Journal of Memory and Language 127 (2022) 104357

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Does grammatical gender affect object concepts? Registered replication of Phillips and Boroditsky (2003)

Nan Elpers[a], Greg Jensen[a,b], Kevin J. Holmes[a,*]

[a] Department of Psychology, Reed College, United States
[b] Department of Neuroscience and Zuckerman Institute, Columbia University, United States

ARTICLE INFO

Keywords:
Grammatical gender
Linguistic relativity
Object concepts
Replication

ABSTRACT

Many languages assign nouns to grammatical gender categories (e.g., masculine and feminine), and inanimate objects often have different genders in different languages. In a seminal study, Phillips and Boroditsky (2003) provided evidence that such "quirks of grammar" influence how people conceptualize objects. Spanish and German speakers judged person-object picture pairs as more similar when their biological and grammatical genders matched than when they did not, and English speakers showed the same pattern of similarity judgments after learning gender-like categories. These widely cited findings were instrumental in vindicating the Whorfian hypothesis that language shapes thought, yet neither the original study nor any direct replications have appeared in a peer-reviewed journal. To examine the reliability of Phillips and Boroditsky's findings, we conducted a high-powered replication of two of their key experiments (total $N = 375$). Our results only partially replicated the original findings: Spanish and German speakers' similarity judgments exhibited no effect of grammatical gender when accounting for key sources of error variance, but English speakers trained on gender-like categories rated same-gender pairs more similar than different-gender pairs. These results provide insight into the contexts in which grammatical gender effects occur and the mechanisms driving them.

Introduction

Languages divide up the world in strikingly different ways. Even for inanimate objects like clocks and toasters, languages differ not only in their lexical distinctions (Malt & Majid, 2013) but also in the categories imposed by their grammars (Lucy, 2016). Many languages, for example, have a *grammatical gender* system whereby all nouns are assigned to a gender category, most commonly masculine or feminine (Corbett, 2003). Grammatical gender assignment, though usually predictable in the case of humans and other entities with biological sex, is largely arbitrary and semantically illogical in the case of inanimate objects: *clock* is masculine in Spanish and feminine in German, while *toaster* has the opposite genders in the two languages (Boutonnet et al., 2012).[1] Moreover, such "quirks of grammar" (Phillips & Boroditsky, 2003) are unnecessary for communication—speakers of English and Chinese, among many other languages, get by just fine without them (Samuel, Cole, & Eacott, 2019).

But for speakers of grammatical gender languages, the gender category of an object cannot be ignored. Gender is marked obligatorily in such languages via a range of morphosyntactic devices, including adjective inflections, determiners, and pronouns, all of which must agree with the gender of nouns (Corbett, 1991). Compulsory attention to gender in all manner of linguistic contexts suggests an intriguing possibility: that speaking a grammatical gender language might lead people to conceptualize objects as gendered even when they are *not* using language. Grammatical gender thus offers a useful testbed for investigating the classic but oft-maligned Whorfian hypothesis that language shapes thought (i.e., *linguistic relativity*; Whorf, 1940/2012). After decades of spirited debate on this proposal (Gentner & Goldin-Meadow, 2003; Gumperz & Levinson, 1996), including recurring critiques against it (e.g., McWhorter, 2014; Pinker, 1994), a growing consensus has emerged that language modulates cognition in a variety of ways

[*] Corresponding author at: Department of Psychology, Reed College, 3203 SE Woodstock Blvd., Portland, OR 97202, United States.
E-mail address: kjholmes@reed.edu (K.J. Holmes).

[1] In sex-based grammatical gender languages such as Spanish and German, gender assignment for object nouns often depends on phonological or morphological, rather than semantic, information. In a sizeable minority of grammatical gender languages (comprising 25% of a sample of 112 grammatical gender languages from the World Atlas of Language Structures), gender assignment is driven largely by the semantic property of animacy, with no link to biological sex at all (Corbett, 2003).

https://doi.org/10.1016/j.jml.2022.104357
Received 31 December 2020; Received in revised form 8 July 2022; Accepted 19 July 2022

(Wolff & Holmes, 2011), though the nature and extent of this influence remain unresolved (Athanasopoulos & Casaponsa, 2020; Casasanto, 2016; Lupyan, Rahman, Boroditsky, & Clark, 2020; Malt, 2020; Ünal & Papafragou, 2016).

In the case of grammatical gender, a series of seminal experiments by Phillips and Boroditsky (2003; henceforth, P&B) were instrumental in vindicating the Whorfian hypothesis. In one experiment, native Spanish and native German speakers, both proficient in English, were asked (in English) to rate the similarity of pairs of pictures, with each pair consisting of one male or female person and one inanimate object or animal with masculine or feminine grammatical gender. Both groups judged the pictures as more similar when the person's presumptive biological sex was congruent with the object or animal's grammatical gender in their native language, compared to when the two were incongruent. Notably, participants honored the gender distinctions of their native language despite being tested in English. P&B interpreted this finding as evidence that grammatical gender affects object concepts even in an ungendered language context (i.e., beyond merely "thinking for speaking"; Slobin, 1996).

In another experiment, P&B probed this causal claim more directly by manipulating native English speakers' experience with grammatical gender. After learning to classify people and objects according to gender-like categories in a fictional language (e.g., all females and some objects were "oosative," while all males and other objects were "soupative"), English speakers judged picture similarity just as the Spanish and German speakers had: person-object pairs from the same category were rated as more similar than pairs from different categories. P&B concluded that grammatical gender shapes object concepts even in the absence of associated cultural factors.

Perhaps owing to this unabashedly Whorfian conclusion, P&B's findings have received exceptional attention in the linguistic relativity literature, as well as in popular science books on the subject (e.g., Deutscher, 2010; Shariatmadari, 2020). This reception comes despite the fact that the findings were never published in a peer-reviewed journal. The only empirical report of P&B's experiments is a Cognitive Science Society conference paper, and the findings were also featured prominently in a chapter by Boroditsky, Schmidt, & Phillips (2003) in an influential edited volume that revitalized interest in linguistic relativity (Gentner & Goldin-Meadow, 2003). Together, the conference paper and chapter have been cited 1029 times, with 416 citations since 2018 alone (per Google Scholar, as of July 8, 2022).

Despite their enduring impact, P&B's findings stand in contrast to more recent research suggesting that the impact of grammatical gender is more limited than P&B proposed. In a systematic review of 43 grammatical gender studies, Samuel et al. (2019) found that only 32% of the results (accounting for sample size and repeated measures) offer unambiguous support for gender congruency effects and that 43% offer no support. Moreover, the majority of supporting data come from tasks in which gender is highly salient, such as assigning a male or female voice to an object, for which there is no objectively correct choice (e.g., Belacchi & Cubelli, 2012; Kurinski & Sera, 2011). Such tasks may invite participants to engage grammatical gender strategically (Ramos & Roberson, 2011; cf. Gleitman & Papafragou, 2013; Pinker, 1994), rather than revealing its chronic influence on conceptual representations. P&B's similarity task is not immune to these criticisms: each trial requires comparing an object to a gendered human (Ramos & Roberson, 2011), and similarity is no less subjective than the gender of an object's imagined voice (Samuel et al., 2019).

