Gender congruency from a neutral point of view:

The roles of gender classes and conceptual connotations

Andrea Bender[1,2], Sieghard Beller[1,2], & Karl Christoph Klauer[3]

[1] Department of Psychosocial Science, University of Bergen, Norway

[2] SFF Centre for Early Sapiens Behaviour (SapienCE), University of Bergen, Bergen, Norway

[3] Department of Psychology, University of Freiburg, Germany

Corresponding author's address

Andrea Bender

Department of Psychosocial Science

University of Bergen, N-5020 Bergen, Norway

Tel: +47 55 58 90 81

email: Andrea.Bender@uib.no

Abstract

The question of whether language affects thought is long-standing, with grammatical gender being one of the most contended instances. Empirical evidence focuses on the gender congruency effect, according to which referents of masculine nouns are conceptualized more strongly as male and those of feminine nouns more strongly as female. While some recent studies suggest that this effect is driven by conceptual connotations rather than grammatical properties, research remains theoretically inconclusive due to the confounding of grammatical gender and conceptual connotations in gendered (masculine or feminine) nouns. Taking advantage of the fact that German also includes a neuter gender, the current study attempted to disentangle the relative contributions of grammatical properties and connotations to the emergence of the gender congruency effect. In three pairs of experiments, neuter and gendered nouns were compared in an *Extrinsic Affective Simon Task* based on gender associations, controlled for a possible role of gender-indicating articles. A congruency effect emerged equally strongly for neuter and gendered nouns, but disappeared when including connotations as covariate, thereby effectively excluding grammatical gender as the (only) driving force for this effect. Based on a critical discussion of these findings, we propose a possible mechanism for the emergence of the effect that also has the potential to accommodate conflicting patterns of findings from previous research.

Keywords: Cognition; Language; Grammatical Gender; Extrinsic Affective Simon Task (EAST); Linguistic Relativity

Introduction

How strongly do properties of the words we use interfere with the way in which we conceptualize the things we talk about? This question is of core interest in psychology, as it touches upon the relationship between language and thought, also known as the *linguistic relativity principle* (Whorf, 1956); yet research on it has undergone waves of approval and dismissal for almost a century (Lucy, 2016; Wolff & Holmes, 2011). Since its revival in the 1990s (Gentner & Goldin-Meadow, 2003; Gumperz & Levinson, 1996; Lucy, 1992), it has been gaining empirical support in various domains (e.g., Boroditsky & Gaby, 2010; Dolscheid, Shayan, Majid, & Casasanto, 2013; Gilbert, Regier, Kay, & Ivry, 2008; Haun, Rapold, Janzen, & Levinson, 2011; Imai & Gentner, 1997; Majid, Bowerman, Kita, Haun, & Levinson, 2004)—to the extent that "the burden of proof has shifted" (Lucy, 2016, p.498), with evidence increasingly being required to sustain claims on the absence rather than the existence of such effects. Accordingly, questions have shifted away from whether relativity effects exist to a focus on when and how they operate and on the factors that affect their strength and durability.

Evidence in favor of relativity effects is particularly convincing for moderate readings of linguistic relativity, which hold that thinking will be affected by language (a) just *before*, (b) *while*, and (c) *after* using language (Wolff & Holmes, 2011). Version (b) refers to cases in which language provides an essential tool for a given cognitive task, as with number words, which are already indispensable for counting and calculating (Frank, Everett, Fedorenko, & Gibson, 2008), not to mention its influence on numerical cognition more generally (Bender & Beller, 2017). More relevant to the question at hand are versions (a) and (c). The former describes the fact that the grammatical rules in every language define which information needs to be specified in order to

produce correct sentences. Planning to express something linguistically therefore compels us to focus our attention on certain aspects of the world, while we are free to neglect other aspects. In the long run, this "thinking for speaking" (Slobin, 1996, 2003) may also give rise to the latter version (c), in that habits emerging from it are likely to persist even in contexts in which respective information is not (yet) immediately needed for verbalization. Insofar as languages use grammatical distinctions to prioritize certain pieces of information over others, they may shift attention—like a spotlight—to particular information in a regular and sustained manner.

Wolff and Holmes (2011) mention grammatical gender as one instance of this "spotlight effect" (see also Bassetti, 2007). Originally being simply an abstract system of noun classes that determines belonging to a declensional paradigm and ensures agreement in formal properties of words associated with the noun (Comrie, 1999; Corbett, 1991), the occurrence of masculine and feminine gender in most European languages has motivated an inviting hypothesis: If properties of a word do affect how we think about things, then the grammatical gender of a noun should increase gender-congruent associations in its referent (*gender congruency effect*). More concretely: The referents of nouns with a masculine gender should be conceptualized as somewhat more male and the referents of feminine nouns as somewhat more female.

The Case for Gender Congruency: Pros and Cons

This assumed influence of grammatical gender on conceptualization is contentious, though, both on theoretical and empirical grounds. On theoretical grounds, a relativistic effect of language on thought would be most plausible for domains in which the boundaries between stimuli are fuzzier in perception than in linguistic categorization (Cibelli, Xu, Austerweil, Griffiths, & Regier, 2016). The underlying mechanism for effects in these domains would be *referent codability:*

Having a systematic set of labels for a concept not only facilitates encoding and classification (Lucy, 2016), but may be instrumental for category formation itself (Lupyan, 2008) as well as for interactions between different levels of processing (Lupyan & Clark, 2015). In such instances, language exerts its influence on cognition through linguistic and conceptual development, by changing an initially implicit representational system into a more explicit system, and by facilitating attention to dimensions of experience for which no initial preference is set (Gomila, 2015). A prime example here are the color terms (for an overview of domains, see Malt & Majid, 2013).

On the continuum of how experiences may be classified via language, referent codability is situated at the concrete end, being directly connected to properties of the referents. Its counterpart at the relational end is the *analogical transfer* of category properties, such as when large-scale grammatical patterns are carried over through metaphors to a different domain (Lucy, 2016). The prototypical example of this mechanism is grammatical gender, which is largely meaningless in terms of semantics (Corbett, 1991). The gender class into which a noun falls is often arbitrary, both within and across languages (Foundalis, 2002), and hence does not reflect real-world differences (Bassetti, 2007). While this renders relativistic effects less likely (Cibelli et al., 2016), it does not rule them out: As a large-scale grammatical pattern, gender classes may still be taken as the basis for categorizing referents in a semantically meaningful way through analogical transfer (Lucy, 2016). After all, when acquiring a first language with a gender system, users may not be aware of its arbitrariness (Bassetti, 2014) and may form assumptions about an underlying relatedness of nouns based on their belonging to the same class. For this very reason, Bassetti considers grammatical gender to be a particularly appropriate test bed for investigating the "pure effects of language" (2007, p. 254), as confounding effects of reality or perception can be excluded.

On empirical grounds, findings on this issue have been heterogeneous from the start, with some early studies reporting a gender congruency effect (Clarke, Losoff, McCracken, & Still, 1981; Ervin, 1962; Mills, 1986), albeit only for some of their samples (Konishi, 1993, 1994), and others finding no confirmation at all (Guiora, 1983; Guiora & Sagi, 1978; Hofstätter, 1963). While this early wave of studies used tasks like the *semantic differential* to measure indirectly sex-related connotations of gendered nouns (such as valence or potency), different paradigms were used in later years. The following overview of findings is ordered based on the approaches adopted in the various studies.

The voice assignment task (VAT). A task that became particularly popular in the field of language development, and later on in research on bilingualism, is the *voice assignment task* (VAT) developed by Sera, Berge, and Castillo Pintado (1994). It aims at a more direct assessment of effects of grammatical gender on cognition, specifically on object categorization, and yielded a range of valuable insights. The studies by Sera and colleagues (1994), for instance, demonstrated that speakers of Spanish do assign male or female voices to objects in line with the gender distinction in their language, even when the stimuli are presented as pictures only. While this applied for adults and for 8-year-old children, it did not hold for younger children. This pattern of findings was replicated for speakers of Spanish, Italian, and French, but not German (Bassetti, 2007; Flaherty, 2001; Sera, Elieff, Forbes, Burch, Rodríguez, & Dubois, 2002), and indicates that an initially semantic-based categorization is replaced (or superimposed) before or at about age 8 by a gender-based categorization—at least in languages with a two-gender system.

Knowledge of more than one language complicates the picture slightly. When children acquire a second gendered language in addition to a first gendered language, the gender congruency effect disappears, arguably due to an increasing awareness of the arbitrariness of the gender system (Bassetti, 2007). In contrast, those who acquire a gendered and a non-gendered

language do show the effect of gender on object categorization from age 8 in both languages (Nicoladis & Foursha-Stevenson, 2012), although the order of acquisition may matter for whether the effect generalizes to the non-gendered language (Forbes, Poulin-Dubois, Rivero, & Sera, 2008). Similar observations were made for native speakers of non-gendered languages, who learn a gendered language as adults. After several weeks of language instruction, these speakers were influenced by the gender categories in the newly learned language (Kurinski, Jambor, & Sera, 2016; Kurinski & Sera, 2011).

