Language and Cognition (2025), 17, e28, 1–19
doi:10.1017/langcog.2024.76

ARTICLE

A conceptual replication of an implicit test of grammatical gender effects on inanimate concepts

Devyani Mahajan and Frank H. Durgin

Department of Psychology, Swarthmore College, Swarthmore, PA, USA
Corresponding author: Frank H. Durgin; Email: fdurgin1@swarthmore.edu

(Received 05 March 2024; Revised 02 December 2024; Accepted 06 December 2024)

Abstract

It has been argued that the incidental and arbitrary use of gender markings for inanimate concepts in language may affect the conceptualization or semantics of those inanimate concepts. The present article sought to replicate the findings of a classic paper that made this argument. Konishi used the potency dimension of the semantic differential method as an implicit measure of perceived gender. He reported that words for inanimate concepts of masculine grammatical gender were rated as higher in potency than words for the same concepts that had feminine grammatical gender. Two preregistered replication studies are reported here. The first was a conceptual replication of Konishi's study that was conducted with 240 bilingual native speakers of either German or Spanish. Included in the study was a follow-up with 120 monolingual native English speakers. This data was used to test whether the grammatical gender in the native languages of German and Spanish speakers affected their sense of the potency of common inanimate categories when tested in a second language (English) in which they were fluent and the nouns had no grammatical gender. A second version of the study was conducted in the native languages of Spanish and German speakers, as a closer attempt at a replication of Konishi's original study. The results of both studies provided evidence against the grammatical-gender hypothesis. Bayesian tests of both studies strongly favored the null hypothesis that there were no grammatical gender effects on implicit measures of perceived potency.

Keywords: bilingual; grammatical gender; Whorf

1. Introduction

Does grammatical gender in one's native language influence how speakers perceive the meanings of words in their own or other languages? Many of the world's most commonly-spoken languages have grammatical gender, where even inanimate nouns are gendered as masculine or feminine (Corbett, 1991). A number of authors have argued that the arbitrary assignment of grammatical gender affects the conceptualization of inanimate nouns, like ‘key’ and ‘bridge’ (e.g., Konishi, 1993). This has been suggested to be a kind of Whorfian effect (Whorf, 1956) in which arbitrary grammatical gender comes to affect the semantics of words (Boroditsky et al., 2003). A recent review has expressed skepticism about this perspective (Samuel et al., 2019), and another recent replication study failed to find evidence consistent with the Whorfian hypothesis (Elpers et al., 2022). The present study was conducted to follow up on an early report concerning grammatical gender (Konishi, 1993).

1.1. Personification

Konishi’s early work (Mackay & Konishi, 1980) was focused on the non-neutrality of masculine pronouns. Within a language, there is some evidence that grammatical gender can influence personification. For example, Konishi (1993) notes an observation that Russian speakers asked to personify the days of the week did so according to their grammatical gender. Segel and Boroditsky (2011) examined personification in European visual art (including Italian, French, German, Spanish, and Dutch painters), concluding that in 78% of cases, gendered personification was consistent with grammatical gender of the artists’s language.

However, it is unclear whether or not there are sometimes correspondences between grammatical gender and what is often called semantic gender. For example, Mackay and Konishi (1980) found that the words sun and moon were consistently differently gendered when personified in English (sun as masculine; moon as feminine) – a gendering that is consistent with their grammatical genders in Latin, French and Spanish. Konishi (1993) suggested that the grammatical genders in these languages might be non-accidental, an idea that has been explored most recently by Bender et al. (2018).

In contrast to these observations from personification, published norms for explicit measures of perceived semantic gender of nouns in English (Scott et al., 2019) and in Dutch (Vankrunkelsven et al., 2024) do not show such differences. In these studies participants were asked to rate words on a semantic scale of association from very feminine to very masculine. In both languages, the word moon (Dutch: maan) was rated slightly less feminine than the word sun (Dutch: zon): 3.4 versus 3.2 in English (where 4 was neutral), and 2.9 versus 2.4 in Dutch (where 3 was neutral). The apparent discrepancy between personification (Mackay & Kinoshi, 1980), and explicit judgments of the semantic gender of moon and sun suggests that it is possible that explicit judgments of semantic gender do not capture the same thing that is captured by implicit measures such as personification.

1.2. The semantic differential method

In his study of grammatical gender, Konishi (1993) chose to use the potency dimension as an implicit measure of perceived masculinity. The potency dimension was identified by Osgood et al. (1957) as an emergent dimension of human judgment. Osgood et al. developed the semantic differential technique to try to quantitatively study the structure of meaning in language, and originally employed dozens of scales. They collected ratings along dozens of Likert scales in which the two ends were

https://doi.org/10.1017/langcog.2024.76 Published online by Cambridge University Press

indicated by words that form polar opposites, intended to create a semantic dimension. These scales included many different adjective pairs. Many of these likely invited more abstract interpretation when applied to concepts to which they did not apply literally (e.g., large versus small, soft versus hard, sweet versus bitter, loud versus quiet; fast versus slow, warm versus cold, sharp versus dull, old versus young). Some scales also included some adjective pairs that seem to invite personification when applied to inanimate concepts (e.g., friendly versus unfriendly). According to factor analysis, the results of semantic differential studies that use large numbers of scales tended to produce three primary dimensions along which judgments of many objects align. Although the goal of Osgood's (1952) project was to quantify semantic meaning in general, the three factors he identified came to be understood as measures of connotative or affective meaning (Osgood et al., 1975) rather than of denotative (objective) meaning.

The first factor, labeled 'evaluative' tended to register what might be regarded as an attitude – the positive or negative evaluative property of the thing to be rated (so words like, good, sweet, friendly, kind would all be aligned with this first factor). The second factor that typically showed up was something that seemed to include perceived size and strength and was labeled 'potency'. Intuitively, whether evaluating politicians (Osgood et al., 1957) or odors (Dalton et al., 2008), both valence and potency are the most salient factors of judgments made along multiple scales. The third factor that emerged was described as 'activity' (as opposed to passivity), though that factor will be of limited concern in the present study.

These three factors have been shown to turn up across many languages and cultures; Osgood et al. (1975) reported the results of a cross-cultural study examining 21 languages (including a variety of European, Asian and Middle-Eastern languages). For each language, lists of qualifiers (e.g., adjectives) were first generated in response to 100 common nouns by native speakers of the language. Subsequently, 50–60 scales based on the generated qualifiers were then used as rating scales for each of the original 100 items in each language. Across the ratings collected in each of the 21 languages, factor analysis consistently found an *evaluative* dimension, a *potency* dimension (typically including size), and an *activity* dimension (typically including speed) as the first, second and third factors.