At the same time, P&B's methodology has several unique strengths. First, any effect of grammatical gender on judgments of picture similarity is notable given the many other dimensions that might be invoked, including immediate perceptual features. Second, P&B's unlabeled picture stimuli minimize online language processing—regarded by many as critical for a strong test of the Whorfian claim that language affects nonlinguistic representations (Casasanto, 2016; Wolff & Holmes, 2011). Indeed, P&B was one of only two studies classified as "low" in language salience in the Samuel et al. (2019) review.

Finally, P&B's methods have been regarded as particularly useful for illuminating the mechanisms driving grammatical gender effects (Samuel et al., 2019). A common interpretation of such effects, by scientists and the general public alike, is that inanimate objects are conceptualized as gendered—that Spanish speakers, for example, "imagine a table as… having little skirts on its legs" (Garfield & Vuolo, 2014). Another possibility—arguably just as Whorfian, if less evocative—is that these effects are driven by the *statistical* association between grammatical gender and biological sex (Sato & Athanasopoulos, 2018; Vigliocco, Vinson, Paganelli, & Dworzynski, 2005). For Spanish speakers, tables may be judged similar to human females not because they are mentally imbued with stereotypically feminine features, but simply because *mesa* ('table') and nouns denoting females co-occur with the same gender markers (e.g., the feminine determiner *la*). An extension of P&B's category-learning paradigm has been proposed as a way of disentangling these two accounts. If gender-congruent similarity judgments reflect mere statistical association rather than gendered conceptualization, then following category learning, inanimate objects should be judged no more similar to same-category males or females than to other same-category exemplars equated for their co-occurrence frequency with the objects during learning. Such a pattern, along with the potential for strategic use of grammatical gender information, would render gendered conceptualization an unlikely explanation for gender congruency effects, at least in tasks like P&B's (Samuel et al., 2019).

The trailblazing status of P&B's findings as support for linguistic relativity, and their potential to guide future research on grammatical gender of the sort just discussed, rests heavily on their reliability. Yet no direct replications of P&B's experiments have ever been published, to our knowledge. Establishing reliability is particularly important in this case because peer review is generally more rigorous for journals than for the conference proceedings volume in which the original work appeared. Indeed, P&B's conference paper is missing many details needed to properly evaluate the findings, including key descriptive statistics. Moreover, the experiments were likely underpowered (*N*s = 10–55; *M* = 25), suggesting that the observed effects may have been false positives or inflated relative to their true size in the population (Brysbaert, 2019). For these reasons, P&B's findings have high replication value (Nosek, Spies, & Motyl, 2012).

We therefore conducted direct replications of the two experiments by P&B described above, with native speakers of Spanish and German (Experiment 1) and category-trained native English speakers (Experiment 4). P&B reported three additional experiments, one with participants proficient in both Spanish and German and the others showing that the results were unaffected by a concurrent verbal shadowing task designed to interfere with language processing.[2] These experiments supplement the gender congruency effects observed in speakers of a single grammatical gender language and in category-trained English speakers in the absence of verbal interference. Our replications focused on the reliability of these key effects. As in P&B, the Whorfian prediction of interest in both of our experiments was that pictures from the same grammatical gender category—Spanish and German gender categories in Experiment 1 and trained gender-like categories in Experiment 2—would be rated more similar than pictures from different categories.

Experiment 1: Spanish, German, and English speakers

In P&B's Experiment 1, native Spanish speakers and native German

[2] This null result is difficult to interpret because the shadowing task may not have disabled all linguistic processes covertly recruited for judging similarity (as P&B acknowledged), may not have selectively interfered with linguistic processes (as opposed to categorical processes more generally; Holmes & Wolff, 2012), and lacked a nonverbal counterpart to control for the demands of performing concurrent tasks (Perry & Lupyan, 2013).

Table 1
Demographic Data for Both Experiments.

	Experiment 1			Experiment 2
N (sampled / analyzed)	333/225			189/150
Native language	Spanish	German	English	English
Final *n*	75	75	75	150
Mean age (*SD*)	26.6 (8.1)	32.7 (11.1)	33.7 (12.6)	40.3 (14.6)
Female / male	29% / 68%	41% / 56%	68% / 32%	59% / 38%
Country of origin	57% Mexico, 17% Spain, 15% Chile	92% Germany, 5% Austria, 2% Switzerland	44% U.S., 24% U.K., 12% Canada	64% U.K., 25% U.S., 7% Canada
Country of residence	57% Mexico, 21% Spain, 16% Chile	79% Germany, 9% U.K., 3% Austria	43% U.S., 24% U.K., 12% Canada	67% U.K., 25% U.S., 5% Canada

speakers who were also proficient in English rated the similarity of pictures of objects and animals to pictures of human males and females. The object and animal stimuli were chosen to have opposite grammatical genders in Spanish and German, and all participants were tested in English. The method of our first experiment was the same as P&B's (including their original stimuli), with three exceptions. First, given that some studies have found stronger congruency effects in two-gender languages like Spanish than in three-gender languages like German (which has a neuter gender; e.g., Sera et al., 2002; Vigliocco et al., 2005), we included a control group of English monolinguals whose similarity judgments, unaffected by grammatical gender, provided a baseline for comparison with the Spanish and German groups. Second, we added a labeling task following the similarity task in order to verify the grammatical genders of the object and animal stimuli in Spanish and German. Finally, participants completed the experiment online on their own devices, rather than in a lab. We employed several measures to ensure comparable or higher data quality (Palan & Schitter, 2018).

Method

Materials for both experiments are available on the Open Science Framework (OSF): https://osf.io/f2qjc/.

Participants

Demographics and exclusion criteria. Three groups of participants were recruited using the Prolific participant-sourcing platform (https://www.prolific.co; Palan & Schitter, 2018): native Spanish, German, and English speakers (target *n* per group = 75). The first two groups were also fluent in English. All participants (N = 333) were at least 18 years old and had a good performance record on Prolific (≥95 % approval rate for at least 50 previous studies). As an additional quality filter, participants who failed an initial attention check ("… to demonstrate that you are a participant who reads the study instructions carefully and thoroughly, please check the option 'Other' below and enter the number 8 in the text box of this option") were prevented from completing the study (n = 13). Following our registered protocol, data were excluded from (a) participants who rated their proficiency in any grammatical gender language other than their native language as higher than 2 (low) on a scale from 0 (none) to 10 (perfect) (n = 74), (b) participants who did not complete all measures (n = 12), and (c) Spanish and German speakers who answered <2 of 3 comprehension questions correctly on an auditory measure of native language proficiency (see Procedure; n = 0). We also excluded data from Spanish and German speakers who did not follow instructions on the labeling task (n = 9). These participants used adjectives or English nouns that did not indicate grammatical gender, and therefore we were unable to classify picture pairs in the similarity task as same- or different-gender for these participants. Excluded participants were replaced with other participants who met all inclusion criteria. Upon completing the study, participants received a payment between $1.50 and $1.59. Table 1 shows participant demographic data. Methods for both experiments were approved by the Institutional Review Board at Reed College.