While this line of research has produced coherent evidence for the assumption that a grammatical gender system affects object categorization, a major concern with the applied method is that it asks for sex-related associations (male vs. female voices) rather explicitly, meaning that a strategic usage of gender information in solving the task cannot be excluded. To address this concern, more implicit techniques were probed.

Implicit techniques. One study tested two implicit techniques (Vigliocco, Vinson, Paganelli, & Dworzynski, 2005): a task based on triadic similarity judgments, in which participants were asked to pick out the odd item in triplets presented either as pictures or words, and a task that used a continuous naming paradigm to induce semantic substitution errors. A gender congruency effect was obtained in both tasks, but only for speakers of Italian (not of German), only for animal items (not for artifacts), and only when stimuli were presented as words (not as pictures). The authors concluded that "the mechanisms assumed by the sex and gender hypothesis apply to language development but do not extend to conceptual structures" (p. 511).

This conclusion is in line with two earlier psycholinguistic studies: one suggesting that the distinction of nouns based on grammatical gender is not incorporated into the conceptual representation of their referents (Bowers, Vigliocco, Stadthagen, & Vinson, 1999), and one suggesting that consideration of gender information depends on a linguistic context that requires it

(Vigliocco, Vinson, Indefrey, Levelt, & Hellwig, 2004). It also corresponds to a more recent study measuring event-related potentials, which found no behavioral effects on object categorization in English-Spanish bilinguals, despite spontaneous and automated access to gender information (Boutonnet, Athanasopoulos, & Thierry, 2012).

Vigliocco and colleagues went a step further, by also scrutinizing their conclusion with monolingual and bilingual speakers of Italian and English. Again using a substitution error induction task, but this time with picture stimuli of animals only, they were able to replicate gender effects for the Italian monolinguals as well as for the bilinguals, but for the latter only in an Italian-speaking context. This intra-speaker relativity in semantic representations within bilinguals was taken as converging support for the claim that the grammatical gender of nouns does not alter the conceptualization of their referents (Kousta, Vinson, & Vigliocco, 2008).

The same conclusion was also drawn by three additional, largely unrelated studies. Using similarity judgments for pairs of object nouns with either the same or different gender, and similarity judgments for triads of words or pictures, Ramos and Roberson (2010) observed a gender congruency effect among their Portuguese-speaking participants only for words, and an asymmetry between congruent and incongruent pairs, which they took as evidence for effects of linguistic processing. Likewise, in a category decision task with pictures, Cubelli, Paolieri, Lotto, and Job (2011) found faster responses to same-gender pairs in both Italian and Spanish speakers, but not when articulation was suppressed. And Bender, Beller, and Klauer (2011) found priming effects on lexical and gender decisions in German for grammatical primes, but not for semantic primes.

The only study to use implicit measures and still obtain effects of grammatical gender was a set of experiments reported by Boroditsky and colleagues. In these experiments, for instance, participants memorized pairs of nouns and names, and were later asked to recall the gender of the name that was paired with a given noun (Boroditsky & Schmidt, 2000; and see Boroditsky,

Schmidt, & Phillips, 2003). In this case, native speakers of Spanish and German were tested in English on nouns for which grammatical gender diverged in the two languages, and a significant interaction was reported. However, two subsequent studies had difficulties replicating such findings with speakers of German (Koch, Zimmermann, & Garcia-Retamero, 2007), or with speakers of either German or Spanish, even if participants were tested in their native language (Mickan, Schiefke, & Stefanowitsch, 2014). This seems to indicate that the observed effect is rather unstable in this context.

Finally, Imai, Saalbach, and colleagues investigated whether gender congruency may carry over to the inferences that people draw regarding biological properties of animals. Animals are a specific category insofar as most species encompass both male and female individuals, while the grammatical gender of the noun referring to the species is typically used in a generic manner. Taking advantage of this conflicting classification, the authors found that 5-year-old preschoolers speaking (gendered) German are influenced in their inferences by the generic gender of the nouns, compared to preschoolers speaking (non-gendered) Japanese (Saalbach, Imai, & Schalk, 2012). Such an influence was also found for adult speakers of German, even when an adjective indicated the specific sex of the animal (Imai, Schalk, Saalbach, & Okada, 2014). As this effect depended on the presence of the gender-indicating article, however, the authors concluded that it is not the conceptual representations of animals per se that are affected by gender, but that the effect is carried by the article.

Emerging Patterns. Taken together, the studies reported here paint a rather complex picture, although with a few consistent patterns. Moreover, the discrepant findings can be partially attributed to methodological differences insofar as findings of a gender congruency effect have been more likely with explicit measures than with implicit measures. Studies adopting *explicit measures* like the VAT have produced coherent evidence for the assumption that grammatical

gender affects object categorization, not only for words, but also for pictures. This influence is observed in native speakers from 8 years of age onwards—that is, a few years after children have acquired the gender system—and even in adult speakers of a non-gendered language upon learning a gendered language. In addition, the presence of the effect also seems to depend on the investigated language and the number of gender classes it contains. Specifically, the effect appears to emerge only if the language distinguishes exclusively between masculine and feminine, and if participants have not acquired a second language with a diverging gender system. While two-gender languages allow for a straightforward mapping between grammatical and biological gender, this mapping is rendered less consistent, and hence less likely, in a three-gender system (Sera et al., 2002; see also Koch et al., 2007; Vigliocco et al., 2005)[1].

Even if this pattern is replicated with similar explicit tasks by studies that additionally employ *implicit measures*, findings with their implicit measures also suggest that the impact of grammatical gender is confined to the lexical level, as its effect is more stable for words than for pictures (Bowers et al., 1999; Ramos & Roberson, 2010; Vigliocco et al., 2005), is more likely to emerge for some categories (notably animals) than others (Cook, 2016; Saalbach et al., 2012; Vigliocco et al., 2005), depends on language context (Kousta et al., 2008; Vigliocco et al., 2004), disappears under articulatory suppression (Cubelli et al., 2011), or is not manifest on the behavioral level at all (Bender et al., 2011; Boutonnet et al., 2012). Finally, when investigating inferences on the biological properties of animals, effects of grammatical gender are found beyond those of the

[1] This is true, of course, only in languages whose two genders are masculine and feminine. Languages like Dutch, Swedish, and parts of Norwegian, by contrast, distinguish one gender common for (formerly) masculine and feminine nouns from the neuter gender. In such cases, the three-gender language (variant) may exhibit a gender congruency effect, while the two-gender language (variant) does not (Beller, Brattebø, Lavik, Reigstad, & Bender, 2015).

VAT (e.g., for a three-gender language, and for younger children), but again depend on the linguistic context, at least in adults (Imai et al., 2014; Saalbach et al., 2012).

How can these two patterns of findings be reconciled? One possibility which is compatible both with the supportive findings from studies using the VAT and with the qualifications implied by the studies using implicit measures is that the gender congruency effect does emerge at the lexical level in tasks that allow for linguistic processing, but does not reach the conceptual level (Cubelli et al., 2011; Ramos & Roberson, 2010; Vigliocco et al., 2005; and see Arnon & Ramscar, 2012). This is supported by studies showing that gender is spontaneously, yet unconsciously, accessed together with the nouns for objects, even when verbalization is not required (Boutonnet et al., 2012; Cubelli et al., 2011; and see Cubelli, Lotto, Paolieri, Girelli, & Job, 2005; Müller & Hagoort, 2006). Speakers of gendered languages may then, deliberately or not, utilize this information in tasks that explicitly ask for sex associations, throughout the trials or as a fall-back position for instances in which they are uncertain.

Drawing on gender information in this manner may be more inviting in two-gender than in three-gender languages. Yet, languages differ not only in how many gender classes they comprise, but also in how they convey information on a noun's belonging to one of these classes. One option is by morphology. In several Romance languages, for instance, gender is morphologically marked across several grammatical categories through word endings. In contrast to these 'gender-loaded' languages, gender assignment in German is largely nontransparent and cannot be gleaned from endings alone. Several of the previous studies made use of stimuli with morphologically marked gender in Romance languages, and this probably contributed to the rather stable effect in object categorization for these languages (Bassetti, 2007; Kurinski & Sera, 2011; Kurinski et al., 2016; Sera et al., 1994, 2002). For German, on the other hand, the largely nontransparent gender assignment renders the accompanying article as the only reliable indicator of gender (for more

details, see below). Its crucial role for the emergence of effects in German is attested to in several studies (Beller et al., 2011; Imai et al., 2014; Konishi, 1994; Vigliocco et al., 2004), hinting at the possibility that, when producing behavior that is compatible with a gender congruency effect, speakers of German, too, may draw on gender information. While the strategic usage of gender information may resemble an effect of gender on categorization, it actually blurs the effect of interest, namely the immediate, unreflected impact of a noun's gender on how its referent is conceptualized.