When Konishi (1993) chose the potency scale as a measure of masculinity of meaning, the association of potency with masculinity was well established (Heise, 1971, cited in Konishi, 1993; Osgood et al., 1957). The association of gender and potency is still present in current populations. A recent study used the semantic differential method to examine perceived race and gender bias in the USA (Billups et al., 2022). Participants in the study judged how society viewed 18 different social groups ('rich people', 'parents', 'Christians', etc.) and used 10 rating scales that included 2 scales each for the dimensions of Evaluation, Potency and Activity, but also included scales for Warmth and Competence, which are favored scales for social comparison (e.g., Kervyn et al. 2013). Billups et al. used principal components analysis (Dunteman, 1989), to convert the ratings data into principal underlying dimensions of judgment. This identified both an evaluative dimension (PC1) and a potency dimension (PC2) across all 10 of the rating scales. The results of their study showed that adding gender information to descriptions of social roles had produced differences in the potency dimension (increased for men; decreased for women), but not the evaluative dimension, whereas adding (Black or White) race information produced differences in the evaluative dimension, but not the potency dimension. This result suggests that the potency dimension is still a potentially valuable way of measuring whether native-language grammatical gender might produce implicit gender-stereotyped perceptions of concepts.

1.3. Recent critiques

Although several studies of grammatical gender effects have used tests in a non-gendered second language (e.g., Boroditsky et al., 2003; Boroditsky & Schmidt, 2000; Phillips & Boroditsky, 2003; Semenuks et al., 2017), most were conducted and reported prior to the implementation of open science practices that limit the likelihood of false positives (Simmons et al. 2011). Indeed, Elpers et al. (2022) recently reported a preregistered replication of the much-cited work of Phillips and Boroditsky (2003). Phillips and Boroditsky had reported the similarity judgments between non-verbal representations of concepts (images) and images of men or of women showed systematic effects of their grammatical gender. Phillips and Boroditsky had used this method, administered in English, with German or Spanish bilingual speakers of English using items that were differently gendered in German and in Spanish. They did this in order to remove the possible concerns of being tested in the native language. In their well-powered, replication of this task however, Elpers et al. found little evidence of grammatical gender effects.

That is, Elpers et al. (2022) replicated the study using large samples of bilingual (English-speaking) native speakers of German or Spanish. The results of the replication did not show any evidence that similarity judgments made between images of inanimate concepts and images of men or women in a study conducted in English were affected by the grammatical gender for the names of the inanimate objects (as later named by the participants) in their native language. The only positive effect from the original paper that replicated 'gender' effects, was for a second study where native English speakers were trained to use two novel articles ('sou' and 'oos') in a fictional language with two lists of words that included gendered items (aligned with the two articles) and non-gendered items. Here categorization effects did seem to align with gender. However, Elpers et al. concluded that this effect seemed likely due to experimental demands, which were explicitly articulated by some participants. Thus, the existing evidence suggests that image categorization tasks simply do not show compelling evidence that grammatical gender controls categorization behavior for non-gendered objects.

1.4. The present study

Because objects may be categorized along many dimensions, perhaps conceptual similarity to gendered images of humans was too subtle a test. The present study sought to do a similar strong bilingual test, also using a preregistered design, that implemented a conceptual replication of the original study by Konishi (1993). Konishi selected the potency dimension specifically because it was an indirect measure of gender. However, Konishi's study was conducted in the native languages of the Spanish speakers and German speakers he tested. As Samuel et al. (2019) and others have pointed out, when testing concepts in the native language of the participant, it is possible that any gender-stereotypical effects observed are mediated by awareness of the grammatical gender of the words presented. To avoid invoking the grammatical

https://doi.org/10.1017/langcog.2024.76 Published online by Cambridge University Press

gender of the rated nouns, the present conceptual replication differed from that of Konishi in that the concepts were tested in a non-gendered second language (English) rather than in the gendered native language.

2. Experiment 1: Conceptual replication of Konishi (1993) in English

The present study sought to use the semantic differential method employed by Konishi (1993) to address a question that has often previously been answered in the positive. However, past methods used, including Konishi's may have been open to experimenter bias, whether by stimulus selection, by flexible definitions of the dependent measure, or by other common practices that artificially inflate the likelihood of a false positive (Simmons et al., 2011). Specifically, the goal of the present study was to test whether the connotative semantics of non-animate concepts tested in a second language are affected by the grammatical genders applied to the equivalent words in one's native language.

It was hypothesized that, if grammatical gender really affects speakers' concepts, the data should show differences in the potency ratings (higher for masculine) based on grammatical gender. To accomplish this, inanimate nouns that were oppositely gendered in German and in Spanish when native speakers of these languages were used. These concepts were tested in a second language, English, in which the participants were also fluent.

The four main prongs of the research strategy were (1) algorithmic selection of the stimuli to be tested in order to avoid experimenter bias, (2) preregistration of the design, to avoid p-hacking, (3) the use of a sensitive, but implicit measure (to avoid experimental demand), that has previously been used successfully, if more crudely, to provide evidence of Whorfian effects, and (4) the use of a test language without grammatical gender (English) in bilingual native speakers of gendered languages that were used previously (Spanish and German).

2.1. Methods

2.1.1. Preregistration and availability of materials

The main study was preregistered at https://aspredicted.org/3Q2_GXV. The follow-up was preregistered at https://aspredicted.org/ZQF_CD1. Complete data and analysis files for the study are available at: https://osf.io/dr9h2/?view_only=a9080eeeba274cd399622ca373e79332, as well as sample images of the online surveys used and code.

All procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008, and the methods were approved by the local IRB (IRB-FY22–23-25).

2.1.2. Stimulus selection

The 4000 nouns with the highest frequency in English according to the Subtlex word frequency dataset (Brysbaert & New, 2009) were considered for investigation. The automated process involved two steps: (1) Each noun was translated into both German and Spanish using both Bing and Google; items that had different translations in the two services for either language were automatically excluded. (2)

https://doi.org/10.1017/langcog.2024.76 Published online by Cambridge University Press

Table 1. Sample stimuli

English	Spanish	German	Spanish gender	German gender
chair	silla	Stuhl	F	M
moon	luna	Mond	F	M
sun	sol	Sonne	M	F
time	tiempo	Zeit	M	F

The Yandex API was then used to determine the grammatical gender of each translated noun, retaining all and only the English nouns whose German and Spanish translations were masculine in one language and feminine in the other.