Power analysis. P&B did not report their effect size in Experiment 1, nor sufficient information for computing it (e.g., means and SDs by picture pair type), and we are unaware of other grammatical gender studies for which the dependent measure of interest is the similarity of person-object picture pairs. Moreover, recent work on power analysis in cognitive psychology cautions against relying on effect sizes from prior studies, which may be inflated due to publication bias (Brysbaert & Stevens, 2018), and on those from small-scale pilot studies, which tend to be unreliable (Brysbaert, 2019). Therefore, we adopted Brysbaert's (2019) recommendation (derived from the statistical guidelines of the Psychonomic Society) to select a sample size that provides sufficient power to detect the smallest, non-negligible effect size of theoretical interest in psychological research (d = 0.4).

For P&B's single-variable, repeated-measures design with two levels (same-gender vs. different-gender pairs), a sample of 150 speakers of grammatical gender languages (half Spanish speakers and half German speakers) is more than capable of providing strong evidence in favor of the experimental hypothesis in both frequentist (p <.005) and Bayesian (BF > 10) analyses, with d =.4 and power =.9 (Benjamin et al., 2018; Wagenmakers et al., 2018). These analyses require, respectively, only N = 109 (computed via G*Power; Faul et al., 2007) and N = 130 (assuming that the prior distribution on effect size is a positive-only folded Cauchy with scale r =.707; Brysbaert, 2019). A sample of 150 Spanish and German speakers is also more than capable of providing moderate, non-anecdotal evidence in favor of the null hypothesis (BF < 1/3, which requires N = 80); strong evidence for the null hypothesis (BF < 1/10) is impractical even with a simple repeated-measures design (requires N = 1,800; Brysbaert, 2019).

Power is also enhanced with more observations per participant, especially in repeated-measures designs (Brysbaert, 2019). P&B's Experiment 1 task consisted of 112 unique picture pairs, 56 per pair type (same-gender or different-gender). With 150 Spanish and German speakers, the total number of observations per pair type is 8,400, far exceeding current recommendations for experiments with arguably noisier data (e.g., 1,600 per condition in reaction time tasks; Brysbaert & Stevens, 2018).

As the primary goal of our first experiment was to replicate P&B's Experiment 1, the power considerations discussed above focused on our ability to detect their key gender congruency effect: greater similarity for same-gender than different-gender picture pairs, across both Spanish and German speakers. Data from the monolingual English control group were included only in supplemental analyses comparing the magnitude of the congruency effect between the three groups. With this in mind, we sought the same number of monolingual English speakers as each of the other groups (n = 75), for a total target sample size of 225.

Materials

We used the original P&B stimuli, which consist of 22 pictures, 8 of people and 14 of objects and animals. Of the pictures of people, 4 depicted females (woman, ballerina, bride, girl) and 4 depicted males (man, king, giant, boy). Of the pictures of objects and animals, 7 depicted items classified by P&B as masculine in German and feminine in Spanish (toaster, moon, spoon, broom, whale, frog, fox),[3] and 7 depicted items that are feminine in German and masculine in Spanish (clock, sun, fork, toothbrush, mouse, snail, cat). Each picture was 300 × 300 pixels.

Procedure

The experiment was created using Qualtrics online survey software and presented in English. First, participants read the following instructions for the similarity task (verbatim from P&B): "In this study, you will see pairs of pictures appear on the screen. In each pair, there will be a picture of a person on the left and a picture of an object or animal on the right. Your task is to tell us how similar you think the two things being depicted are. You will see a scale where 1 = not similar and 9 = very similar. For each pair of pictures, please choose a number between 1 and 9 to indicate how similar you think the two things are. Please use the whole scale (give some 1's and some 9's and some of all the numbers in-between." Next, participants provided similarity ratings for all 112 person-object picture combinations, presented individually in a randomized order. Each pair of pictures was shown until participants pressed one of 9 numbered buttons to make their response.

Following the similarity task, participants were asked to report their native language(s). Those who reported Spanish or German then provided labels for the object and animal stimuli in that language, based on the following instructions: "For each picture below, **please type the one word in [Spanish/German] (*not* English)** that is the most appropriate label for the object or animal depicted." The 14 object and animal stimuli were presented on the same page in a randomized order.

Next, Spanish and German speakers completed a brief auditory task as an additional check of their self-reported native language proficiency. They were asked to listen to a 39-second audio clip of a dialogue between two speakers in their native language, adapted from a standardized assessment (Common European Framework of Reference for Languages, C1 level: "proficient user"; Council of Europe, 2001). After listening to the clip, participants answered 3 multiple-choice questions in English assessing their comprehension of the dialogue.

Finally, all participants provided their age, gender, race/ethnicity, country of origin, country of residence, and highest level of education completed. They also listed all languages they knew, indicated how many years of experience they had with each language, and rated their proficiency in each language on a scale from 0 (none) to 10 (perfect).

Registered analyses

Coding of picture pairs

Although the object and animal stimuli were selected by P&B to have opposite grammatical genders in Spanish and German, the labels provided by participants for some stimuli had a different grammatical gender than intended (see Table 2). Therefore, following our registered protocol, picture pairs were classified as same-gender or different-gender for each Spanish and German speaker based on the grammatical gender of the object or animal label provided by the participant in their native language.

Frequentist analyses

P&B's Experiment 1 analysis consisted of traditional paired-samples *t*-tests comparing mean similarity ratings for same-gender and different-gender picture pairs across participants (collapsing across Spanish and German speakers; t_1) and across items (apparently collapsing across the pictures of people, based on the degrees of freedom; t_2). However, mixed-effects models are more powerful and provide a better fit to the data because they account for multiple sources of error variance as random variables (Brysbaert & Stevens, 2018). The present experiment's design had four main sources of error variance: participants, native language, person items (i.e., the 8 unique pictures of people, 1 presented on each trial), and object/animal items (i.e., the 14 unique pictures of objects and animals, 1 presented on each trial). Therefore, for our main analysis of the gender congruency effect in Spanish and German speakers, we used a mixed-effects model to predict similarity ratings from pair type (same-gender vs different-gender), with random slopes and intercepts for participants, native language, person items, and object/animal items. This model accounts for possible differences in the magnitude of the gender congruency effect and/or overall similarity as a function of these factors, via random slopes and intercepts, respectively.