A second alternative to this immediate impact of gender is connotations of the noun's referent on the conceptual level.

The Role of Connotations

The body of research reviewed above attests to the fact that objects may indeed be conceptualized as more male or more female, and the VAT is a particularly valuable tool for assessing such associations. One of the important—and still open—questions, though, is to what extent these associations arise from the gender of the noun as compared to biases grounded in individual, culturally, or perhaps even universal connotations, which are not part of the lexical-semantic meaning of the noun, but are linked to its referent on the conceptual level (Beller et al., 2015; Cubelli et al., 2011; Guiora & Sagi, 1978). As Nicoladis and Foursha-Stevenson put it, "it is important to test for the possibility of cultural biases *before* concluding that it is something about the structure of the language that affects thought" (2012, p.1106; emphasis added).

Evidence for the existence of such non-linguistic, conceptual connotations is manifold. For instance, a rather stable finding has been that speakers of both gendered and non-gendered languages tend to associate artifacts more strongly with male properties and natural kinds more

strongly with female properties (Sera et al., 1994, 2002; and see Bassetti, 2014; Mullen, 1990). In language development, this is present as early as age 5 and is only later partly superimposed by gender distinctions in two-gender languages. Cultural influences are also acknowledged by Flaherty (2001), who notes that before the gender system 'creeps in', people's categorizations are influenced by "perceived attributes" (p.29); by Nicoladis and Foursha-Stevenson (2012), who report biases of monolingual English speakers for classifying specific objects as either boys and girls; and most explicitly by Bassetti (2014), who discusses the possibility that grammatical gender may to some extent even be motivated by such connotations and cultural representations.

More generally, such conceptual connotations may be variable across individual speakers when reflecting personal experiences and feelings (Cubelli et al., 2011), but may also be culturally shared. They may have developed along stereotypical lines, such as when things associated with kids and kitchen are perceived as more female and hammers and axes as more male (Boroditsky et al., 2003; Guiora & Sagi, 1978; Leinbach, Hort, & Fagot, 1997); they may arise from personified allegories encountered, for instance, in fairy tales where frogs become princes, or in cultural symbols such as Lady Liberty (Bender, Beller, & Klauer, 2016b; Mills, 1986; Segel & Boroditsky, 2011); or they may have simply happened to co-occur with persons of one sex more often than the other, either in reality or in linguistic contexts (Nicoladis & Foursha-Stevenson, 2012).

With few exceptions, most previous studies did not control for such connotations. For instance, while Bassetti (2007) took care "to ensure that no object had male or female connotations, avoiding objects such as skirts or perfumes which are generally associated with women" (p. 261), the results of an item analysis still indicated that some of the twelve items they employed "may have masculine connotations in an Italian environment" (p.265). The same argument used to account for items not producing the expected effect (i.e., gender-incongruent connotations) may also account for items producing the effect (through gender-congruent connotations). And while

Kurinski and Sera (2011) concede that possible cultural influences on gender attribution should not be ignored (p.206), their materials included items that did evoke such connotations based, for instance, on stereotypical gender roles or on the color of stimuli (p.215). The authors conclude that "it is crucial to control for potential effects of culture" (p.207) and therefore confined their sample to participants from a single culture.

An even more conclusive step would be to directly control one's materials for such connotations, and harnessing the VAT with its explicit focus on sex-related associations is a viable option to achieve this goal. In two recent studies with native speakers of German, we therefore combined both explicit and implicit measures (Bender, Beller, & Klauer, 2016a, 2016b): an assignment task in which participants were asked for the biological gender they associated with a set of nouns, and the *Extrinsic Affective Simon Task*, EAST (De Houwer, 2003), which is a variant of the *Implicit Association Test*, IAT (Greenwald, McGhee, & Schwartz, 1998), but requires only one content-related categorization (for details see Figure 1 and explanations there). In an attempt to disentangle contributions of gender and of connotations to the gender congruency effect, we contrasted nouns that share gender, but differ in the strength and/or direction of their connotations.

While these studies did yield a gender congruency effect even for German, they also suggested that the effect is carried, at least to a substantial extent, by conceptual connotations as obtained in the explicit task. More concretely, items with strong gender-congruent associations produced a gender congruency effect, items with weak associations did not, and items with incongruent associations tended to reverse the effect (Bender et al., 2016a, 2016b).

However, since these two studies—like most studies in this field—only made use of masculine and feminine items, their findings remain inconclusive. When using only nouns that have either masculine or feminine gender as well as male or female associations, the possibility remains that the associations are, directly or indirectly, still driven by the grammatical gender.

After all, the associations we measure in any of our tasks may have a number of different sources, including conceptual connotations as well as grammatical gender, possibly interacting in a complex manner. Moreover, even some of those associations *proximally* originating in non-linguistic, conceptual connotations may still be indirectly inspired by the grammatical gender of the noun. For instance, liberty presumably became a lady because the feminine gender of the Latin word *libertas* facilitated portrayals of its tutelary deity as goddess.

This potential confound in stimuli cannot be resolved in two-gender languages, all the more so as associations between grammatical and biological gender are a priori more consistent in these languages than they are in languages with more than two genders. To assess the proportion of grammatical gender in such associations, in the current study, we therefore take advantage of the fact that German is a three-gender language, by including nouns that have neuter gender while still evoking male or female associations.

Overview of the Current Study

The main goal of this study is to assess the degree of direct influence that the grammatical gender of a noun may have on the conceptualization of its referent—which would constitute a genuine impact of language on thought—as compared to largely non-linguistic, conceptual connotations, while controlling for participants' deliberate usage of gender information. By giving rise to a quasi-effect of gender congruency, the deliberate usage of gender information poses challenges primarily to the empirical investigation of those other sources on which a genuine gender congruency effect may draw.

Several means may be conducive for eliminating the deliberate usage of grammatical gender as a possible confound of the gender congruency effect. One is to adopt an implicit measure (the

EAST), which in no way alludes to the decisive role of grammatical gender, and to combine it with an explicit measure (an assignment task) to check gender congruency. Another is to manipulate the extent to which information on grammatical gender is made available to begin with, by varying whether and how articles are presented in combination with the nouns.

Teasing apart grammatical gender and conceptual connotations as possible sources of the gender congruency effect is the actual challenge we aim to tackle here. To this end, we contrast gendered nouns with neuter nouns, which were picked so as to evoke either male or female associations, regardless of their grammatical gender. Keeping apart two categories of nouns (generic nouns referring to animates and nouns referring to non-animates) defines pairs of experiments.

The critical contrast: gendered versus neuter nouns. One way to disentangle the relative contributions of grammatical properties from conceptual connotations to the emergence of the gender congruency effect is by contrasting two item categories that share one of these features but differ with regard to the other. In contrast to previous studies, in which we compared targets that shared gender, but differed in the strength or direction of associations, here, we compare targets that share associations but differ in gender. More specifically, we now use items that are semantically similar and equally loaded either with strong male or female associations, while differing in whether or not they have grammatical gender: masculine and feminine nouns with strong male or female associations, respectively, on the one hand, and neuter nouns with strong male or female associations on the other. While for the former, grammatical gender and conceptual connotations remain difficult to disentangle, the latter are clear-cut: As neuter nouns, they lack the grammatical gender of interest, but may still come with strong connotations. If these neuter nouns produce an effect in the implicit task, this effect cannot, by definition, originate in grammatical gender, but will necessarily be driven by the non-linguistic, conceptual connotations; and if this

effect is of similar size to that produced by masculine and feminine nouns, we may conclude that these connotations also play a major role for the gendered nouns.

Keeping semantic categories separate. From a theoretical point of view, two categories of nouns need to be distinguished: nouns referring to non-animates (mostly objects) that have grammatical gender but no sex (henceforth abbreviated as NON-ANIMATES), and nouns referring to animates (mostly animals) that have both grammatical gender and sex, with the two being potentially in conflict, such as when one particular gender is generically used for individuals of both sexes (GENERIC ANIMATES). Although our own work did not yield diverging effects for the two categories (Bender et al., 2016a), there is a body of research pointing to such differences—most notably that an effect may emerge for animals but not objects (Cook, 2016; Saalbach et al., 2012; Vigliocco et al., 2005), and that natural kinds are more strongly associated with female properties and artifacts more strongly with male properties (Mullen, 1990; Sera et al., 1994, 2000). Although our category of GENERIC ANIMATES is, by definition, confined to natural kinds and our category of NON-ANIMATES consists largely (albeit not exclusively) of artifacts, we ensured that items in each category evoke associations in line with the particular grammatical gender.

