has its precedents in the short-term memory literature (see, e.g., Papagno et al., 1991, p. 339) and it can be empirically tested in future cross-cultural memory studies.

Third, as can be seen in Table 3, Swahili word length negatively correlated with recall accuracy on Trials 2 and 3. However, Table 3 also shows that wordlikeness was negatively correlated with Swahili word length. The latter result is consistent with findings from developmental studies which showed that children’s performance on a nonword repetition task may be independently influenced by mnemonic (i.e., nonword length) and linguistic (i.e., wordlikeness) factors (e.g., Gathercole et al., 1991). Next, we further explored whether the relationship between Swahili word length and recall accuracy on Trials 2 and 3—as well as the relationship between wordlikeness and recall accuracy on Trial 1, showed in the previous paragraph—remain significant after controlling for the relationship between wordlikeness and Swahili word length. These exploratory analyses showed (a) that wordlikeness and recall accuracy on Trial 1 remained significantly correlated after controlling for the effect of Swahili word length on wordlikeness, $r_{\text{semipartial}} = .24, p = .03$; and (b) that Swahili word length significantly correlated with recall accuracy on Trial 3 ($r_{\text{semipartial}} = -.24, p = .04$), but not on Trial 2 ($r_{\text{semipartial}} = -.18, p = .11$), after controlling for the effect of wordlikeness on Swahili word length. These semipartial correlations suggest that wordlikeness and Swahili word length shared variance with recall accuracy (on Trials 1 and 3, respectively) that is independent from the variance shared between themselves (wordlikeness and Swahili word length).

Table 4 shows model summaries for simultaneous multiple regression analyses for predicting recall accuracy on Trials 1, 2, and 3. As is shown in Table 4, all models approached, but did not reach, statistical significance ($p = .052–.095$). As previously described, it should be noted that the valence and arousal estimates were strongly correlated (see Table 3). To minimize multicollinearity concerns, alternative models were tested. In these alternative models, only one of these two measures was entered at a time. All analyses indicated the same pattern of results (i.e., all models remained nonsignificant), with the exception that the Swahili word length was a significant predictor of recall accuracy on Trial 3, but only when arousal was not entered into the model. As can be seen in Table 4, all unstandardized coefficients had values that approached zero ($|b| \leq 0.10$). The same is true for alternative models not reported in Table 4. Thus, we avoid making strong claims of predictive relationships based on our simultaneous multiple regression analyses. To sum up, although different variables correlated with recall accuracy at different time points (Table 3), multiple regressions indicate that none of the word-attribute variables predicted recall accuracy across trials. It is possible that word attributes that predict recall on free-recall tasks, such as imagery, availability, and emotionality (Rubin & Friendly, 1986), may not be so important in foreign–native word learning, as in multitrial learning task. By nature, word-attribute estimates from Studies 1a and 1b were made at the word level, not at the word pair level. However, the task that participants were asked to perform in Study 2 involves learning the association between an unknown Swahili word and a probably known Brazilian Portuguese word. We raise the possibility that variables that can make this foreign–native association easier or more difficult may be potentially useful word-pair attributes that predict recall at least at the beginning of the task. Recently, Dikmans et al. (2020) found that mediated retrieval diminishes after frequent retrieval attempts, which suggests that even word-pair attributes can have their effects on cued recall at different points in time. Future studies may benefit from asking whether attributes measured at the word pair level could predict recall accuracy in cued-recall tasks.

**Word Pair Analyses in Retrieval Practice Experiments**

A sample of 40 Swahili–Portuguese word pairs from the present norms was used in retrieval practice experiments reported elsewhere (Lage, 2019, $N = 59$; Lima et al., 2020, Experiment 1, $N = 51$, and Experiment 2, $N = 28$). In these experiments, 20 pairs were easy and 20 were difficult, based on the average recall accuracy on Trials 1, 2, and 3 (computed from the spreadsheet available in the Supplemental Material). These word pairs were split into two sets, matched for familiarity, concreteness, valence, arousal, wordlikeness, and average recall accuracy, and they were later assigned to one of the two conditions, nominally defined by the type of practice that participants engaged during experiments: Participants studied all word pairs, then repeatedly restudied half of them and retrieval practiced another half, according to word pair assigned condition. Two days (Lima et al., 2020, Experiments 1 and 2) or seven days (Lage, 2019) later, all participants took a final cued-recall test.