We also conducted a supplemental analysis to compare the gender congruency effect between language groups, including the English monolingual control group. For this analysis, we used a mixed-effects model with both pair type and native language as predictors, and with random slopes and intercepts for participants, person items, and object/animal items. For this analysis, picture pairs were classified as same-gender or different-gender for the English control group based on the grammatical genders of the object and animal labels provided by the German group (i.e., the gender that corresponds to the modal label given for each item).[4] A significant interaction between pair type and native language would indicate that the magnitude of the gender congruency effect differed between language groups. For all mixed-effects analyses, we used the lme4 package (Bates et al., 2015) in R.

Bayesian analyses

To quantify evidence for the null hypothesis of no gender congruency effect in Spanish and German speakers, we also conducted Bayesian analyses. These analyses compared the predictive adequacy of the null hypothesis and the experimental hypothesis that the effect size is positive (i.e., higher similarity for same-gender than different-gender pairs) across Spanish and German speakers and across person items, analogous to P&B's frequentist analyses. In line with our power analysis, the experimental hypothesis assigns effect size a positive-only prior distribution, defined statistically as a Cauchy distribution folded on zero with scale $r = .707$ (Brysbaert, 2019). Below we report the Bayes factors from these analyses, interpret these values qualitatively (compelling support: $BF > 6$ or $< 1/6$; moderate support: $6 > BF > 3$ or $1/3 > BF > 1/6$; inconclusive: $3 > BF > 1/3$; Schönbrodt & Wagenmakers, 2018), and report a robustness region for each (i.e., the range of scale values that would yield the same interpretation; Dienes, 2019). For comparison with P&B, we also report the analogous frequentist *t*-tests and effect sizes.

Results and discussion

Data and analysis code for both experiments are available on the OSF project site: https://osf.io/3fnhm/.

[3] According to Google Translate and a native Spanish-speaking consultant, the most common Spanish translation of *fox* (*zorro*) is masculine, contrary to P&B's classification. The feminine form (*zorra*) has a pejorative, sexual connotation and is typically used to refer to a female person rather than an actual fox. The labeling task (see Procedure) was used to verify the grammatical genders of all of the object and animal stimuli for analysis purposes. As shown in Table 2, all of our Spanish-speaking participants gave the fox picture a masculine label.

[4] As the object and animal stimuli were chosen to have opposite grammatical genders in Spanish and German, either language could be used to classify picture pairs for the English group, for whom similarity ratings were not expected to differ systematically by pair type. We also report exploratory analyses comparing the English group separately to each of the other two groups, with picture pairs classified based on grammatical gender in each of the latter two languages (see Results and Discussion).

Table 2
Grammatical Genders of Object and Animal Stimuli in Experiment 1.

	Spanish				German			
		Modal				Modal		
Stimulus	P&B	Label	Gender	%	P&B	Label	Gender	%
toaster	f	*tostadora*	f	64	m	*Toaster*	m	100
moon	f	*luna*	f	99	m	*Mond*	m	100
spoon	f	*cuchara*	f	99	m	*Löffel*	m	100
broom	f	*escoba*	f	97	m	*Besen*	m	100
whale	f	*ballena*	f	97	m	*Wal*	m	100
frog	f	*rana*	f	77	m	*Frosch*	m	96
fox	f	*zorro*	m	100	m	*Fuchs*	m	97
clock	m	*reloj*	m	100	f	*Uhr*	f	77
sun	m	*sol*	m	100	f	*Sonne*	f	100
fork	m	*tenedor*	m	100	f	*Gabel*	f	100
toothbrush	m	*cepillo*	m	100	f	*Zahnbürste*	f	100
mouse	m	*ratón*	m	88	f	*Maus*	f	100
snail	m	*caracol*	m	99	f	*Schnecke*	f	100
cat	m	*gato*	m	100	f	*Katze*	f	97

Note. For each stimulus, the following information is provided: P&B's gender classifications, the modal label given by Spanish and German speakers in Experiment 1, the gender of that label, and the percentage of participants whose label (modal or otherwise) had the modal gender; m = masculine; f = feminine.

Language proficiency

On the 3 comprehension questions that followed the audio clip in participants' native language, mean accuracy was 96.9% (*SD* = 9.8) for Spanish speakers and 99.6% (SD = 3.8) for German speakers. These results corroborate participants' self-reported language experience.

Main analysis

For our registered main analysis of the gender congruency effect for Spanish and German speakers, we entered their similarity ratings into a mixed-effects model with pair type (same-gender vs. different-gender) as a predictor, and with random slopes and intercepts for participants, native language, person items, and object/animal items. The effect of pair type was not significant, $\chi^2(1) = 2.00$, $p = .16$. Same-gender pairs ($M = 2.57$, $SD = 1.07$) and different-gender pairs ($M = 2.49$, $SD = 1.01$) were rated similarly, providing no evidence for a gender congruency effect. (The descriptive statistics for same-gender and different-gender pairs cannot be compared to those of P&B, who did not report them for Experiment 1.)

Supplemental and exploratory analyses

For our registered supplemental analysis assessing whether the magnitude of the gender congruency effect differed between language groups (including the English monolingual control group), we used a mixed-effects model with pair type and native language as predictors of similarity ratings, with random slopes and intercepts for participants, person items, and object/animal items. The interaction between pair type and native language did not reach significance, $\chi^2(2) = 4.46$, $p = .11$, indicating that the difference in similarity between same-gender and different-gender pairs did not differ markedly across the three language groups (see Fig. 1).

However, Fig. 1 indicates that the German group rated same-gender pairs ($M = 2.83$, $SD = 1.14$) as numerically more similar than different-gender pairs ($M = 2.67$, $SD = 1.00$). To explore whether this difference was significantly larger than that of the English control group (same-gender: $M = 2.48$, $SD = 1.19$; different-gender: $M = 2.39$, $SD = 1.11$) as predicted by P&B's account, we repeated the supplemental analysis described above, but with the Spanish group excluded. The interaction between pair type and native language was not significant, $\chi^2(1) = 2.43$, $p = .12$, providing no evidence that the magnitude of the gender congruency effect differed between groups. That both groups found German same-gender pairs more similar suggests that German grammatical gender categories may capture conceptual or visual similarities that are apparent even to speakers of languages without grammatical gender (Boroditsky & Schmidt, 2000; Foundalis, 2002; but see Vigliocco et al., 2005), at least for the items included in P&B's study and our replication.