Controlling for strategic usage of gender. As mentioned above, several studies indicate that the definite article plays a crucial role in the emergence of the gender congruency effect in German speakers (Beller et al., 2011; Imai et al., 2014; Konishi, 1994; Vigliocco et al., 2004). With a few exceptions, the assignment of grammatical gender in German is opaque. While it does follow statistical regularities to some extent (Köpcke & Zubin, 1983, 1984; Zubin & Köpcke, 1986), native speakers typically acquire their gender-related knowledge implicitly and are not aware of the underlying regularities (Corbett, 1991; Hohlfeld, 2006; Schwichtenberg & Schiller, 2004). In order to test which grammatical gender a noun has, native speakers therefore tend to use the definite article as the only reliable indicator to assess agreement. Each of the three grammatical genders

used in German is indicated by a distinct definite article in the singular (i.e., *der, die*, and *das* in the nominative case), hence for instance: *der Löffel* ("the$_{[masc]}$ spoon"), *die Gabel* ("the$_{[fem]}$ fork"), and *das Messer* ("the$_{[neuter]}$ knife"). In the plural, this distinction disappears, as all nouns, regardless of gender, take the same definite article in the nominative case, namely *die* ("the$_{[plural]}$").

In the current study, we refrain from pairing the nouns with the correct article, as this would be overly inviting for a strategic usage of grammatical gender, and we seek to prevent its usage in three steps, by making them less available and less informative. For the first of three pairs of experiments, the nouns used as stimuli are co-presented with a randomly selected article, which participants are asked to ignore as irrelevant to the task at hand. The rationale for both the random selection of the article and its postposition, which is generally at odds with German syntax, is to interfere with German speakers' dominant strategy for gender checking. In the second pair of experiments, we drop the article entirely. And in the third pair of experiments, we use nouns in the plural, which provide no information about gender. In this latter case, native speakers would first have to convert the plural form of the noun into a singular form before they then can assess its grammatical gender via the gender-specific article, which—if it happened—would be reflected in increased response times compared to the singular condition. If the absence of gender-marking articles, or the absence of gender itself, reduces the gender congruency effect, this would indicate that the effect is at least partly dependent on the article and/or the grammatical information enclosed within it.

Combining implicit and explicit measures. A second means to minimize the deliberate usage of grammatical gender is by using an implicit measure that does not allude to the decisive role of grammatical gender. We therefore combine an assignment task, as our explicit measure of associations, with an EAST, as our implicit measure. The EAST also has an important advantage over other implicit tasks such as similarity judgments: While for the latter, questions have been

raised about whether they are capable of capturing altered mental representations (Cook, 2016), the EAST is a direct, even if implicit, measure of the construct under scrutiny, as it assesses which item category (e.g., masculine or feminine nouns) is associated more strongly with male or female properties.

The basic structure is the same across all three pairs of experiments. The first part of each experiment consists of the EAST, in which participants see nouns from different categories presented in black or one of two colors and are asked for biological gender (if nouns are presented in black) or color (if nouns are presented in color). CONGRUENT ANIMATES (i.e., animates with congruent biological and grammatical gender) serve as either basic or reference category, and either GENERIC ANIMATES (Experiments 1a, 2a, and 3a) or NON-ANIMATES (Experiments 1b, 2b, and 3b) as target categories. Items of the basic category are to be categorized according to the attribute under scrutiny (i.e., the referent's sex), whereas items of the target categories and of the reference category are to be categorized according to a superficial attribute (here: the noun's color).

Assigning the categorization of the two attributes pairwise to the same two keys creates two types of trials. A *congruent trial* is defined as one in which a target of a specific grammatical gender, say feminine, is presented in a color that requests categorization by the same key as the corresponding biological gender of the reference nouns (i.e., female in the example in Figure 1), whereas in an *incongruent trial* it requests categorization by the same key as the opposing biological gender (i.e., male). A *gender congruency effect* is diagnosed when congruent trials generate more accurate and/or faster responses than incongruent trials.

---------- Insert Figure 1 about here ----------

Target categories contain gendered nouns and neuter nouns with either male or female associations. If only the grammatical gender of the noun has an effect on how its referent its conceptualized, then responses should be significantly more accurate and/or faster in congruent trials than in incongruent trials in the case of gendered nouns, but not in the case of neuter nouns (*grammar hypothesis*). If only the connotations evoked by the noun have this effect, the difference between congruent and incongruent trials should emerge for the two classes of nouns to a similar extent (*connotation hypothesis*). And if both grammatical gender and connotations have this effect, the difference should emerge for both classes of nouns, but should be larger for the gendered nouns than for the neuter nouns (*grammar* × *connotation hypothesis*).

The implicit measure is complemented by an explicit measure based on the VAT in the second part of each experiment. For each of the nouns used as target in the EAST, participants are asked whether they would assign a male or female voice to its referent (for the GENERIC ANIMATES) or make associations to men or women (for the NON-ANIMATES).

Ethics. Our research project was guided by the ethical principles as formulated in the WMA Declaration of Helsinki. Ethics approval for social science research is not required in Germany if research objectives do not involve issues regulated by law (typically medical research). Our studies have no such objectives, and therefore, no IRB approval or waiver of permission was sought for these studies.

Experiments 1a and 1b

The experiments aimed at testing the extent to which the emergence of the gender congruency effect in German depends on gender class (specifically, whether it is confined to masculine and feminine nouns, or may also emerge for neuter nouns), as compared to conceptual

connotations. To interfere with the dominant strategy for gender checking among German speakers, the nouns used as stimuli are co-presented with a randomly selected article. As Experiments 1a and 1b followed the exact same protocol except for the stimuli used as target category, they will be described here jointly.

Method

Participants. A total of 37 native speakers of German[2] from the Freiburg area (24 female; age *M* = 25.68 years, *range:* 19 to 44, *SD* = 5.89) participated in Experiment 1a, and 40 (26 female; age *M* = 22.98 years, *range:* 18 to 33, *SD* = 3.63) participated in Experiment 1b. They were rewarded with up to 6.92 Euro (Experiment 1a) and 4.92 (Experiment 1b), contingent on the number of correct and fast decisions made within 800 ms.

Material. Nouns used as stimuli belonged either to a basic or reference category (i.e., CONGRUENT ANIMATES) or to one of two target categories (Experiment 1a: GENERIC ANIMATES, Experiment 1b: NON-ANIMATES). The basic category consisted of first names, 40 for men and 40 for women. The reference category consisted of 20 pairs of masculine and feminine kin terms and other nouns with lexical gender. The two target sets each consisted of 10 pairs of masculine and feminine items with male and female associations, respectively, and of 10 pairs of neuter items with either male or female associations (all items were rated for associations in previous studies). All nouns and ratings are reported in Appendix A(I).

Masculine and feminine CONGRUENT ANIMATES did not differ in the average number of letters (masculine: 5.30 vs. feminine: 4.85; $t(38) = 1.015$; $p = .317$).

---

[2] Due to the German educational system, all participants had learned English, but additional knowledge of (non-gendered) English does not affect the gender-related performance of participants mastering a gendered language (Bassetti, 2014; Boroditsky et al., 2003; Ervin, 1962; Forbes et al., 2008; Kurinski & Sera, 2007).

The four groups of GENERIC ANIMATES (grammatically gendered vs. neuter, with male vs. female associations) were checked for differences in word length, frequency of usage according to the SUBTLEX-DE database for word frequencies (Brysbaert, Buchmeier, Conrad, Jacobs, Bölte, & Böhl, 2011), and strength of associations by means of analyses of variance with the two factors *grammatical gender* (gendered vs. neuter) and *association* (male vs. female). We found no differences in the average number of letters ($m = 5.52$; all $F(1,36) = 0.012$; $p = .915$; $\eta_p^2 < .001$) and no differences in the frequency of usage ($m = 1.75$; all $F(1,36) \leq 1.250$; $p \geq .271$; $\eta_p^2 \leq .034$). For the associations, which were always coded from 1 (female) to 4 (male), the only significant effect was the expected main effect of *association* (male: 3.18 vs. female: 1.78; $F(1,36) = 343.4$; $p < .001$; $\eta_p^2 = .905$). When reversing the polarity of the ratings for the female-associated items, which enabled us to compare the absolute strength of the associations, this main effect disappeared (male: 3.18 vs. female: 3.22; $F(1,36) = 0.329$; $p = .570$; $\eta_p^2 = .009$), indicating that associations to males and females were equally strong.