The final (non-algorithmic) trimming done was to remove 6 words, so that none of the items were animate nouns (‘pal’), none were vulgar (‘dick’, ‘crap’), and none had meanings that were specific to British or American English (‘pants’, ‘trousers’). The word ‘thanks’ was also removed as translating to a verb in German. This approach trimmed the original list down to 82 English nouns, of which 40 were masculine in German and feminine in Spanish, and 42 were feminine in German and masculine in Spanish. Table 1 shows four sample stimuli. The entire stimulus set is provided in Appendix A.

Because each item had to be rated on 10 dimensions, four lists of 20 or 21 items (10 that were masculine in German and 10 or 11 that were masculine in Spanish) were created, so that each participant only had to rate 20 or 21 words. The order of word presentation was randomized for each participant. Note that two English nouns (‘film’ and ‘movie’) were included that had the same primary translations in German and in Spanish. These were presented in separate lists. Eliminating either of them had no effect on the statistical conclusions below.

2.1.3. Participants

Participants (240) were recruited using Prolific, and tested online using PsyToolkit (Stoet, 2010, 2017). Participants received $3.40 for participation (~$12/h), plus a $1 bonus if their performance merited inclusion in the final data set.

The participants included 120 (30 per list) native speakers of German who were living in Germany and had previously reported to Prolific that they were fluent in English. There were also 120 (30 per list) native speakers of Spanish who were living in Spain had previously reported to Prolific that they were fluent in English. A preregistered exclusion criterion was used to remove low-quality raters (those with a low correlation with other raters with the same native language rating the same list). Consequently, the final sample included 99 native Spanish speakers and 95 native German speakers. The participants were roughly balanced in gender (97 women, 95 men, 2 non-binary), and 19–71 years of age (M ± SD: 33 ± 11 years), with a mean age of acquisition of English of 8.5 ± 3.8 years. In one measure of fluency in the survey, the mean self-rated fluency in English was 5.6 on a 1–7 scale (and did not differ between Spanish and German speakers). Data were also collected concerning the types of activities in which they employed English. 67% reporting that they chose to watch movies in English 50% of the time or more, and 48% reporting they chose to read books in English 50% of the time or more. Entropy scores, representing measures of multilingualism in different contexts, were computed for five different contexts (Gullifer & Titone, 2018, 2020). Mean language entropy was low (i.e., indicating fairly monolingual activity) for family situations (M = 0.16; SD = 0.34), but higher for work (0.46 ± 0.48), social situations (0.47 ± 0.50), reading (0.52 ± 0.49), and for audio while movie-watching (0.55 ± 0.50), illustrating that English was used fairly extensively across these latter situations.

2.1.4. Exclusion criterion

The preregistered exclusion criteria were designed to remove inattentive participants by eliminating those whose data vectors (either 200 or 210 unique ratings) had low correlations with the mean vector for their cell. Specifically, participants whose judgments had less than a medium correlation (0.5) with the mean ratings of the same-language group tested with the same list were excluded. This was iteratively implemented by eliminating those with the lowest correlation and then re-computing the mean vector until the lowest correlation across the remaining participants was at least 0.5. Had fewer than 20 participants out of any of the 8 sets of 30 remained after the exclusion process, the preregistered procedure was to recruit additional participants to replace all those initially excluded for that list. However, within each of the 8 cells of the design there were between 22 and 27 participants with acceptable data.

2.1.5. Semantic differential survey

Ratings of each item were made on 7-point scales between polar terms, for 10 dimensions (in interleaved order): including four *evaluative* scales: bad – good, awful – nice, sweet – bitter (reversed), and light – dark (reversed); three *potency* scales: weak – strong, fragile – sturdy, and large – small (reversed); and three *activity* scales: warm – cool, active – passive (reversed), and dull – sharp. An eleventh dimension, abstract to concrete, was added for exploratory purposes, with the idea of testing whether there would be more of a gender effect for abstract or for concrete concepts. This dimension is not included in the analyses here. Because the list of dimensions was not preregistered, we ran a version of the analysis both with and without the concrete-abstract dimension, to ensure that the results were unaffected by its exclusion (they were; moreover, concrete/abstract loaded highest on PC3).

Although nominally each rating scale was expected to contribute to one of the three expected dimensions (evaluative, potency, activity), the preregistration specified that the ratings data would be first averaged by item and native language, and then subjected to Principal Components Analysis (PCA, Dunteman, 1989). This was done with the expectation that the second principal component (PC2) would be the potency dimension, as is typically found for semantic-differential studies. PC2 loadings were then used to convert each included participant's raw ratings into a potency score for each item they had rated.

2.1.6. Observed principal components

Following the method of Billups et al. (2022), principal components analysis with orthogonal rotation computed with R (version 4.2.1; R Core Team, 2022) was used to extract the primary sources of variance in the ratings averaged by item and native language. The full PCA structure is shown in Table 2, along with eigenvalues for each dimension. As is normally found with the semantic differential method, there was an

https://doi.org/10.1017/langcog.2024.76 Published online by Cambridge University Press

Table 2. Principal component loadings[a]

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Bad–good	0.459	0.076	−0.067	0.212	−0.105	0.215	−0.068	0.250	−0.474	−0.620
Awful–nice	0.468	0.063	−0.084	0.167	−0.077	0.124	0.018	0.306	−0.215	0.762
Sweet–bitter	−0.473	−0.001	0.001	−0.035	0.062	−0.007	0.029	−0.231	−0.831	0.164
Light–dark	−0.450	−0.080	−0.089	−0.057	0.187	0.065	−0.228	0.829	0.029	−0.025
Weak–strong	−0.023	0.598	−0.110	0.032	0.133	0.143	−0.735	−0.208	0.063	0.054
Fragile–sturdy	−0.032	0.447	−0.524	0.039	0.463	0.083	0.544	0.007	0.057	−0.049
Large–small	0.083	−0.480	0.041	0.485	0.670	0.111	−0.182	−0.174	0.027	0.018
Active–passive	0.010	−0.249	−0.724	0.155	−0.233	−0.532	−0.222	−0.053	−0.022	−0.017
Warm–cool	−0.345	−0.046	−0.178	0.517	−0.455	0.569	0.095	−0.096	0.171	0.029
Dull–sharp	−0.126	0.363	0.369	0.627	−0.044	−0.536	0.121	0.134	0.010	−0.020
Eigenvalues	4.24	2.40	1.30	0.79	0.48	0.41	0.18	0.12	0.07	0.01