For completeness, we compared the Spanish and English groups in the same manner, with picture pairs reclassified for the English group based on the grammatical genders corresponding to the modal Spanish object/animal labels. Once again, there was no significant interaction, $\chi^2(1) = 1.05$, $p = .31$, indicating broadly similar response patterns across groups.[5]

Bayesian analyses

Our registered Bayesian analyses yielded different outcomes than the mixed-effects models because they were analogous to P&B's *t*-tests, which did not account for inter-item or inter-participant variation as random effects (Barr, Levy, Scheepers, & Tily, 2013). For the Spanish and German speakers, the by-participant analysis (averaging over items) yielded a Bayes factor (BF_{10}) of 37.42, regarded as compelling support for a gender congruency effect (robustness region: $r = 0.008$–2). The by-person-item analysis (averaging over participants) yielded a Bayes factor of 2.97, regarded as anecdotal (i.e., inconclusive) evidence for a gender congruency effect (robustness region: $r = 0.001$–0.40; 0.658–2). These results converge with those of the analogous *t*-tests, for which the by-participant analysis was significant, $t(149) = 3.35$, $p = .001$, $d = 0.27$, and the by-person-item analysis approached significance, $t(7) = 2.18$, $p = .065$, $d = 0.77$.

[5] Incidentally, our sample of Spanish speakers was younger and had a larger proportion of male participants than the other two groups (see Table 1). Although participant age or gender could, in principle, account for any differences in similarity ratings between groups or between P&B's results and ours, to our knowledge neither of these factors has been proposed or shown to moderate grammatical gender effects, and P&B did not report demographic information for their sample. That evidence for a gender congruency effect was, if anything, stronger for German speakers than Spanish speakers stands in contrast to other research, in which the reverse pattern is often found when comparing speakers of two-gender languages (e.g., Spanish) and three-gender languages (e.g., German; Samuel et al., 2019).

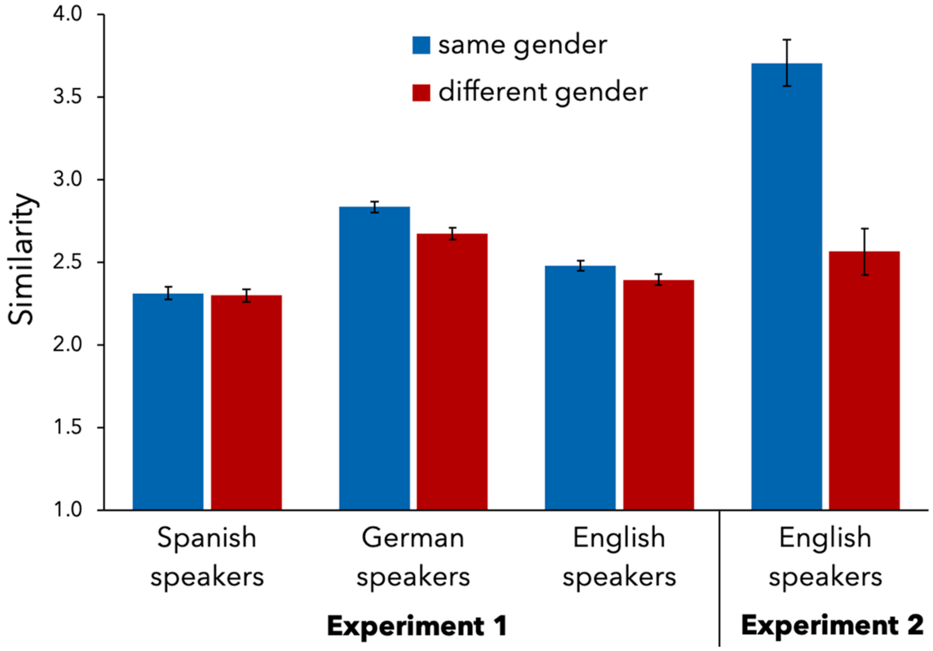

Fig. 1. Similarity ratings by pair type and language group in Experiments 1 and 2. For English speakers, picture pairs were classified based on the grammatical gender corresponding to the modal German object/animal label. Error bars represent 95% within-subjects confidence intervals, computed separately for each experiment.

Although both the Bayesian analyses and *t*-tests meet the assumption of independence by averaging over participants or items, the mixed-effects models reported in the previous sections account for both sources of variance at once (Winter, 2013). Per reviewer suggestion, we conducted exploratory Bayesian analyses that incorporated these sources of variance. To do so, we estimated Bayes factors derived from the Bayesian Information Criterion (BIC) values in each of our mixed-effects models (Lindeløv, 2018).[6] For the effect of pair type for Spanish and German speakers (see Main Analysis above), the BIC-based Bayes factor (BF_{10}) was 0.021, favoring the null model over the experimental model by a factor of 48. For the interaction between pair type and native language for all three language groups (see Supplemental and Exploratory Analyses above), the BIC-based Bayes factor was <.001, favoring the null model by a factor of >2700. For the comparison between the German and English groups, the BIC-based Bayes factor for the interaction was .026, favoring the null model by a factor of 39. For the comparison between the Spanish and English groups, the BIC-based Bayes factor for the interaction was .013, favoring the null model by a factor of 77. Taken together, these exploratory analyses suggest that when inter-participant and inter-item variance is properly accounted for in statistical models of our data, there is substantial evidence against a gender congruency effect overall and against group differences in the magnitude of this effect.

Summary

The results of Experiment 1 fail to replicate P&B’s Experiment 1. Our primary mixed-effects analyses showed that Spanish and German speakers rated same-gender and different-gender picture pairs similarly and that neither group exhibited a larger gender congruency effect than the English monolingual control group. Although Bayesian and frequentist analyses that disregard inter-item variance provided some support for a gender congruency effect, there was strong evidence against such an effect from Bayesian analyses that accounted for both inter-participant and inter-item variance as random effects.

Experiment 2: Training English speakers on gender-like categories

In P&B’s Experiment 4, native English speakers were taught a grammatical distinction (“oosative” vs “soupative”) in a fictional language (“Gumbuzi”). The oosative and soupative categories were distinguished by the presumptive biological sex of the people in them (i.e., all females were in one category and all males in the other), but each category also included inanimate objects. After participants had mastered the distinction, they rated the similarity of each person-object pair. The general method of Experiment 2 was the same as P&B’s Experiment 4, with three exceptions. First, although the majority of the stimuli were P&B’s originals, we were unable to obtain three object pairs used in their Experiment 4 and substituted our own normed stimuli for them. To ensure that any differences between our results and P&B’s were not due to stimulus differences, we conducted analyses both with and without the replacement stimuli. Second, as in Experiment 1, participants completed the experiment online on their own devices. Finally, to prevent attrition from online participants, the category learning phase had a limited number of trials. Participants who failed to adequately learn the oosative/soupative distinction during this phase were excluded from analyses.