The four groups of NON-ANIMATES were checked in the same way. We found no differences in the average number of letters ($m = 6.00$; all $F(1,36) \leq 1.166$; $p \geq .287$; $\eta_p^2 \leq .031$) and no differences in the frequency of usage ($m = 1.97$; all $F(1,36) \leq 3.604$; $p \geq .066$; $\eta_p^2 \leq .091$), but the expected main effect of the associations (male: 3.52 vs. female: 1.48; $F(1,36) = 1961.2$; $p < .001$; $\eta_p^2 = .982$). With reversed polarity of the ratings for the female-associated items, this effect disappeared (male: 3.52 vs. female: 3.52; $F(1,36) = 0.000$; $p = 1.0$; $\eta_p^2 = .000$), indicating that associations to males and females were equally strong.

Design and Procedure. The experiment always began with the EAST (Part I), followed by the explicit rating task (Part II).

Part I. For the EAST, stimuli were presented in black, blue, or green color, together with a randomly selected, post-positioned definite article (*der, die,* or *das*). The categorization task

depended on category membership: Items in black (basic category) had to be categorized according to their biological gender. Items in blue or green (one half reference category = CONGRUENT ANIMATES and one half target category = GENERIC ANIMATES [Exp. 1a] or NON-ANIMATES [Exp. 1b]); these had to be categorized according to their color. Responses were made by pressing the inner keys of two computer mice placed to the left and right of the keyboard, respectively. The assignment of computer mouse to gender response (e.g., left = male, right = female) and of computer mouse to color response (e.g., left = blue, right = green) was independently randomized across participants.

Targets were displayed in the centre of a 58.4-cm LCD screen with a 100 Hz refresh rate, subtending about 5º of visual angle horizontally and 1º vertically. They were presented in random order for 0.8 seconds each. During all trials, labels ("male"/"female" in black, and "blue"/"green" in blue/green respectively) were present in the bottom-left and -right corner (Figure 1).

In both experiments, participants were tested individually in a within-subject design. Following instructions on-screen, participants worked on a set of 16 practice items with trial-wise feedback. If a response was incorrect, a red X appeared below the stimulus' position, and the correct response had to be entered to continue (cf. Greenwald, McGhee, & Schwartz, 1998). The same feedback was given in the test trials. Test trials consisted of three complete blocks of 160 items each (80 basic category, 40 reference category, and 40 target category). For each category in the color categorization task, each combination of gender and color occurred equally often, but was otherwise randomized anew for each block.

Part II. The rating task subsequent to the EAST consisted of a list of nouns, including the nouns of the target categories, for which participants had to indicate whether they would assign a male or female voice to its referent (Experiment 1a: GENERIC ANIMATES) or whether its referent evokes associations to men or women (Experiment 1b: NON-ANIMATES). Instructions in the style of

Sera and colleagues (2002) were provided, illustrated here for Experiment 1a with GENERIC ANIMATES (the German translation of all rating instructions is provided in Appendix A[I]):

> *For a cartoon film, various animate beings will be used as protagonists, which you will find listed in the following. For each of these animate beings, please tick whether a female or a male voice fits better. If you are unsure, you may attenuate your judgment somewhat ("more female voice" or "more male voice").*
>
> *If, for instance, you consider a guinea pig to be female, please tick "female voice"; if you consider a groundhog to be more male than female, please tick "more male voice"; and if you consider a hippo to be fairly male, please tick "male voice".*
>
> *Important: Please always tick exactly one of the four options for each animate being. You can correct your choice at any time.*

Each item had the same four response options (from left to right): *female voice, more female voice, more male voice,* and *male voice* in Experiment 1a, and *clearly woman, more woman, more man,* and *clearly man* in Experiment 1b, respectively. To control for order effects, the items were presented in a randomized order.

Results and Discussion

As explained in the introduction, if a gender congruency effect is based on grammatical gender only, it should emerge for gendered nouns but not for neuter nouns. If it is based on connotations only, it should emerge for both, and should be of a similar extent. And if is based on grammatical gender and connotations, it should also emerge for both, but should be weaker for neuter nouns. In view of our previous findings, we expected a gender congruency effect for all our

target items alike (in line with the connotation hypothesis), and we thus expected that the effect would be mediated by the strength of the association. In the following, the results of the explicit rating tasks (Part II) are presented first as they pertain to the material used in the EAST (Part I), which is presented and discussed subsequently.

Explicit measure: Rating tasks (Part II). The ratings that we obtained for the target items in Experiment 1 correlated nearly perfectly with those from the material selection, both for the GENERIC ANIMATES in Experiment 1a ($r = .972$; $p < .001$; $N = 40$) and for the NON-ANIMATES in Experiment 1b ($r = .992$; $p < .001$; $N = 40$); the mean ratings for all items are provided in Appendix A(I). As for the material selection, the four groups of target items were checked for differences in the strength of associations by means of analyses of variance with the two factors *grammatical gender* (gendered vs. neuter) and *association* (male vs. female).

For the GENERIC ANIMATES in Experiment 1a, the only significant effect was the expected main effect of the associations (male: 3.21 vs. female: 1.93; $F(1,36) = 209.2$; $p < .001$; $\eta_p^2 = .853$). With reversed polarity of the ratings for the female-associated items, this effect disappeared (male: 3.21 vs. female: 3.07; $F(1,36) = 2.369$; $p = .132$; $\eta_p^2 = .062$), indicating that associations to males and females were equally strong across gendered and neuter nouns.

For the NON-ANIMATES in Experiment 1b, we also found the expected main effect of the associations (male: 3.44 vs. female: 1.53; $F(1,36) = 1013.6$; $p < .001$; $\eta_p^2 = .966$), but in addition—and deviating from the pretest of the material—we found an interaction with grammatical gender; $F(1,36) = 7.690$; $p = .009$; $\eta_p^2 = .176$. The analysis with reversed polarity of the ratings for the female-associated items indicated that the interaction was driven by a difference in the absolute strength of the associations between gendered and neuter nouns. For the current sample, the associations of gendered nouns were significantly stronger than those of neuter nouns (gendered: 3.54 vs. neuter: 3.37; $F(1,36) = 7.690$; $p = .009$; $\eta_p^2 = .176$), whereas the two types of

association did not differ in absolute strength (male: 3.44 vs. female: 3.47; $F(1,36) = 0.418$; $p = .522$; $\eta_p^2 = .011$).

Overall, these results validate our selection of target items both for the GENERIC ANIMATES and NON-ANIMATES, with strong differences in the explicit ratings of associations to male versus female characteristics. For the NON-ANIMATES (Experiment 1b), however, they also indicate an imbalance between gendered and neuter nouns.

Implicit measure: EAST (Part I). Using Tukey's criterion for extreme outliers (extreme values being values three times the interquartile range below the first or above the third quartile; Clark-Carter, 2004, Chap. 9), we first examined whether any participant was an extreme outlier in terms of percentage of correct responses or mean correct-response reaction time (RT). In Experiment 1a, two participants (below 51% correct in a sample with $M = 93\%$, $SD = 12\%$) were excluded for the analyses of both accuracy and reaction time data, and one further participant (mean correct-response RT 1317 ms in a sample with $M = 525$ ms, $SD = 155$) was excluded for the analysis of the reaction time data. In Experiment 1b, one participant (53% correct in a sample with $M = 91\%$ correct, $SD = 7\%$) was excluded for the analyses of both accuracy and reaction time data.

The data were then analyzed as follows (see Bender, Beller, & Klauer, 2016a, 2016b). First, we estimated mixed linear models (for the accuracy data: generalized mixed linear models with logistic link function) with *participants* and *target items* as random factors in order to determine the best fitting random effects structure (Jaeger, 2008; Judd, Westfall, & Kenny, 2012). The strategy for selecting a model with the most appropriate structure and the final model used for the subsequent analyses are reported in Appendix B. Based on the final model, we then checked the fixed effects of three within-participant factors: *type* of target (CONGRUENT ANIMATE vs. gendered target vs. neuter target [with target in Experiment 1a: GENERIC ANIMATE; Experiment 1b: NON-ANIMATE]), *gender association* of target (male vs. female), and *response association* (whether

the trial required pressing the "male" or "female" key)[3]. A significant interaction *gender* × *response association* indicates gender congruency. Delta chi-square statistics are used for the accuracy data and *F* statistics with Kenward-Roger approximated degrees of freedom (according to Judd, Westfall, & Kenny, 2012) for the reaction time data; these tests were two-tailed. In addition, we checked each noun category for a gender congruency effect; these tests were one-tailed in line with the directed gender congruency hypothesis. Finally, we re-analyzed the data with the mean rating of the targets (from the explicit task; centered on the scale midpoint 2.5 and scaled to a standard deviation of 1.0) and its interaction with response association as covariate in order to check whether or not the observed gender congruency effects are mediated by the explicit ratings of biological gender. Again, we first determined the best fitting model with the most appropriate structure of random effects (reported in Appendix B), and then used this model to test the fixed effects. All analyses were conducted in the programming language R (R Core Team, 2014) using the packages *lme4* (Bates, Maechler, Bolker, & Walker, 2014) and *afex* (Singmann, 2014).