[a] The sign of each dimension in PCA is arbitrary (i.e., all the signs in any column can be reversed). The signs shown here for the first three PCs were set so that the ‘good’ was positive for PC1, ‘strong’ was positive for PC2, and ‘passive’ was negative for PC3. A negative loading means the left end of the scale was more positively related than the right end of the scale.

https://doi.org/10.1017/langcog.2024.76 Published online by Cambridge University Press

Evaluative dimension (PC1), a Potency dimension (PC2), and an Activity dimension (PC3). Specifically, 'sweet', 'nice', 'good', and 'light' loaded highly on PC1 (which accounted for 42% of the variance), 'strong', 'large', and 'sturdy' loaded highly on PC2 (which accounted for 24% of the variance); 'active' and 'fragile' loaded highly on PC3 (which accounted for 13% of the variance). The eigenvalue for PC2, the intended dependent measure, was 2.4, which is well above a typical criterion of 1. Potency scores were then computed for each item, for each participant based on weighting their ratings on the 10 scales by the 10 loadings for PC2.

2.1.7. Analysis

The PC2 (Potency) data were analyzed with a preregistered LMER using the lmerTest package (version 3.1.3; Kuznetsova et al., 2017) in R (version 4.2.1; R Core Team, 2022) across all participants and items, using the Satterthwaite approximation for degrees of freedom (Luke, 2017). Native Language and Native Language Gender were used as planned predictors of potency with the items and participants as error terms, as well as the slope of Native Language Gender with respect to items and participants. Because the preregistered analysis produced a singular fit, Native Language was eliminated from the model, given that the main question concerned Native Language Gender.

2.2. Results

If masculine grammatical gender makes inanimate nouns more potent than feminine gender does, there should have been a positive effect of native language masculinity on potency. Instead, there was a trend in the wrong direction, which goes against Konishi's (1993) findings. Specifically, the trend suggested that natively masculine words tended to be rated slightly lower on potency than natively feminine words, $\beta = -0.05$, $t(77.5) = 1.78$, $p = .079$.

Although there was no stipulated cut-off for English fluency in the study preregistration, 11% of participants rated their English fluency as 4 or lower on a 7-point scale. Eliminating participants with these low fluency ratings did not affect the outcome of the analysis: $\beta = -0.055$, $t(168.1) = 2.02$, $p = .045$ (though only the slope for subjects could be included in this model because other models gave singular fits).

Given this result, an exploratory Bayesian analysis was conducted using the BFPack library (version 1.2.3; Mulder et al., 2019) in R. Specifically, a Bayesian t-test on the mean potency difference (native masculine – native feminine) showed that the odds against the Grammatical Gender hypothesis (i.e., masculine > feminine) were 24.5 to 1 (Figure 1).

As a final exploratory analysis, a test limited to the 28 nouns from the algorithmically-produced set that were also used in Konishi's (1993) study was conducted. The observed beta value for potency was essentially the same for this subset of the items ($\beta = -0.06$) as for the whole set. Thus, it seemed unlikely that the difference between the results of the present study and Konishi's was due to item selection. There was very little overlap between the algorithmically-derived stimuli used in this study and those selected by Boroditsky and Schmidt (2000), so it was not possible to meaningfully examine a subset representing their stimuli.

https://doi.org/10.1017/langcog.2024.76 Published online by Cambridge University Press

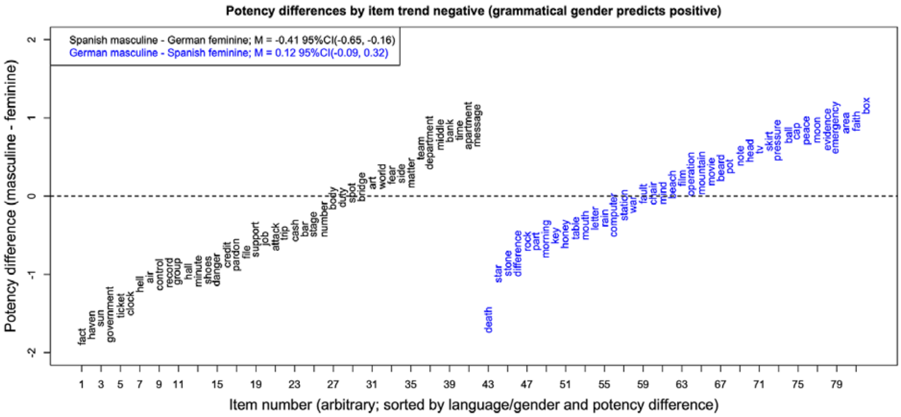

Figure 1. Representation of potency differences (masculine native language – feminine in native language) by item according to native language gender with Spanish masculine words (black), and German masculine words (blue) demonstrating no evidence of increased potency based on masculine grammatical gender in the native language. Error bars represent 95% confidence intervals.

2.2.1. Follow-up with native English speakers

Vigliocco et al. (2005) have suggested that grammatical gender effects may be more evident in some languages than others. Could it be possible, for example, that there were effects in Spanish, but these were masked by something unusual about German (which includes neutral gender as well)? The original study design included no baseline against which to measure each of the languages separately. A second preregistration was therefore made for the collection of an additional set of 120 participants who were identified on Prolific as monolingual native speakers of English in the US. This was done in order to establish a baseline potency rating in English for each word, against which the other two sets of speakers could be compared. A total of 99 native English speakers (50 women; 48 men; 1 non-binary; ages 19–91, M ± SD = 40.5 ± 14.4 years old) provided data that were of at least medium correlation with the means for their cells (the preregistered inclusion criterion).

Potency scores were generated by conducting separate PCAs for each language group (as specified in the second preregistration), so as to take into account any language-specific interpretations of the rating scales. The resulting PC2 loadings were highly correlated across language groups: Between native German speakers and native Spanish speakers, the correlation between the 10 loadings affecting PC2 was 0.99; between monolingual English speakers and native German speakers the correlation was 0.91; between monolingual English speakers and native Spanish speakers the correlation was 0.90. The amount of variance accounted for by PC2 was similar across the three different language groups (ranging from 23% to 26%); the eigenvalues for PC2 ranged from 2.3 to 2.6.