Method

Participants

Demographics and exclusion criteria. We recruited 189

[6] Because estimates of BF_{10} based on BIC are only an approximation, we used the brms package (Bürkner, 2017) as a second way to numerically estimate Bayes factors. These analyses are included on our OSF project site. In all cases, the brms-based BF_{10} approximations were consistent with those provided by the BIC method.

monolingual English speakers who were at least 18 years old via Prolific (target $N = 150$). As in Experiment 1, participants had a good performance record on Prolific (≥95% approval rate on at least 50 previous studies) and were prevented from completing the study if they failed an initial attention check ($n = 4$). Following our registered protocol, data were excluded from participants who (a) did not classify at least 80% of items correctly in the final round of test trials in the learning phase (see below; $n = 27$), or (b) did not complete all measures ($n = 7$). We also excluded data from one participant who reported very good proficiency in a second language on the post-experiment demographic questionnaire. Excluded participants were replaced with other participants who met all inclusion criteria. Upon completing the study, participants received a payment of $3.25. See Table 1 for participant demographics.

Power analysis. P&B did not report a standardized effect size in Experiment 4, nor sufficient information for computing it (they reported means, but not SDs, by pair type). Therefore, as in Experiment 1, we relied on Brysbaert's (2019) recommendation and sought a sample of 150 participants, more than capable of providing strong evidence in favor of the experimental hypothesis ($p < .005$ and $BF > 10$) and moderate evidence in favor of the null hypothesis ($BF < 1/3$), with $d = 0.4$ and power = .9 (Benjamin et al., 2018; Wagenmakers et al., 2018). P&B's Experiment 4 similarity task consisted of 96 unique picture pairs, 48 per pair type. With $N = 150$, the total number of observations per pair type is 7,200, far exceeding current recommendations (Brysbaert & Stevens, 2018).

Materials

Following P&B, the stimuli were 20 pictures, 8 of people (the same as in Experiment 1) and 12 of objects. The object stimuli, sized as in Experiment 1, consisted of pairs of similar items: fork and spoon, guitar and violin, apple and pear, pen and pencil, bowl and cup, chair and table. The first 3 pairs were the original stimuli from P&B's Experiment 4; as we were unable to obtain the remaining pairs, we chose substitutes similar in style to the rest. In a norming study conducted on Amazon Mechanical Turk ($n = 34$ native English speakers), each of the 6 replacement pictures received a single dominant label (>97% agreement). One of the pairs (chair and table) replaced a pair used by P&B (pot and pan) because several candidate pictures of the latter objects did not elicit consistent labels.

As in P&B, the members of each pair were assigned to different grammatical categories (*oosative*: fork, violin, pear, pen, cup, chair; *soupative*: spoon, guitar, apple, pencil, bowl, table), such that no two objects in a category came from the same superordinate class (e.g., writing instruments). The categories were anchored by gender: for the pictures of people, participants learned either that all 4 females were *oosative* and all 4 males *soupative*, or vice versa. As a result, the objects within each category were grouped with females for half of the participants, and with males for the other half.

Design and procedure

The experiment was created using Qualtrics online survey software.

Learning phase. First, participants read the following instructions for the learning phase (verbatim from P&B, except that "chair" and "table" were not used as examples because they now served as test items): "In this study you will learn a bit about the Gumbuzi language. In Gumbuzi, there are two different words for 'the.' For example, in order to say 'the pan' you would say 'sou pan,' and in order to say 'the pot' you would say 'oos pot.' This is called the oosative/soupative distinction. Some nouns are always preceded by 'sou' and some are always preceded by 'oos.'" Each picture was then shown individually with its Gumbuzi article and label (e.g., the pear picture accompanied by "oos pear") for 3 s.

After the 20 pictures were each shown 3 times in a randomized order, participants were tested on the oosative/soupative distinction. Each picture was shown individually without a label, and participants indicated whether its label was oosative or soupative by clicking on one of two buttons on the screen. Following a correct response, the next trial began. Following an incorrect response, an error message was shown ("Incorrect. Please select the correct answer.") and the next trial began after participants clicked on the correct button. Test trials continued until participants had correctly classified all items in a given round of 20 trials, or until 3 rounds were completed.[7]

Similarity task. After the learning phase, participants rated the similarity of pairs of pictures. There were 96 trials, one for each person-object combination. In P&B's Experiment 4, participants responded by selecting a number on the keyboard. Our participants pressed a button on the screen, as in Experiment 1. All other aspects of the procedure were identical to Experiment 1.

Registered analyses

Because 6 of our object stimuli were not part of P&B's original set, we conducted all analyses both with and without these stimuli. In all cases, the two sets of analyses yielded similar results. The analyses excluding the replacement stimuli are reported below in Footnote 8.

Frequentist analyses

P&B's main analysis included data from participants who completed the similarity task described above (labeled Experiment 4) as well as from those who completed it while performing a concurrent verbal shadowing task (Experiment 5). However, they also reported separate paired-samples *t*-tests for each group, comparing mean similarity ratings for same-gender and different-gender picture pairs across items only (apparently averaging over the pictures of objects, based on the degrees of freedom). Similar to Experiment 1, we used a mixed-effects model to predict similarity ratings from pair type (same-gender vs different-gender), with random slopes and intercepts for participants, person items, object items, and accuracy in the final round of test trials in the learning phase (ranging from 80 to 100%, in 5% increments).

Bayesian analyses

We conducted Bayesian analyses comparing the predictive adequacy of the null and experimental hypotheses across participants and across object items. The latter were analogous to P&B's frequentist analysis but different from Experiment 1, for which the item analysis was across person items as in P&B. Our registered analyses otherwise mirror those of Experiment 1.

Results and discussion

Learning phase

In participants' final round of test trials in the learning phase, mean accuracy in classifying pictures as oosative or soupative was 95.8% (*SD* = 5.8). Eighty-four participants (56%) achieved perfect accuracy and took an average of 1.5 rounds to do so (*SD* = 0.8).

Main analysis

For our registered main analysis of the gender (oosative/soupative) congruency effect, we entered participants' similarity ratings into a mixed-effects model with pair type as a predictor, and with random

[7] In pilot testing, the majority of participants classified ≥90% of items correctly after 3 or fewer rounds of test trials. Our analyses account for variability in classification accuracy in the final sample (see Registered Analyses). Our learning phase also deviated slightly from P&B's due to its web-based format: P&B's participants responded by pressing computer keys rather than on-screen buttons, and incorrect responses were followed by a beep rather than an error message. P&B did not specify the duration of the familiarization trials or the order in which they appeared.

slopes and intercepts for participants, person items, object items, and accuracy in the final round of test trials in the learning phase. There was a significant gender congruency effect, $\chi^2(1) = 11.80$, $p = .001$, with same-gender pairs ($M = 3.71$, $SD = 2.04$) rated more similar than different-gender pairs ($M = 2.56$, $SD = 1.21$; see Fig. 1). This 1.15-unit difference (on a 9-point rating scale) is descriptively larger than the 0.66-unit difference in P&B's Experiment 4 (same-gender pairs: $M = 4.63$; different-gender pairs: $M = 3.97$), and its 95% confidence interval (0.86–1.42) does not overlap with it.