Of particular interest here, all three experiments found significant effects of difficulty, so that more easy items were recalled than difficult ones (Lage, 2019; Lima et al., 2020). These “easy” and “difficult” items were dichotomized variables. In addition, Lima et al. reported the proportion of participants who correctly recalled each Brazilian Portuguese word on final cued-recall tests—similar to recall accuracy here reported, except that this Lima et al.’s measurement occurred after a 2-day retention interval. Lima et al. found strong item-wise correlations between their 2-day recall accuracy and our average recall accuracy across Trials 1, 2,
and 3, in both Experiment 1, \( r = .88 \), and Experiment 2, \( r = .72 \), \( p < .001 \), providing converging evidence of the adequacy of their difficulty manipulation. Lage (2019) did not present this item-wise correlation, but we reanalyzed data from three experiments reported in both Lage (2019) and Lima et al. (2020).

The top half of Table 5 shows item-wise correlations among our word attributes and the 2-day and the 7-day recall accuracies from Lage (2019) and Lima et al. (2020). The top half of Table 5 shows that, with a data set restricted to 40 word pairs, four word attributes correlated with recall accuracy. First, wordlikeness correlated with recall accuracy on Trial 1 (as already showed earlier in Table 3). Second, familiarity correlated with recall accuracy on Trials 2 and 3 of the present study, also replicating earlier correlational analyses (see Table 3); more important, Table 5 shows that familiarity also correlated with recall accuracy from Lima et al.’s (2020) Experiments 1 and 2, with the magnitude of correlations remaining approximately at the same levels as they were in a single session (see Table 3). Although these results should be interpreted with caution, as they came from studies with different designs, they seem to suggest that the influence of familiarity remains considerably stable from Trial 2 to a 2-day retention interval.

Third, providing converging evidence for the importance of word frequency (either subjective or objective), log frequency of occurrence (NILC, 2005) also correlated with 2-day recall accuracy from Lima et al.’s (2020) Experiments 1 and 2. Fourth, valence significantly correlated with 2-day recall accuracy from Lima et al.’s (2020) Experiment 1. Finally, we should note that no word attribute significantly correlated with Lage’s (2019) recall accuracy. This is not to say, however, that none of them is related to recall accuracy after retention intervals longer than 2 days. Participants from Lage (2019) had a lower average recall (\( M = .24, SD = .20 \)) than both Experiments 1 (\( M = .41, SD = .19 \)) and 2 (\( M = .43, SD = .20 \)) from Lima et al. (2020). A median-split suggest that 50% of word pairs from Lage (2019) were recalled by a proportion less than .16 of the participants, which may partially explain those null results.

The bottom half of Table 5 shows correlations among recall accuracy across studies. All correlations among recall accuracies were significant and ranged from .58 to .98. Taken together, these correlations further suggest that the present

| Table 4. Summary of Simultaneous Multiple Regressions for Predicting Recall Accuracy on Trials 1, 2, and 3. |
| --- |
| **Criterion variable and predictor variables** | **b** | **SE (b)** | **t** | **p** |
| **Criterion variable: Recall accuracy, Trial 1 - \( F(8, 71) = 1.88 \), \( p = .078 \), \( R^2 = .17 \), \( \Delta R^2 = .081 \)** | | | | |
| Familiarity | 0.010 | 0.033 | <1[1] | .76 |
| Concreteness | 0.004 | 0.015 | <1[1] | .77 |
| Valence | 0.045 | 0.019 | 2.35 | .022 |
| Arousal | 0.05 | 0.03 | 1.66 | .10 |
| Wordlikeness | 0.043 | 0.018 | 2.45 | .017 |
| Frequency of occurrence | -0.036 | 0.033 | -1.07 | .29 |
| Swahili word length | -0.005 | 0.015 | <1[1] | .74 |
| Portuguese word length | -0.009 | 0.011 | <1[1] | .44 |
| **Criterion variable: Recall accuracy, Trial 2 - \( F(8, 71) = 1.78 \), \( p = .095 \), \( R^2 = .17 \), \( \Delta R^2 = .073 \)** | | | | |
| Familiarity | 0.04 | 0.044 | <1[1] | .36 |
| Concreteness | 0.007 | 0.02 | <1[1] | .72 |
| Valence | 0.045 | 0.026 | 1.76 | .08 |
| Arousal | 0.048 | 0.04 | 1.18 | .24 |
| Wordlikeness | 0.022 | 0.024 | <1[1] | .35 |
| Frequency of occurrence | -0.017 | 0.045 | <1[1] | .71 |
| Swahili word length | -0.027 | 0.02 | -1.36 | .18 |
| Portuguese word length | -0.014 | 0.015 | <1[1] | .35 |
| **Criterion variable: Recall accuracy, Trial 3 - \( F(8, 71) = 2.05 \), \( p = .052 \), \( R^2 = .19 \), \( \Delta R^2 = .096 \)** | | | | |
| Familiarity | 0.039 | 0.04 | <1[1] | .34 |
| Concreteness | 0.019 | 0.019 | <1[1] | .32 |
| Valence | 0.031 | 0.024 | 1.30 | .20 |
| Arousal | 0.032 | 0.037 | <1[1] | .39 |
| Wordlikeness | 0.013 | 0.022 | <1[1] | .56 |
| Frequency of occurrence | 0.013 | 0.041 | <1[1] | .75 |
| Swahili word length | -0.037 | 0.019 | -1.96 | .054 |
| Portuguese word length | -0.007 | 0.014 | <1[1] | .63 |