Separate LMERs compared (1) potency ratings of native Spanish speakers with those of native English speakers, and (2) potency ratings of native German speakers with those of native English speakers. Compared to the potency ratings of native English speakers, the potency ratings of native Spanish speakers did not differ significantly for either feminine (in Spanish translation) words, $\beta = -0.023$, $t(92.8) = 0.30$, $p = .76$, or masculine (in Spanish translation) words, $\beta = 0.025$, $t(116.9) = 0.35$, $p = .73$. And

https://doi.org/10.1017/langcog.2024.76 Published online by Cambridge University Press

compared to the potency ratings of native English speakers, the potency ratings of native German speakers did not differ significantly for either feminine (in German translation) words, $\beta = 0.065$, $t(102.0) = 0.98$, $p = .33$, or masculine (in German translation) words, $\beta = -0.079$, $t(91.3) = 1.15$, $p = .25$.

2.3. Discussion

Konishi (1993) was one of the first to investigate the hypothesis that the meanings of inanimate concepts are affected by their grammatical gender in one's native language. He used the potency dimension as his dependent measure and compared ratings by native speakers of German and of Spanish. He reported evidence supporting the Whorfian idea that grammatical gender affected even inanimate concepts. We attempted a conceptual replication of Konishi's study. Like Elpers et al., (2022), the present study used open science methods including preregistration of (1) the conditions, (2) the exclusion rules, (3) the dependent measures, (4) the stopping rule for data collection, as well as (5) the main analysis.

Although we did not replicate Konishi (1993) exactly due to testing our participants in English, the present data provided evidence contrary to the hypothesis. This result held even when only the subset of items also used by Konishi was considered. Would a closer replication of Konishi's study produce evidence of grammatical gender effects?

3. Experiment 2: Attempted Replication of Konishi (1993) in Spanish and in German

One hypothesis concerning grammatical gender effects is that they are sometimes induced by the activation of gender information in the native language of testing (Samuel et al., 2019). Given that Konishi's (1993) experiment was conducted in German and in Spanish, we sought to determine whether we could replicate his observation if we conducted the study by testing native German and Spanish speakers in their native languages. As in Experiment 1, we tested participants who were native speakers of German who resided in Germany and of Spanish who resided in Spain, and we used the same set of 81 nouns, but now in their German and Spanish forms.

For the rating scales, we used the 12 semantic differential scales that were used by Konishi (1993), shown here in Table 3, with the exception that in the Spanish scale for large-small, the word 'chico' was replaced with the word 'pequeño' as being more appropriate for Spain (Konishi had conducted the Spanish version of his study in Mexico).

As shown in Table 3, Konishi (1993) did not use German and Spanish scales that were translations of one another. Rather, he used separate German and Spanish 'pancultural scales' (Konishi, 1993, p. 525) that were based on personal communications from other scientists (see also Osgood et al., 1975, Table 4:18, for the basis for the Spanish scale recommendation). Each set of scales contained four scales for each of three semantic-differential dimensions. This meant that the ratings made in the two languages referred to some of the same concepts, but also to some very different concepts across the two languages. We replicated this choice of scales in case this was an essential detail. However, there remained (in addition to the Spanish speaking country tested) many differences between the studies.

First, Konishi tested University students in Mexico and in Germany using paper surveys. We tested adults online in Spain and Germany using Prolific, which meant

https://doi.org/10.1017/langcog.2024.76 Published online by Cambridge University Press

Table 3. The scales used in the German and Spanish versions of the survey

Dimension	German	Spanish	(English)
Evaluative	schön – hässlich	–	beautiful – ugly
Evaluative	gut – schlecht	bueno – malo	good – bad
Evaluative	angenehm – unangenehm	agradable – desagradable	pleasant – unpleasant
Evaluative	freundlich – unfreundlich	simpático – antipático	friendly – unfriendly
Evaluative	–	admirable – despreciable	admirable – despicable
Potency	kraftvoll – zart	–	powerful – delicate
Potency	schwer – leicht	–	heavy – light
Potency	stark – schwach	fuerte – débil	strong – weak
Potency	gross – klein	grande – pequeño	big – small
Potency	–	gigante – enano	giant – dwarf
Potency	–	mayor – menor	major – minor
Activity	bewegt – ruhig	–	agitated – calm
Activity	lebhaft – gemessen	–	lively – measured
Activity	geräuschvoll – still	–	noisy – silent
Activity	schnell – langsam	rápido – lento	fast – slow
Activity		activo – pasivo	active – passive
Activity		joven – viejo	young – old
Activity		duro – blando	hard – soft

that our participants were older, and were also almost all bilingual in English. Second, our word lists differed (though we could address this both by looking at the shared subset of 28 words his study, and also by filtering our words based on concreteness, as he had done).

In some cases, we were unsure what Konishi did. We chose not to include gendered articles when presenting the nouns to be rated. It is not evident from the published record whether or not Konishi (1993) presented the nouns with their gendered articles. An additional detail we considered was that Konishi (1993) included the words for man, woman, and thing among the items to be rated, but it is unclear whether these were randomly mixed in with the other words or presented at the beginning or at the end. Our (preregistered) decision was to manipulate the ordering so that our participants either (1) rated all three of those words before they rated the experimental words (thus, possibly priming a focus on gender) or (2) rated all three of them only after having rated the experimental words. Our preregistration treated these two orderings as separate attempts to replicate Konishi's result so that we would have two chances at replication.

3.1. Methods

3.1.1. Preregistration and availability of materials

This replication study was preregistered at https://aspredicted.org/r9rt-3ttn.pdf. Data and analysis files, as well as images and code from the surveys are available at: https://osf.io/dr9h2/?view_only=a9080eeeba274cd399622ca373e79332.

The methods were approved by the local IRB under protocol IRB-FY22–23-25.

3.1.2. Design

Ratings were collected for all 81 of the Spanish and German nouns in the Appendix (excluding the duplicates for 'film' and 'movie'). Individual surveys were limited to

30 items (including man, woman, and thing), so three lists of 27 words (13 masculine in German; 14 feminine in German) were created. Each list was administered in German or in Spanish in random order with the German or Spanish words for man, woman, and thing, also randomly shuffled among themselves, but presented either at the beginning or at the end of the survey. Thus, we conducted two versions of the attempted replication. In one version, male and female gender was primed by the early presentation of the words for man and woman. In the other version, the words for man, woman, and thing were not presented for rating until all the experimental items had already been rated.