Figure 2 and Appendix C summarize the findings for Experiments 1a and 1b, and Appendix B(I and II) reports the model comparisons; the fixed effects and the results for the noun categories are presented in the following.

---------- Insert Figure 2 about here ----------

Experiment 1a (GENERIC ANIMATES): For the accuracy data, the analysis of the fixed effects indicated a general gender congruency effect (gender × response association: $\chi^2[df = 1] = 66.63$;

[3] As indicated in Appendix B, the final model in all of the reported analyses includes random slopes for response association as a function of *participants*. This indicates inter-individual differences with regard to the preference for the "male" or "female" response key, perhaps due to an interaction between the participant's handedness and the balanced assignment of the response keys.

$p < .001$) that interacted with the *type* of targets (type × gender × response association: $\chi^2[df = 2] = 32.10$; $p < .001$), and an interaction *type* × *response association* ($\chi^2[df = 2] = 6.92$; $p = .031$). Testing the congruency effect for the three noun categories revealed a significant effect for CONGRUENT ANIMATES ($M = 10.67$ %; $\chi^2[df = 1] = 137.04$; $p < .001$; Cohen's $d_z = 1.27$) and for the gendered GENERIC ANIMATES ($M = 4.29$ %; $\chi^2[df = 1] = 13.18$; $p < .001$; $d_z = 0.61$), but not for the neuter GENERIC ANIMATES ($M = 1.52$ %; $\chi^2[df = 1] = 1.85$; $p = .087$; $d_z = 0.23$). However, the difference between gendered and neuter GENERIC ANIMATES did not reach statistical significance (type × gender × response association: $\chi^2[df = 1] = 2.26$; $p = .133$). The general gender congruency effect observed in the initial analysis disappeared when the ratings were included as covariates (gender × response association: $\chi^2[df = 1] = 0.02$; $p = .896$), and was independent of the *type* of target (type × gender × response association: $\chi^2[df = 2] = 2.03$; $p = .362$). The covariate *rating* × *response association* was significant ($\chi^2[df = 1] = 5.26$; $p = .022$), and its inclusion effectively eliminated the gender congruency effect, consistent with the argument that the effect is mediated by the explicit ratings of biological gender.

For the reaction time data, the analysis of the fixed effects indicated a general gender congruency effect (gender × response association: $F(1,7077.1) = 7.13$; $p = .008$), but no interaction with the *type* of target (type × gender × response association: $F(2,7076.8) = 1.06$; $p = .35$). Testing the congruency effect for the three noun categories revealed a significant effect for CONGRUENT ANIMATES ($M = 12.69$ ms; $F(1,3462.3) = 9.05$; $p = .002$; $d_z = 0.34$), but no effect for the gendered GENERIC ANIMATES ($M = 7.12$ ms; $F(1,1728.6) = 2.47$; $p = .06$; $d_z = 0.19$), and no effect for the neuter GENERIC ANIMATES ($M = 1.29$ ms; $F(1,1764.8) = 0.27$; $p = .30$; $d_z = 0.03$). The general gender congruency effect observed in the initial analysis disappeared when the ratings were included as covariates (gender × response association: $F(1,7071.2) = 0.81$; $p = .37$), and was independent of the *type* of target (type × gender × response association: $F(2,7075.6) = 0.63$;

$p = .53$). The covariates were not significant, but nevertheless eliminated the gender congruency effect.

Experiment 1b (NON-ANIMATES): For the accuracy data, the analysis of the fixed effects indicated a general gender congruency effect (gender × response association: $\chi^2[df = 1] = 67.56$; $p < .001$) that interacted with the *type* of target (type × gender × response association: $\chi^2[df = 2] = 6.30$; $p = .043$), and an interaction *type* × *response association* ($\chi^2[df = 2] = 7.45$; $p = .024$). Testing the congruency effect for the three noun categories revealed a significant effect for CONGRUENT ANIMATES ($M = 6.84$ %; $\chi^2[df = 1] = 77.91$; $p < .001$; $d_z = 0.94$), for the gendered NON-ANIMATES ($M = 3.33$ %; $\chi^2[df = 1] = 9.73$; $p < .001$; $d_z = 0.49$), and for the neuter NON-ANIMATES ($M = 3.85$ %; $\chi^2[df = 1] = 14.51$; $p < .001$; $d_z = 0.56$). The difference between gendered and neuter NON-ANIMATES was not significant (type × gender × response association: $\chi^2[df = 1] = 0.39$; $p = .532$). The general gender congruency effect observed in the initial analysis disappeared when the ratings were included as covariates (gender × response association: $\chi^2[df = 1] = 0.02$; $p = .893$), and was independent of the *type* of target (type × gender × response association: $\chi^2[df = 1] = 0.79$; $p = .375$). The covariates were not significant, but nevertheless eliminated the gender congruency effect.

For the reaction time data, the analysis of the fixed effects indicated a general gender congruency effect (gender × response association: $F(1,8339.2) = 14.15$; $p < .001$), but no interaction with the *type* of target (type × gender × response association: $F(2,8338.6) = 0.19$; $p = .83$). Testing the congruency effect for the three noun categories revealed a significant effect for CONGRUENT ANIMATES ($M = 9.89$ ms; $F(1,4099.1) = 9.52$; $p = .001$; $d_z = 0.34$), no effect for the gendered NON-ANIMATES ($M = 6.90$ ms; $F(1,2040.2) = 2.42$; $p = .06$; $d_z = 0.22$), but a significant effect for the neuter NON-ANIMATES ($M = 10.53$ ms; $F(1,2058.0) = 4.66$; $p = .015$; $d_z = 0.35$). The general gender congruency effect observed in the initial analysis disappeared when the ratings were

included as covariates (gender × response association: $F(1,4170.7) = 0.00$; $p = .97$), and was independent of the *type* of target (type × gender × response association: $F(1,4168.8) = 0.41$; $p = .52$). The covariates were not significant, but nevertheless eliminated the gender congruency effect.

Taken together, the findings from the first pair of experiments suggest the following conclusions. The congruency effect was generally greatest and most consistent for the CONGRUENT ANIMATES and emerged both for accuracy and reaction time (this is unsurprising, given that for items of this category grammatical gender and biological sex coincide). For the GENERIC ANIMATES, a gender congruency effect occurred only in accuracy and only for the gendered nouns (not for the neuter nouns), which would be consistent with the hypothesis that the effect is based on grammar only (grammar hypothesis). On the other hand, for the NON-ANIMATES, a congruency effect of similar size occurred in accuracy both for gendered and neuter nouns, and additionally in reaction time for neuter, but not gendered nouns, which would be consistent with the hypothesis that the effect is based on connotations only (connotation hypothesis). Consistent with the latter hypothesis, all congruency effects disappeared when the explicit ratings were included in the analyses as covariate. This is largely in line with what we observed in previous studies for CONGRUENT ANIMATES as well as gendered GENERIC ANIMATES and NON-ANIMATES with randomly assigned articles in post-position (Bender, Beller, & Klauer, 2016a, Exp. 3).

While the articles were primarily intended to discourage deliberate gender-checking strategies among our participants, both their random selection and non-canonical position may have been confusing to participants. To address these concerns, the experiments were repeated without articles.

Experiments 2a and 2b

Experiments 2a and 2b tested whether the randomly assigned article in post-position may have distorted the results, by dropping the article. If the effect remained largely unchanged, a distortion could be regarded as unlikely.

Method

Participants. A total of 39 native speakers of German from the Freiburg area (23 female; age $M = 25.36$ years, *range:* 18 to 40, $SD = 5.08$) participated in Experiment 2a, and 39 (26 female; age $M = 26.51$ years, *range:* 19 to 44, $SD = 6.73$) participated in Experiment 2b. They were rewarded with up to 4.92 Euro, contingent on the number of correct and fast decisions made within 800 ms. None of them had participated in the previous experiments.

Material. The material was the same as in Experiments 1a and 1b, respectively, except that stimuli in the EAST were not accompanied by an article and the corresponding sentence in the instructions was therefore dropped.

Design and Procedure. These were the same as in Experiments 1a and 1b, respectively.

Results and Discussion

The data were analyzed in the same way as described for the previous experiments.