Note. \( b = \) unstandardized coefficient.
recall accuracy estimates are reliable. Importantly, the 7-day recall accuracy from Lage (2019) correlated with the other recall accuracies, suggesting that, although Lage’s (2019) participants showed a trend to floor effects, it was still possible to demonstrate that the difficulty of word pairs was relatively retained at least for 7 days.

Concluding Remarks

A number of limitations of the present study should be noted. First, compared to previous recall accuracy normative studies (Bangert & Heydarian, 2017; Cho et al., 2020; Grimaldi et al., 2010; Nelson & Dunlosky, 1994), our Study 2’s sample size was rather unusual. However, the strong positive correlations between word-attribute estimates for Studies 1a and 1b (see Figure 3) seems to point to the idea that even smaller samples are capable of producing reliable and stable estimates of word attributes. Admittedly, this is a demonstration we made only for Studies 1a and 1b, while the sample size limitation refers to Study 2. Nonetheless, we believe that our word-pair analyses in retrieval practice experiments (Lage, 2019; Lima et al., 2020) helped us to circumvent this sample size issue by showing that the difficulty of word pairs was relatively retained at least for 7 days.

Second, due to our methodological decision to use word pairs normed for recall accuracy by Nelson and Dunlosky (1994), we cannot guarantee that the normed words would reflect the full range of different word attributes, which seems to have been the case for familiarity. Despite that, we were still able to find relationships between familiarity and other measures (see Tables 3–5). Third, our estimated attributes were made only at the word level and, perhaps for that reason, they accounted for a small portion, if any, of the variance of the recall accuracy across trials. We raised the possibility that word-pair attributes related to the ease of association between a foreign and a native word could potentially predict recall, and we suggested that this may be further explored in future studies.

The use of normed word pairs in human memory research has been increasingly common (e.g., Pyc & Rawson, 2009, 2010). Knowing several word attributes and how they relate with recall accuracy is important for experimental planning. This is the first study to gather normative measures for recall accuracy for word pairs in which either cue or target, or both are in Brazilian Portuguese. More importantly, as far as we know, among the normative studies for recall accuracy (Bangert & Heydarian, 2017; Cho et al., 2020; Grimaldi et al., 2010; Nelson & Dunlosky, 1994), this is the first one that gathered data with participants from a non-WEIRD society. Recently, some scholars have said that psychologists should emphasize close replications, asking what findings generalize to participants from non-WEIRD societies and what findings differ across cultures (Henrich et al., 2010; LeBel & Peters, 2011; Papagno et al., 1991; Roediger & Yamashiro, 2018). Previously, we mentioned mediator-based accounts of retrieval practice effects (e.g., Pyc & Rawson, 2010). One possibility for future studies is to ask whether words that act as more effective mediators in different languages—wing, for wingu–cloud, for English-speaker samples; salame for sahani–prato, for Brazilian-Portuguese-speaker samples—produce similar results across cultures. These norms may prove fruitful not only for Brazilian human memory researchers but also for international research teams, as it will enable the development of more controlled cross-cultural studies in this field.