3.1.3. Participants

A total of 360 participants were recruited using Prolific using the gender-balance feature. Participants received $5.00 for participation (~$12/h). As in Experiment 1, we recruited equal numbers of native German speakers living in Germany, and Native Spanish speakers living in Spain (excluding participants who had participated in Experiment 1) with 30 assigned to each of the surveys. Each participant made a total of 360 ratings (12 ratings for each of 30 words). As preregistered, participants whose rating vector showed a correlation of less than 0.5 with the mean rating vector in each survey were excluded from analysis, yielding 149 German participants (mean age: 32; range 18–61 years old; 72 men, 73 women, and 4 non-binary) and 164 Spanish participants (mean age 31; range 18–65 years old; 80 men, 82 women, 1 non-binary, and 1 unspecified).

3.1.4. Analysis plan

We preregistered the experiment as two attempted replications (one with a gender contrast primed by words for man and woman near at the beginning). Our primary analyses were to be conducted as in Experiment 1, on the PCA-derived potency dimension. For the German participants, the PCA revealed the expected dimensions, with potency as PC2. However, for the Spanish participants, the PCA revealed a different set of dimensions: PC1 seemed to represent something like a contrast of male vs female gender stereotypes, and there was no other candidate for a potency dimension. This led us to add an exploratory analysis in which we simply computed averages based on the intended potency-dimension rating scales. This exploratory analysis was akin to the analysis that Konishi (1993) had performed. We also considered subsets of the stimuli that (a) overlapped with Konishi's, or that were (as preregistered) relatively concrete nouns.

3.1.5. Principal components analysis

Principal components analysis with orthogonal rotation computed with R (version 4.2.1; R Core Team, 2022) was used to extract the primary sources of variance in the ratings separately for German and for Spanish. In both analyses, only the first 3 PCs had eigenvalues greater than 1. For the German analysis, PC2 clearly represented a potency dimension, with high loadings from three of the four intended potency scales (strong, big, and powerful). The four evaluative scales loaded highest on PC1, and the four activity scales loaded highest on PC3, as expected. In the Spanish analysis, however, the four evaluative scales loaded highest on PC2, whereas PC1 reflected a set of contrasts across several different scales that seems fairly consistent

https://doi.org/10.1017/langcog.2024.76 Published online by Cambridge University Press

with age-laden gender stereotypes. Specifically (in order of loadings) the Spanish PC1 contrasted soft, young, friendly, and small with hard, old, unfriendly, and large. Note that soft/hard and young/old are both meant to be activity scales, while small/large is meant to be a potency scale. The remaining two activity scales (fast and active) loaded highest on PC3. Because PC1 better represents the intended potency scale than does PC2 in the Spanish data, we used PC1 for the potency dimension of the Spanish data in the main analysis.

3.1.6. Preregistered analyses

We preregistered two primary analyses using standard hypothesis testing. We also preregistered both (1) a follow-up Bayesian analyses in the case of an absence of evidence from the primary analyses and (2) a follow-up test with a limited subset of stimuli that excluded the most abstract concepts, because Konishi (1993) reported using only concrete nouns.

We first considered the gender-primed version of the experiment. As in Experiment 1, we had to remove native language from the planned LMER in order to avoid a singular fit. The resulting LMER sought to predict potency with native grammatical gender of word with random slopes for participants and for items (concepts). No effect was found: Words that were natively masculine in grammatical gender were not judged more potent than words that were natively feminine, $\beta = 0.005$, $t(79.6) = 0.04$, $p = .97$.

In the version of the experiment without gender priming, the full model converged (including native language as predictor). Once again, however, words that were natively masculine in grammatical gender were not judged more potent than words that were natively feminine, $\beta = 0.004$, $t(79.0) = 0.04$, $p = .97$.

Following the plan of our preregistration, we next used a Bayesian test based on mean difference scores for each concept between its masculine language version and its feminine language version, collapsing across both versions of the experiment. This was done using the BayesFactor R library with the default prior. This analysis showed odds in favor of the null hypothesis of no effect of grammatical gender that were 8.0 to 1.

Because Konishi (1993) reported using only concrete nouns, we had also preregistered a version of the LMER analysis where we eliminated the most abstract third of the of concepts as defined by the ratings collected in Experiment 1. In this case, the analysis was, again, across all the data, and a singular fit with the full model required dropping the random slopes with respect to participants, and again found no evidence of a grammatical gender effect, $\beta = -0.07$, $t(52.0) = 0.75$, $p = .46$.

3.1.7. Exploratory analyses

Given the unexpected structure of the PCA in the Spanish data, it seemed wise to rerun the analyses using the means of the raw scales that were intended to measure potency; this is what Konishi (1993) measured. This exploratory analysis was done as a Bayesian test on the full data set using the item level means, and again the analysis found evidence in favor of the null hypothesis of no effect of grammatical gender. The odds were 7.3 to 1 against the hypothesis that potency differed by grammatical gender.

The same Bayesian analysis was then conducted while limiting consideration to the 28 items that overlapped with those used in Konishi's study. Here, with only about 1/3 the items, the odds still favored the null hypothesis, 4.3 to 1. (The reduced odds against the hypothesis may simply have been the result of including fewer items.)

https://doi.org/10.1017/langcog.2024.76 Published online by Cambridge University Press

3.2. Discussion

In Experiment 2, we sought to replicate Konishi's (1993) result by testing native speakers of Spanish and native speakers of German in their native languages. While many details of our procedure differed from those of Konishi's, and some details of his procedure were simply unknown, our study used 50% more items and four times as many participants. Bayesian analyses of our replication found strong evidence against the hypothesis that the perceived potency of inanimate concepts (that differ in grammatical gender between German and Spanish) was affected by their grammatical gender.

4. General discussion

Does grammatical gender affect the semantics of words referring to inanimate concepts? Konishi (1993) tested this question using lists of words in Spanish and in German for inanimate concepts that had opposite genders in the two languages. He used multi-dimensional ratings of the words to test the hypothesis that concepts given masculine gender in the native language grammar would be rated higher along potency scales embedded in the ratings than those given feminine gender. He concluded that potency ratings were higher for words which were grammatically masculine than for those grammatically feminine. In the 30 years since Konishi (1993) published his observations concerning semantic effects of grammatical gender, interest in this area has grown as a possible example of a kind of Whorfian hypothesis. The methods used to study this question have become more sophisticated in some ways, such as testing bilingual speakers of a gendered language in a language without grammatical gender. However, many of the most commonly cited studies that used this technique (e.g., Phillips & Boroditsy, 2003) were published before the shift toward preregistration.