Explicit measure: Rating tasks (Part II). The ratings that we obtained for the target items in Experiment 2 correlated strongly with those from the material selection (GENERIC ANIMATES: $r = .973$; NON-ANIMATES: $r = .995$) and those obtained in Experiment 1 (GENERIC ANIMATES: $r = .988$; NON-ANIMATES: $r = .996$; all $p < .001$; $N = 40$); the mean ratings are provided in Appendix A(II).

As for Experiment 1, the different groups of items were checked for differences in the strength of associations by means of analyses of variance with the two factors *grammatical gender* (gendered vs. neuter) and *association* (male vs. female).

For the GENERIC ANIMATES in Experiment 2a, the only significant effect was the expected main effect of the associations (male: 3.25 vs. female: 1.91; $F(1,36) = 307.4$; $p < .001$; $\eta_p^2 = .895$). With reversed polarity of the ratings for the female-associated items, this effect did not disappear completely, and indicated that the absolute associative strength was a little greater for items with male associations than for those with female associations (male: 3.25 vs. female: 3.09; $F(1,36) = 4.512$; $p = .041$; $\eta_p^2 = .111$).

For the NON-ANIMATES in Experiment 2b, we also found the expected main effect of the associations (male: 3.53 vs. female: 1.44; $F(1,36) = 1413.6$; $p < .001$; $\eta_p^2 = .975$), but, similar to Experiment 1b, also an interaction with grammatical gender; $F(1,36) = 6.547$; $p = .015$; $\eta_p^2 = .154$. According to the analysis with reversed polarity of the ratings for the female-associated items, the absolute strength of associations was significantly greater for gendered nouns than for neuter nouns (gendered: 3.62 vs. neuter: 3.47; $F(1,36) = 6.547$; $p = .015$; $\eta_p^2 = .154$) whereas the two types of association did not differ in strength (male: 3.53 vs. female: 3.56; $F(1,36) = 0.154$; $p = .697$; $\eta_p^2 = .004$).

Overall, these results replicate the findings of Experiment 1 and revalidate our selection of target items, with small imbalances both for GENERIC ANIMATES (Experiment 2a) and NON-ANIMATES (Experiment 2b).

Implicit measure: EAST (Part I). In Experiment 2b, one participant (mean correct-response RT 841 ms in a sample with $M = 490$ ms, SD = 92) was excluded from the reaction time analyses. The findings for Experiments 2a and 2b are summarized in Figure 3 and

Appendix C, and the model comparisons are reported in Appendix B(III and IV); the results for the fixed effects and for the noun categories are described below.

---------- Insert Figure 3 about here ----------

Experiment 2a (GENERIC ANIMATES): For the accuracy data, the analysis of the fixed effects indicated a general gender congruency effect (gender × response association: $\chi^2[df = 1] = 82.39$; $p < .001$) that interacted with the *type* of target (type × gender × response association: $\chi^2[df = 2] = 10.77$; $p = .005$). Testing the congruency effect for the three noun categories revealed a significant effect for CONGRUENT ANIMATES ($M = 8.55$ %; $\chi^2[df = 1] = 106.16$; $p < .001$; $d_z = 1.52$), for the gendered GENERIC ANIMATES ($M = 4.27$ %; $\chi^2[df = 1] = 17.48$; $p < .001$; $d_z = 0.59$), and for the neuter GENERIC ANIMATES ($M = 3.68$ %; $\chi^2[df = 1] = 10.14$; $p < .001$; $d_z = 0.61$). The difference between gendered and neuter GENERIC ANIMATES was not significant (type × gender × response association: $\chi^2[df = 1] = 0.69$; $p = .406$). As in Experiment 1a, the general gender congruency effect observed in the initial analysis disappeared when the ratings were included as covariates (gender × response association: $\chi^2[df = 1] = 0.13$; $p = .718$), and was independent of the *type* of target (type × gender × response association: $\chi^2[df = 2] = 0.54$; $p = .763$). The covariate *rating* × *response association* was significant ($\chi^2[df = 1] = 6.52$; $p = .011$). Its inclusion thus effectively eliminated the gender congruency effect, consistent with the argument that the effect is mediated by the explicit ratings of biological gender.

For the reaction time data, the analysis of the fixed effects indicated a general gender congruency effect (gender × response association: $F(1,8186.1) = 28.46$; $p < .001$), but no interaction with the *type* of target (type × gender × response association: $F(2,8180.6) = 1.82$; $p = .16$). Testing the congruency effect for the three noun categories revealed a significant effect

for CONGRUENT ANIMATES (M = 19.60 ms; $F(1,4007.6) = 31.09$; $p < .001$; $d_z = 0.63$), for the gendered GENERIC ANIMATES (M = 12.47 ms; $F(1,2022.4) = 6.00$; $p = .005$; $d_z = 0.33$), and for the neuter GENERIC ANIMATES (M = 9.01 ms; $F(1,2002.9) = 3.53$; $p = .03$; $d_z = 0.27$). As in Experiment 1a, the general gender congruency effect observed in the initial analysis disappeared when the ratings were included as covariates (gender × response association: $F(1,8184.3) = 0.00$; $p = .96$), and was independent of the *type* of target (type × gender × response association: $F(2,8179.7) = 0.07$; $p = .93$). The covariates were not significant, but nevertheless eliminated the gender congruency effect.

Experiment 2b (NON-ANIMATES): For the accuracy data, the analysis of the fixed effects indicated a general gender congruency effect (gender × response association: $\chi^2[df = 1] = 70.61$; $p < .001$) that interacted with the *type* of target (type × gender × response association: $\chi^2[df = 2] = 24.28$; $p < .001$). Testing the congruency effect for the three noun categories revealed a significant effect for CONGRUENT ANIMATES (M = 8.08 %; $\chi^2[df = 1] = 109.28$; $p < .001$; $d_z = 1.12$), and for the gendered NON-ANIMATES (M = 5.38 %; $\chi^2[df = 1] = 28.74$; $p < .001$; $d_z = 0.68$), but not for the neuter NON-ANIMATES (M = 0.94 %; $\chi^2[df = 1] = 0.736$; $p = .195$; $d_z = 0.19$). In contrast to Experiment 1b, the difference between gendered and neuter NON-ANIMATES was significant (type × gender × response association: $\chi^2[df = 1] = 10.32$; $p = .001$). The general gender congruency effect observed in the initial analysis disappeared when the ratings were included as covariates (gender × response association: $\chi^2[df = 1] = 2.12$; $p = .145$), but the interaction with the *type* of target remained (type × gender × response association: $\chi^2[df = 1] = 11.11$; $p < .001$). The covariates were not significant, and could not completely eliminate the gender congruency effect.

For the reaction time data, the analysis of the fixed effects indicated a general gender congruency effect (gender × response association: $F(1,8107.0) = 36.65$; $p < .001$), but no interaction with the *type* of target (type × gender × response association: $F(2,8106.2) = 0.73$;

p = .48). Testing the congruency effect for the three noun categories revealed a significant effect for CONGRUENT ANIMATES (M = 17.64 ms; F(1,3971.5) = 28.52; p < .001; d_z = 1.05), for the gendered NON-ANIMATES (M = 11.49 ms; F(1,1990.2) = 7.31; p = .004; d_z = 0.43), and for the neuter NON-ANIMATES (M = 13.96 ms; F(1,1999.8) = 9.07; p = .002; d_z = 0.49). As in Experiment 1b, the general gender congruency effect observed in the initial analysis disappeared when the ratings were included as covariates (gender × response association: F(1,4040.4) = 0.73; p = .39), and was independent of the *type* of target (type × gender × response association: F(1,4056.5) = 0.01; p = .93). The covariates were not significant, but nevertheless eliminated the gender congruency effect.

Taken together, the findings from the second pair of experiments suggest the following conclusions. The rather strong and consistent congruency effect for CONGRUENT ANIMATES was replicated. For the GENERIC ANIMATES, the effect occurred in accuracy and reaction time, both for gendered and neuter nouns, and was of similar extent, thus consistent with the connotation hypothesis. For the NON-ANIMATES, the congruency effect occurred in reaction time both for gendered and neuter nouns, also consistent with the connotation hypothesis, but in accuracy only for the gendered nouns, which would be consistent with the grammar hypothesis. Except for this latter case, the congruency effect disappeared when the explicit ratings were included in the analyses as covariate, again consistent with the connotation hypothesis.

Especially for the GENERIC ANIMATES, reaction time was shorter in Experiment 2a than in Experiment 1a. This may indicate that the randomly assigned article slowed down the processing of the stimuli in Experiment 1a, but did not distort the categorization patterns. Still, even in the current trials without accompanying article, participants may have used grammatical gender for strategic purposes by *retrieving* the fitting article. The fact that neuter items generated a congruency effect of similar size to the gendered items makes this less likely, but preventing

gender-checking in the first place would provide a more convincing argument. We therefore repeated the experiments with target nouns in the plural, which lack grammatical gender altogether.