| Measure | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---------|---|----|----|----|----|----|----|
| 1. Familiarity | .08 | .32* | .38* | .29 | .23 | .40* | .38* |
| 2. Concreteness | .05 | .07 | .10 | .08 | .14 | .04 | .12 |
| 3. Valence | .12 | .28 | .31 | .26 | .23 | .38* | .14 |
| 4. Arousal | -.008 | -.13 | -.17 | -.12 | -.16 | -.15 | -.02 |
| 5. Wordlikeness | .32* | .24 | .28 | .29 | .16 | .14 | .20 |
| 6. Frequency of occurrence | .16 | .25 | .27 | .25 | .21 | .40* | .34* |
| 7. Swahili word length | -.30 | -.23 | -.30 | -.29 | -.14 | -.20 | -.25 |
| 8. Portuguese word length | -.15 | -.23 | -.21 | -.21 | .04 | -.04 | -.02 |
| 9. Recall accuracy, Trial 1 | | I | | | | | |
| 10. Recall accuracy, Trial 2 | .83** | | I | | | | |
| 11. Recall accuracy, Trial 3 | .76** | .94** | | I | | | |
| 12. Recall accuracy, average | .89** | .98** | .96** | | I | | |
| 13. Lage (2019) | .59** | .58** | .58** | .62** | | I | |
| 14. Lima et al. (2020, Exp. 1) | .76** | .87** | .87** | .88** | .69** | | I |
| 15. Lima et al. (2020, Exp. 2) | .69** | .68** | .69** | .72** | .62** | .77** | I |

Note. Correlations among word attributes (columns 1 to 8) were omitted for the sake of simplicity. Item-wise correlations were restricted to a sample of 40 Swahili–Portuguese word pairs.

*p < .05. **p < .01.
Author’s Note
This manuscript is based on a Master’s thesis submitted to the University of Brasília by the first author under the supervision of the second author.

Acknowledgments
We thank Carlos Eduardo D. C. Lage (stimulus translation), Gabriela Y. Iwama, Sebastião Venâncio, and Tatiana Litvin (data collection), Carlos Biagolini-Jr (data analyses), and Beatriz A. Cavendish (data collection, response scoring, and comments on a previous version of this manuscript).

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The study was supported by the National Council for Scientific and Technological Development (CNPq).

ORCID iDs
Marcos Felipe Rodrigues de Lima https://orcid.org/0000-0002-5922-2543
Luciano Grüdtner Buratto https://orcid.org/0000-0002-7003-7824

Supplemental Material
Supplemental material for this article is available on the Open Science Framework website (https://osf.io/ucx7h/)