The present study sought to replicate Konishi's method using more modern methods. Conducting this study today was facilitated by (1) the ready availability of algorithmic methods for stimulus selection and (2) relatively easy access to international populations online that were simply not available 30 years ago. Most importantly, however, the value of preregistration as an important tool for making statistical conclusions meaningfully reproducible was also harnessed to the task. We first attempted to update Konishi's method by testing bilingual speakers (native speakers of German or Spanish) in English, a language which does not have grammatical gender, as a stronger test of his hypothesis. When that failed to show any semantic effects of grammatical gender, we then attempted to replicate his method by testing the same populations of German and Spanish speakers, but tested in their native languages, as Konishi had done. Across both of the studies we conducted, no evidence was found for effects of grammatical gender on the conceptualization of inanimate concepts. Indeed, in both cases Bayesian analyses found strong evidence against the hypothesis.

Like the study by Elpers et al. (2022), the present study may prove useful as a corrective to the many studies that have purported to show effects of grammatical gender. Most were conducted using methods that may have been deeply flawed either by their methods of stimulus selection or by the possibility of p-hacking during post-hoc analyses. Konishi's studies of personification (MacKay & Konishi, 1980) deserve further consideration, but his influential work on grammatical gender effects does not seem to replicate using modern methods. The meanings of nouns may vary from language to language, but it does not seem that grammatical gender interacts much with semantic information.

We believe that Konishi's (1993) study was well conceived. The use of the semantic differential technique provides a sensitive implicit measure of connotative meaning, and, even today, seems like an effective *implicit* method for testing for gender effects. For one, there is a strong association of gender with potency in judgments of persons in English (e.g., Billups et al. 2022). Moreover, explicit ratings of dominance (another way of interpreting the potency dimension) show a small, but reliable correlation with judgments of semantic gender (Vankrunkelsven et al., 2024). The observations reported here are consistent with the conclusions of Montefinese et al. (2019) who used a set of affective concepts (valence, arousal, and dominance) in a comparison between German and Italian, finding no evidence of differential grammatical gender effects on these affective variables.

The use of explicit measures of semantic gender ratings may at some point be useful as an additional methodology, but the risk of effects of experimental demand remains high for such methods. At present, implicit measures that correlate (even weakly) with semantic gender, but clearly correlate with gender stereotypes, seem a useful method for studying Whorfian hypotheses because they seem less susceptible to concerns of experimental demand. Although our results were primarily negative, the present study has contributed to the recent evidence that robust grammatical gender effects on semantic understanding remain elusive when tested rigorously.

Data availability statement. Complete data and analysis files (anonymized) https://osf.io/dr9h2/?view_only=a9080eeeba274cd399622ca373e79332 files for all experiments, as well as sample code and sample images from the surveys are available at https://osf.io/dr9h2/?view_only=a9080eeeba274cd399622ca373e79332.

Preregistration. The study conducted here was preregistered on aspredicted.org. The anonymized preregistration documents may be accessed at: https://aspredicted.org/m6w6-b3rw.pdf, https://aspredicted.org/c58y-hyzt.pdf, https://aspredicted.org/r9rt-3ttn.pdf

Acknowledgements. The authors wish to thank Dr. Peter Baumann, Dr. Beatriz Ramirez Canosa, and Chloe Bobal for their help with preparing the materials for Experiment 2.

Competing interest. The author(s) declare none.

References

Bender, A., Beller, S., & Klauer, K. C. (2018). Gender congruency from a neutral point of view: The roles of gender classes and conceptual connotations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(10), 1580. https://doi.org/10.1037/xlm0000534

Billups, S., Thelamour, B., Thibodeau, P., & Durgin, F. H. (2022). On intersectionality: Visualizing the invisibility of Black women. *Cognitive Research: Principles and Implications*, 7:100. https://doi.org/10.1186/s41235-022-00450-1

Boroditsky, L., & Schmidt, L. A. (2000) Sex, syntax, and semantics. *Proceeding of the Annual Meeting of the Cognitive Science Society*, 22(22). https://escholarship.org/uc/item/0jt9w8zf

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. https://doi.org/10.7551/mitpress/4117.003.0010

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

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

https://doi.org/10.1017/langcog.2024.76 Published online by Cambridge University Press

Dalton, P., Maute, C., Oshida, A., Hikichi, S., & Izumi, Y. U. (2008). The use of semantic differential scaling to define the multidimensional representation of odors. *Journal of Sensory Studies*, 23, 485–497. https://doi.org/10.1111/j.1745-459X.2008.00167.x

Dunteman, G. H. (1989). *Principal Components Analysis (No. 69).* Sage.

Elpers, N., Jensen, G., & Holmes, K. J. (2022). Does grammatical gender affect object concepts? Registered replication of Phillips and Boroditsky (2003). *Journal of Memory and Language*, 127, 104357. https://doi.org/10.1016/j.jml.2022.104357

Gullifer, J. W., & Titone, D. (2018). Compute language entropy with {languageEntropy}. Retrieved from https://github.com/jasongullifer/languageEntropy. http://doi.org/10.5281/zenodo.1403272

Gullifer, J. W., & Titone, D. (2020). Characterizing the social diversity of bilingualism using language entropy. *Bilingualism: Language and Cognition*, 23(2), 283–294. https://doi.org/10.1017/S1366728919000026

Heise, D. R. (1971). Evaluation, potency, and activity scores for 1551 words: A merging of three published lists. Unpublished manuscript, Department of Sociology, University of North Carolina, Chapel Hill.