Bayesian analyses

Our registered Bayesian analyses aligned with the results from the mixed-effects model. Both the by-participant and by-object-item analyses yielded Bayes factors that provide compelling support for a gender congruency effect (by participants: $BF_{10} = 5.01 \times 10^{10}$; by object items: $BF_{10} = 6.48 \times 10^8$; robustness region for both: $r = 0.001–2$). These results converge with those of the analogous *t*-tests (by participants: $t(149) = 8.03$, $p < .001$, $d = 0.66$; by object items: $t(11) = 27.49$, $p < .001$, $d = 7.94$). As in Experiment 1, we also computed an exploratory Bayes factor derived from the BIC values from our mixed-effects model. For the effect of pair type (see Main Analysis above), the BIC-based Bayes factor was 3.05, regarded as moderate support for a gender congruency effect.[8]

Summary

The results of Experiment 2 replicate P&B's Experiment 4 and suggest an even larger effect size. Across mixed-effects and Bayesian analyses, monolingual English speakers who had learned novel grammatical gender-like categories rated same-gender picture pairs more similar than different-gender pairs. These results support P&B's conclusion that grammatical gender can exert a causal influence on perceived similarity.

General discussion

P&B reported two now-classic findings on grammatical gender and the conceptualization of inanimate objects. Spanish and German speakers rated person-object picture pairs more similar when their genders were congruent than when not, and English speakers exhibited a similar congruency effect after learning gender-like categories. These findings have been regarded as compelling evidence that grammatical gender affects object concepts and are widely cited as support for the Whorfian hypothesis that language shapes thought (Samuel et al., 2019; Wolff & Holmes, 2011). Given their outsized impact on the linguistic relativity literature despite the small and likely underpowered samples from which they were derived, P&B's findings have high replication value (Nosek et al., 2012). We conducted replications of two of their key experiments. The results revealed a mixed pattern of findings that provide insight into the contexts in which grammatical gender effects occur and the mechanisms driving them.

Experiment 1 did not replicate P&B's gender congruency effect for Spanish and German speakers. Our primary analysis used a mixed-effects model that accounted for several sources of error variance disregarded in P&B's by-participant and by-item comparisons and in analogous Bayesian analyses. The mixed-effects model yielded no significant difference in similarity ratings between same-gender and different-gender picture pairs. This null result cannot be dismissed on methodological or analytic grounds: our sample of Spanish and German speakers was high-powered, all participants demonstrated comprehension of auditory information in their native language and provided sensible native-language labels for the picture stimuli, and our coding of picture pairs was based on the gender of participants' own labels rather than the presumptive dominant label of each picture. These considerations suggest that our participants' language proficiency matched their self-reports and that our methods were sufficient for detecting the gender congruency effect reported by P&B.

Our additional analyses comparing the Spanish and German groups to the English monolingual control group further show that there was no gender congruency effect in Experiment 1. The English group provides a baseline for assessing the effect of grammatical gender above and beyond any aspects of conceptual or visual similarity that are reflected in grammatical gender categories (Vigliocco et al., 2005). Neither registered nor exploratory mixed-effects analyses yielded any significant differences in the magnitude of the gender congruency effect between the English group and either of the other two groups, and exploratory Bayesian analyses that accounted for the same sources of error variance as the mixed-effects models overwhelmingly supported the null hypothesis of no gender congruency effect and no differences between groups. Taken together, then, the results of Experiment 1 indicate that we failed to replicate P&B.

Admittedly, our results fall short of refuting the Whorfian claim that speakers of grammatical gender languages conceptualize objects as gendered. P&B's similarity task is one of many paradigms that have been used to test this claim, and some others (e.g., gendered "voice choice" tasks) have supported it more consistently (but are also open to alternative explanations; see Samuel et al., 2019). The results of Experiment 1 suggest that, at least in the context of judging similarity between pictures of people and objects, grammatical gender may not be as salient for Spanish and German speakers as previously assumed. On the one hand, this conclusion may be surprising given the prominence of gender in the task and the subjective nature of similarity judgments, which may invite strategic engagement of grammatical gender (Ramos & Roberson, 2011; Samuel et al., 2019). On the other hand, P&B's paradigm is one of few in the literature that have relatively little language content, essentially appearing only in the task instructions. Paradigms that incorporate language processing to a greater extent may be more likely to elicit reliable grammatical gender effects, though some question whether such paradigms are capable of revealing the influence of language on conceptual representations or merely "thinking for speaking" (Gleitman & Papafragou, 2013; Vigliocco et al., 2005).

That said, the results of Experiment 2 suggest that P&B's similarity task *can* elicit reliable effects when grammatical gender is rendered inordinately salient. Across all of our analyses, English monolinguals who had just been trained on novel, gender-like categories rated same-gender picture pairs more similar than different-gender pairs, replicating P&B. These results suggest that grammatical gender—when accentuated via a concentrated dose of training—can exert a causal influence on perceived similarity. In P&B's category-learning paradigm, this influence may well reflect strategic use of grammatical gender, as the temporal proximity of the learning phase and the similarity task may induce experimenter demand. Indeed, several participants in Experiment 2 reported at the end of the experiment that they assumed they should rely on their newly-learned gender categories when judging similarity.

The potential for strategic use of grammatical gender makes it unlikely that effects like P&B's—even when found in speakers of grammatical gender languages—reflect the gendered conceptualization of objects. Instead, participants may be recruiting the statistical association between grammatical gender and biological sex in their native language as a proxy for judging similarity (Sato & Athanasopoulos, 2018;

[8] When the 6 replacement object stimuli were excluded, the registered main analysis yielded a significant gender congruency effect, $\chi^2(1) = 9.98$, $p = .002$ (same-gender pairs: $M = 3.80$, $SD = 2.06$; different-gender pairs: $M = 2.57$, $SD = 1.22$), the registered Bayesian analyses provided compelling support for this effect (by participants: $BF_{10} = 5.26 \times 10^{11}$; by object items: $BF_{10} = 7.06 \times 10^4$; robustness region for both: $r = 0.001–2$), the analogous *t*-tests were significant (by participants: $t(149) = 8.45$, $p < .001$, $d = 0.69$; by object items: $t(5) = 39.85$, $p < .001$, $d = 16.27$), and the exploratory BIC-based Bayes factor was 1.73.

Vigliocco et al., 2005). Although an adaptation of P&B's category-learning paradigm has been proposed as a critical test of these two accounts (Samuel et al., 2019), our findings place some limitations on this proposal. The results of Experiment 1 challenge the reliability of P&B's similarity effect for Spanish and German speakers, and any such effect favoring one account over the other in the category-learning paradigm would not necessarily speak to the mechanisms driving gender congruency effects in other tasks. For these reasons, we suggest that a different approach may be more productive: first determine which contexts elicit reliable gender congruency effects in speakers of grammatical gender languages, ideally using a preregistered protocol like ours to replicate previous findings, and then devise tests of candidate mechanisms in those contexts. Such an approach may help illuminate when and how language affects thought more generally, across a range of domains.