References
Baddeley, A. D., Eysenck, M. W., & Anderson, M. C. (2015). Memory (2nd ed.). Psychology Press.
Balota, D. A., Piloti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. Memory & Cognition, 29(4), 639–647. https://doi.org/10.3758/bf03200465
Bangert, A. S., & Heydarian, N. M. (2017). Recall and response time norms for English–Swahili word pairs and facts about children’s repetition of nonwords. Applied Psycholinguistics, 12(3), 349–367. https://doi.org/10.1017/S0142716400009267
Grimaldi, P. J., Pyc, M. A., & Rawson, K. A. (2010). Normative multitrial recall performance, metacognitive judgments, and retrieval latencies for Lithuanian–English paired associates. Behavior Research Methods, 42(3), 634–342. https://doi.org/10.3758/BRM.42.3.634
Hair, J. F. Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). Multivariate data analysis (7th ed.). Pearson Education.
Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2–3), 61–135. https://doi.org/10.1017/S0140525X0999152X
Howell, D. C. (2013). Statistical methods for psychology (8th ed.). Wadsworth Cengage Learning.
Janczura, G. A., Castilho, G. M., Keller, V. N., & Oliveira, N. R. (2017). Free association norms for 1004 Brazilian Portuguese words. Psicologia: Teoria e Pesquisa [Psychology: Theory and Research], 32, 1–7. https://doi.org/10.1590/0102-3772e23ne23
Janczura, G. A., Castilho, G. M., Rocha, N. O., Van Erven, T. J. C., & Huang, T. P. (2007). Normas de concretude para 909palavras da línguaportuguesa [Concreteness norms for 909 Portuguese words]. Psicologia: Teoria e Pesquisa [Psychology: Theory and Research], 23(2), 195–204. https://doi.org/10.1590/0102-3772200700200010
Jia, X., Li, P., Li, X., Zhang, Y., Cao, W., Cao, L., & Li, W. (2016). The effect of word frequency on judgments of learning: Contributions of beliefs and processing fluency. Frontiers in Psychology, 6, Article e01995. https://doi.org/10.3389/fpsyg.2015.01995
Johnson, L. W., & MacKay, D. G. (2019). Relations between emotion, memory encoding, and time perception. Cognition and Emotion, 33(2), 185–196. https://doi.org/10.1080/02699931.2018.1435506
Kristensen, C. H., Gomes, C. F. A., Justo, A. R., & Vieira, K. (2011). Brazilian norms for the affective norms for English words. Trends in Psychiatry and Psychotherapy, 33(3), 135–146. https://doi.org/10.1590/S0159-2237-0040-20000003
Lage, C. E. D. C. (2019). The efficacy of retrieval practice in creating stress-resistant memories [Unpublished master’s thesis]. Universidade de Brasília.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310
LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem’s (2011) evidence of psi as a case study of deficiencies in modal research practice. Review of General Psychology, 15(4), 371–379. https://doi.org/10.1037/a0025172
Leitão, J. A. G., Figueira, A. P. C., & Almeida, A. C. F. (2010). Formas de aquisição para 252 nomes comuns [Imageability, familiarity, and age-of-acquisition norms for 252 common words]. Laboratório de Psicologia [Psychology Laboratory], 8(1), 101–119. https://doi.org/10.14417/lp.hl.651
Lima, M. F. R., Venâncio, S., Feminella, J., & Buratto, L. G. (2020). Does item difficulty affect the magnitude of retrieval practice effect? An evaluation of the retrieval effort hypothesis.
Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili–English translation equivalents. *Memory*, 2(3), 325–335. https://doi.org/10.108/09658219408258951

Núcleo Interinstitucional de Linguística Computacional [Interinstitutional Center for Computational Linguistics]. (2005). *Corpus NILC*. http://www.nilc.icmc.usp.br/nile/tools/corpora.htm

Oliveira, N. R., Janczura, G. A., & Castilho, G. M. (2013). Normas de alerta e valência para 908 palavras da língua portuguesa [Norms of arousal and valence for 908 Portuguese words]. *Psicologia: Teoria e Pesquisa* [Psychology: Theory and Research], 29(2), 185–200. https://doi.org/10.1590/S0102-37722013000200008

Paivio, A. (1968). A factor-analytic study of word attributes and verbal learning. *Journal of Verbal Learning and Verbal Behavior*, 7, 41–49. https://doi.org/10.1016/S0022-5371(68)80161-6

Papagno, C., Valentine, T., & Baddeley, A. (1991). Phonological short-term memory and foreign-language vocabulary learning. *Journal of Memory and Language*, 30(3), 331–347. https://doi.org/10.1016/0749-596X(91)90040-Q

Pexrie, J. W. (2007). PsychoPy: Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, 60(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330(6002), 335–335. https://doi.org/10.1126/science.1191465

Pyc, M. A., & Rawson, K. A. (2012). Why is test–restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(3), 737–746. https://doi.org/10.1037/a0026166

Roediger, H. L., I. I., & Yamashiro, J. K. (2018). History of psychological approaches to studying memory. In R. J. Sternberg & W. E. Pickren (Eds.), *The Cambridge handbook of the intellectual history of psychology* (pp. 165–215). Cambridge University Press.

Rubin, D. C., & Friendly, M. (1986). Predicting which words get recalled: Measures of free recall, availability, goodness, emotionality, and pronunciability for 925 nouns. *Memory & Cognition*, 14(1), 79–94. https://doi.org/10.3758/BF03209231

Soares da Silva, A. (2010). Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese. In D. Geeraerts, G. Kristiansen, & Y. Peirsman (Eds.), *Advances in cognitive sociolinguistics* (pp. 41–84). De Gruyter Mouton.

Witherby, A. E., & Tauber, S. K. (2017). The concreteness effect on judgments of learning: Evaluating the contributions of fluency and beliefs. *Memory & Cognition*, 45, 639–650. https://doi.org/10.3758/s13421-016-0681-0