Kervyn, N., Fiske, S. T., & Yzerbyt, V. Y. (2013). Integrating the stereotype content model (warmth and competence) and the Osgood semantic differential (evaluation, potency, and activity). *European Journal of Social Psychology*, 43(7), 673–681. https://doi.org/10.1002/ejsp.1978

Konishi, T. (1993). The semantics of grammatical gender: A cross-cultural study. *Journal of Psycholinguistic Research*, 22(5), 519–534. https://doi.org/10.1007/bf01068252

Kuznetsova A., Brockhoff P. B., & Christensen R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82 (13), 1–26. https://doi.org/10.18637/jss.v082.i13

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, 49, 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

MacKay, D. G., & Konishi, T. (1980). Personification and the pronoun problem. *Women's Studies International Quarterly*, 3(2–3), 149–163. https://doi.org/10.1016/S0148-0685(80)92092-8

Montefinese, M., Ambrosini, E., & Roivainen, E. (2019). No grammatical gender effect on affective ratings: Evidence from Italian and German languages. *Cognition and Emotion*, 33, 848–854. https://doi.org/10.1080/02699931.2018.1483322

Mulder, J., van Lissa, C., Gu, X., Olsson-Collentine, A., Boeing-Messing, F., Williams, D. R., Fox, J.-P., Menke, J., *et al.* (2019). *Bfpack: Flexible Bayes factor testing of scientific expectations.* (Version 0.2.1) https://CRAN.R-project.org/package=Bfpack

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, 49(3), 197–237. https://doi.org/10.1037/h0055737

Osgood, C. E., May, W. H., & Miron, M. S. (1975). *Cross-cultural universals of affective meaning*. University of Illinois Press.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Phillips, W., & Boroditsky, L. (2003). Can quirks of grammar affect the way you think? Grammatical gender and object concepts. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 25(25). https://escholarship.org/uc/item/31t455gf

R Core Team (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing: Vienna, Austria. https://www.R-project.org/.

Samuel, S., Cole, G., & Eacott, M. J. (2019). Grammatical gender and linguistic relativity: A systematic review. *Psychonomic Bulletin & Review*, 26, 1767–1786. https://doi.org/10.3758/s13423-019-01652-3

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51, 1258–1270. https://doi.org/10.3758/s13428-018-1099-3

Segel, E., & Boroditsky, L. (2011). Grammar in art. *Frontiers in Psychology*, 1:244. https://doi.org/10.3389/fpsyg.2010.00244

Semenuks, A., Phillips, W., Dalca, I., Kim, C., & Boroditsky, L. (2017). Effects of grammatical gender on object description. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 39, 1060–1065.

Simmons J. P., Nelson L. D. & Simonsohn U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Stoet, G. (2010). PsyToolkit – A software package for programming psychological experiments using Linux. *Behavior Research Methods*, 42*(4)*, 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

https://doi.org/10.1017/langcog.2024.76 Published online by Cambridge University Press

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1), 24–31. https://doi.org/10.1177/0098628316677643

Vankrunkelsven, H., Yang, Y., Brysbaert, M., De Deyne, S., & Storms, G. (2024). Semantic gender: Norms for 24,000 Dutch words and its role in word meaning. *Behavior Research Methods*, 56, 113–125. https://doi.org/10.3758/s13428-022-02032-x

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. https://doi.org/10.1037/0096-3445.134.4.501

Whorf, B. (1956). In J.B. Carroll (Ed.), *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. MIT Press.

Appendix

English stimuli used in Experiment 1, and their top translations in German and Spanish, which were used in Experiment 2.

English noun	German noun	Spanish noun	German gender	Spanish gender
area	Bereich	área	m	f
ball	Ball	pelota	m	f
beach	Strand	playa	m	f
beard	Bart	barba	m	f
box	Kasten	caja	m	f
cap	Deckel	gorra	m	f
chair	Stuhl	silla	m	f
computer	Computer	computadora	m	f
death	Tod	muerte	m	f
difference	Unterschied	diferencia	m	f
emergency	Notfall	emergencia	m	f
evidence	Beweis	evidencia	m	f
faith	Glaube	fe	m	f
fault	Fehler	culpa	m	f
film	Film	película	m	f
head	Kopf	cabeza	m	f
honey	Honig	miel	m	f
key	Schlüssel	llave	m	f
letter	Buchstabe	carta	m	f
mind	Geist	mente	m	f
moon	Mond	luna	m	f
morning	Morgen	mañana	m	f
mountain	Berg	montaña	m	f
mouth	Mund	boca	m	f
movie	Film	película	m	f
note	Hinweis	nota	m	f
operation	Betrieb	operación	m	f
part	Teil	parte	m	f
peace	Frieden	paz	m	f
pot	Topf	maceta	m	f
pressure	Druck	presión	m	f
rain	Regen	lluvia	m	f
rock	Felsen	roca	m	f
skirt	Rock	falda	m	f
star	Stern	estrella	m	f
station	Bahnhof	estación	m	f
stone	Stein	piedra	m	f
table	Tisch	mesa	m	f
tv	Fernseher	televisión	m	f

(*Continued*)

(*Continued*)

war	Krieg	guerra	m	f
air	Luft	aire	f	m
apartment	Wohnung	departamento	f	m
art	Kunst	arte	f	m
attack	Attacke	ataque	f	m
bank	Bank	banco	f	m
bar	Bar	bar	f	m
body	Karosserie	cuerpo	f	m
bridge	Brücke	puente	f	m
cash	Kasse	dinero	f	m
clock	Uhr	reloj	f	m
control	Kontrolle	control	f	m
credit	Anerkennung	crédito	f	m
danger	Achtung	peligro	f	m
department	Abteilung	departamento	f	m
duty	Pflicht	deber	f	m
fact	Tatsache	hecho	f	m
fear	Furcht	miedo	f	m
file	Datei	expediente	f	m
government	Regierung	gobierno	f	m
group	Gruppe	grupo	f	m
hall	Halle	salón	f	m
haven	Oase	refugio	f	m
hell	Hölle	infierno	f	m
job	Arbeit	trabajo	f	m
matter	Angelegenheit	asunto	f	m
message	Botschaft	mensaje	f	m
middle	Mitte	medio	f	m
minute	Minute	minuto	f	m
number	Nummer	número	f	m
pardon	Begnadigung	indulto	f	m
record	Aufzeichnung	registro	f	m
shoes	Schuhe	zapatos	f	m
side	Seite	lado	f	m
spot	Stelle	lugar	f	m
stage	Bühne	escenario	f	m
sun	Sonne	sol	f	m
support	Unterstützung	apoyo	f	m
team	Mannschaft	equipo	f	m
ticket	Fahrkarte	boleto	f	m
time	Zeit	tiempo	f	m
trip	Reise	viaje	f	m
world	Welt	mundo	f	m

Cite this article: Mahajan, D., & Durgin, F. H. (2025). A conceptual replication of an implicit test of grammatical gender effects on inanimate concepts, *Language and Cognition*, 17, e28, 1–19. https://doi.org/10.1017/langcog.2024.76