Author Note

We thank Arturs Semenuks, Lera Boroditsky, and Webb Phillips for sharing Phillips and Boroditsky's (2003) stimuli with us, Sarah Wu for assistance with statistical analyses, and Enriqueta Canseco-Gonzalez for helpful discussion. Materials, data, and analysis scripts are available on the Open Science Framework: https://osf.io/f2qjc/.

CRediT authorship contribution statement

Nan Elpers: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Greg Jensen:** Formal analysis, Writing – review & editing. **Kevin J. Holmes:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision.

Declaration of Competing Interest

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

References

Athanasopoulos, P., & Casaponsa, A. (2020). The Whorfian brain: Neuroscientific approaches to linguistic relativity. *Cognitive Neuropsychology, 37*(5–6), 393–412.

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

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

Belacchi, C., & Cubelli, R. (2012). Implicit knowledge of grammatical gender in preschool children. *Journal of Psycholinguistic Research, 41*(4), 295–310.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6–10.

Boroditsky, L., & Schmidt, L. A. (2000). Sex, syntax, and semantics. In L. Gleitman & A. Joshi (Eds.), *Proceedings of the 22nd annual conference of the Cognitive Science Society* (pp. 42–47). Cognitive Science Society.

Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax, and semantics. In D. Gentner, & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 61–79). MIT Press.

Boutonnet, B., Athanasopoulos, P., & Thierry, G. (2012). Unconscious effects of grammatical gender during object categorisation. *Brain Research, 1479*, 72–79.

Bürkner, P. (2017). brms: An R package for Bayesian multilevel models. *Journal of Statistical Software, 80*(1), 1–28.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*(1), 1–38.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1*(1), 1–20.

Casasanto, D. (2016). Linguistic relativity. In N. Riemer (Ed.), *Routledge handbook of semantics* (pp. 158–174). Routledge.

Corbett, G. G. (1991). *Gender*. Cambridge University Press.

Corbett, G. G. (2003). Number of genders (ch. 30); sex-based and non-sex-based gender systems (ch. 31); systems of gender assignment (ch. 32). In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. Retrieved June 21, 2021, from http://wals.info/chapter/30; http://wals.info/chapter/31; http://wals.info/chapter/32.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Deutscher, G. (2010). *Through the language glass: Why the world looks different in other languages*. Metropolitan Books.

Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science, 2*(4), 364–377.

Foundalis, H. E. (2002). Evolution of gender in Indo-European languages. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th annual conference of the Cognitive Science Society* (pp. 304–309). Cognitive Science Society.

Garfield, B., & Vuolo, M. (2014, July 15). Fisherman's Whorf (No. 38) [Audio podcast episode]. In *Lexicon Valley*. Slate. https://slate.com/podcasts/lexicon-valley.

Gentner, D., & Goldin-Meadow, S. (Eds.). (2003). *Language in mind: Advances in the study of language and thought*. MIT Press.

Gleitman, L., & Papafragou, A. (2013). Relations between language and thought. In D. Reisberg (Ed.), *Oxford handbook of cognitive psychology* (pp. 504–523). Oxford University Press.

Gumperz, J. J., & Levinson, S. C. (Eds.). (1996). *Rethinking linguistic relativity*. Cambridge University Press.

Holmes, K. J., & Wolff, P. (2012). Does categorical perception in the left hemisphere depend on language? *Journal of Experimental Psychology: General, 141*(3), 439–443.

Kurinski, E., & Sera, M. (2011). Does learning Spanish grammatical gender change English-speaking adults' categorization of inanimate objects? *Bilingualism: Language and Cognition, 14*(2), 203–220.

Lindeløv, J. K. (2018). How to compute Bayes factors using lm, lmer, BayesFactor, brms, and JAGS/stan/pymc3. [https://rpubs.com/lindeloev/bayes_factors].

Lucy, J. A. (2016). Recent advances in the study of linguistic relativity in historical context: A critical assessment. *Language Learning, 66*(3), 487–515.

Lupyan, G., Rahman, R. A., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in Cognitive Sciences, 24*(11), 930–944.

Malt, B. C. (2020). Words, thoughts, and brains. *Cognitive Neuropsychology, 37*(5–6), 241–253.

Malt, B. C., & Majid, A. (2013). How thought is mapped into words. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(6), 583–597.

McWhorter, J. H. (2014). *The language hoax: Why the world looks the same in any language*. Oxford University Press.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631.

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

Perry, L. K., & Lupyan, G. (2013). What the online manipulation of linguistic activity can tell us about language and thought. *Frontiers in Behavioral Neuroscience, 7*, 122.

Phillips, W., & Boroditsky, L. (2003). Can quirks of grammar affect the way you think? Grammatical gender and object concepts. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th annual conference of the Cognitive Science Society* (pp. 928–933). Cognitive Science Society.

Pinker, S. (1994). *The language instinct*. William Morrow & Co.

Ramos, S., & Roberson, D. (2011). What constrains grammatical gender effects on semantic judgements? Evidence from Portuguese. *Journal of Cognitive Psychology, 23*(1), 102–111.

Samuel, S., Cole, G., & Eacott, M. J. (2019). Grammatical gender and linguistic relativity: A systematic review. *Psychonomic Bulletin & Review, 26*(6), 1767–1786.

Sato, S., & Athanasopoulos, P. (2018). Grammatical gender affects gender perception: Evidence for the structural-feedback hypothesis. *Cognition, 176*, 220–231.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*(1), 128–142.

Sera, M. D., Elieff, C., Forbes, J., Burch, M. C., Rodríguez, W., & Poulin-Dubois, D. (2002). When language affects cognition and when it does not: An analysis of grammatical gender and classification. *Journal of Experimental Psychology: General, 131*(3), 377–397.

Shariatmadari, D. (2020). *Don't believe a word: The surprising truth about language*. W. W: Norton & Company.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking". In J. J. Gumperz, & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge University Press.

Ünal, E., & Papafragou, A. (2016). Interactions between language and mental representations. *Language Learning, 66*(3), 554–580.

Vigliocco, G., Vinson, D. P., Paganelli, F., & Dworzynski, K. (2005). Grammatical gender effects on cognition: Implications for language learning and language use. *Journal of Experimental Psychology: General, 134*(4), 501–520.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58–76.

Whorf, B. L. (1940/2012). *Language, thought, and reality: Selected writings of Benjamin Lee Whorf, 2nd edition* (J. B. Carroll, S. C. Levinson, & P. Lee, Eds.). MIT Press.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. [http://arxiv.org/pdf/1308.5499.pdf].

Wolff, P., & Holmes, K. J. (2011). Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(3), 253–265.