Experiments 3a and 3b

Experiments 3a and 3b tested whether using nouns *without* grammatical gender (i.e., in their plural forms) still generate a gender congruency effect.

Method

Participants. A total of 40 native speakers of German from the Freiburg area (28 female; age $M = 23.15$ years, *range:* 18 to 45, $SD = 4.74$) participated in Experiment 3a, and 40 (29 female; age $M = 24.42$ years, *range:* 19 to 43, $SD = 4.02$) participated in Experiment 3b. They were rewarded with up to 4.92 Euro, contingent on the number of correct and fast decisions made within 800 ms. None of them had participated in the previous experiments.

Material. The set of items for the basic category was the same as in the previous experiments. The items for the other categories, however, had to be adapted due to the fact that not all plural forms in German differ from the singular. For instance, *Messer* ("knife") can be both singular (*das Messer*) and plural (*die Messer*). For the stimuli to lack information on grammatical gender, it was crucial that they were recognizable as plural terms. We therefore replaced a number of items with new ones that conformed to the requirement of distinct plural forms in addition to the requirements described above (for a full list, see Appendix A[III]).

The four groups of plural GENERIC ANIMATES were again checked for differences in word length, frequency of usage, and strength of associations as described for Experiments 1a and 1b. We found no differences in the average number of letters ($m = 6.40$; all $F(1,36) \leq 3.161$; $p \geq .084$;

$\eta_p^2 \leq .081$) and no differences in the frequency of usage ($m = 1.43$; all $F(1,36) \leq 0.056$; $p \geq .814$; $\eta_p^2 \leq .002$). With regard to the strength of associations, the analyses were based on rating data from the respective singular forms. An analysis of variance with the two factors *grammatical gender* (gendered vs. neuter) and *association* (male vs. female) revealed the expected difference between male and female associations (male: 3.25 vs. female: 1.83; $F(1,35)$[4] $= 338.8$; $p < .001$; $\eta_p^2 = .906$), but also an interaction with grammatical gender; $F(1,35) = 9.676$; $p = .004$; $\eta_p^2 = .217$. The analysis with reversed polarity of the ratings for the female-associated items indicated that the interaction was driven by differences in the absolute strength of the associations between gendered and neuter nouns. The associations of gendered nouns were significantly stronger than those of neuter nouns (gendered: 3.33 vs. neuter: 3.09; $F(1,35) = 9.676$; $p = .004$; $\eta_p^2 = .217$), whereas the two types of association did not differ in strength (male: 3.25 vs. female: 3.17; $F(1,35) = 1.136$; $p = .294$; $\eta_p^2 = .031$).

The four groups of plural NON-ANIMATES were checked in the same way as described for the GENERIC ANIMATES. We found no differences in the average number of letters ($m = 6.55$; all $F(1,36) \leq 0.672$; $p \geq .418$; $\eta_p^2 \leq .018$) and no differences in the frequency of usage ($m = 1.61$; all $F(1,36) \leq 0.917$; $p \geq .345$; $\eta_p^2 \leq .025$), but observed the expected main effect of the associations for the singular forms (male: 3.48 vs. female: 1.56; $F(1,36) = 1000.4$; $p < .001$; $\eta_p^2 = .965$). With reversed polarity of the ratings for the female-associated items, this effect disappeared (male: 3.48 vs. female: 3.44; $F(1,36) = 0.002$; $p = .961$; $\eta_p^2 = .008$), indicating that associations to males and females were equally strong across the gendered and neuter nouns.

---

[4] For one neuter noun, “Ponys”, we had no association ratings from a pre-study; therefore, only 39 items were included in these analyses.

Design and Procedure. These were the same as in the previous experiments, except for a modification in the instructions to reflect the fact that items were now presented in the plural (see Appendix A[III]).

Results and Discussion

The data were analyzed in the same way as described for the previous experiments.

Explicit measure: Rating tasks (Part II). The ratings that we obtained for the target items in Experiment 3 (plural forms) correlated strongly with the ratings of the singular forms from the material selection, both for the GENERIC ANIMATES in Experiment 3a ($r = .963$; $p < .001$; $N = 39$) and for the NON-ANIMATES in Experiment 3b ($r = .986$; $p < .001$; $N = 40$); the mean ratings for all items are provided in Appendix A(III). The different groups of items were again checked for differences in the strength of associations by means of analyses of variance with the two factors *grammatical gender* (gendered vs. neuter) and *association* (male vs. female).

For the GENERIC ANIMATES in Experiment 3a, we found the expected main effect of the associations (male: 3.22 vs. female: 1.99; $F(1,36) = 156.8$; $p < .001$; $\eta_p^2 = .813$); the interaction with grammatical gender almost reached significance; $F(1,36) = 4.103$; $p = .050$; $\eta_p^2 = .102$. The analysis with reversed polarity of the ratings for the female-associated items indicated that the absolute associative strength was a little greater for items with male associations than for those with female associations (male: 3.22 vs. female: 3.01; $F(1,36) = 4.635$; $p = .038$; $\eta_p^2 = .114$), and a little greater for gendered items than for neuter items (gendered: 3.21 vs. neuter: 3.02; $F(1,36) = 4.103$; $p = .050$; $\eta_p^2 = .102$).

For the NON-ANIMATES in Experiment 3b, the only significant effect was the expected main effect of the associations (male: 3.39 vs. female: 1.69; $F(1,36) = 649.8$; $p < .001$; $\eta_p^2 = .948$). With

reversed polarity of the ratings for the female-associated items, this effect disappeared (male: 3.39 vs. female: 3.31; $F(1,36) = 1.629$; $p = .210$; $\eta_p^2 = .043$), indicating that associations to males and females were equally strong across gendered and neuter nouns.

Overall, these results validate our selection of plural target items, but also indicate small imbalances for GENERIC ANIMATES (Experiment 3a).

Implicit measure: EAST (Part I). The findings for Experiments 3a and 3b are summarized in Figure 4 and Appendix C, and the model comparisons are reported in Appendix B(V and VI); the results for the fixed effects and for the noun categories are described below.

---------- Insert Figure 4 about here ----------

Experiment 3a (GENERIC ANIMATES): For the accuracy data, the analysis of the fixed effects indicated a general gender congruency effect (gender × response association: $\chi^2[df = 1] = 63.42$; $p < .001$) that interacted with the *type* of target (type × gender × response association: $\chi^2[df = 2] = 11.96$; $p = .003$), and an interaction *type* × *response association* ($\chi^2[df = 2] = 6.19$; $p = .045$). Testing the congruency effect for the three noun categories revealed a significant effect for CONGRUENT ANIMATES ($M = 7.38$ %; $\chi^2[df = 1] = 95.00$; $p < .001$; $d_z = 1.26$), for the gendered GENERIC ANIMATES ($M = 3.00$ %; $\chi^2[df = 1] = 9.89$; $p = .001$; $d_z = 0.41$), and for the neuter GENERIC ANIMATES ($M = 2.50$ %; $\chi^2[df = 1] = 8.27$; $p = .002$; $d_z = 0.36$). The difference between gendered and neuter GENERIC ANIMATES was not significant (type × gender × response association: $\chi^2[df = 1] = 0.02$; $p = .888$). As in Experiments 1a and 2a, the general gender congruency effect observed in the initial analysis disappeared when the ratings were included as covariates (gender × response association: $\chi^2[df = 1] = 0.35$; $p = .554$), and was independent of the *type* of target (type × gender × response association: $\chi^2[df = 2] = 0.62$; $p = .733$). The covariate *rating* × *response association* was

significant ($\chi^2[df = 1] = 10.90$; $p = .001$). Its inclusion thus effectively eliminated the gender congruency effect.

For the reaction time data, the analysis of the fixed effects indicated a general gender congruency effect (gender × response association: $F(1,8596.2) = 12.33$; $p < .001$), no interaction with the *type* of target (type × gender × response association: $F(2,8596.1) = 0.75$; $p = .47$), but an interaction *type* × *response association*; $F(2,8596.2) = 5.20$; $p = .006$. Testing the congruency effect for the three noun categories revealed a significant effect for CONGRUENT ANIMATES ($M = 12.07$ ms; $F(1,4212.6) = 12.97$; $p < .001$; $d_z = 0.47$) and for the gendered GENERIC ANIMATES ($M = 8.76$ ms; $F(1,2116.6) = 3.41$; $p = .03$; $d_z = 0.28$), but not for the neuter GENERIC ANIMATES ($M = 5.98$ ms; $F(1,2114.1) = 1.38$; $p = .12$; $d_z = 0.22$). As in Experiments 1a and 2a, the general gender congruency effect observed in the initial analysis disappeared when the ratings were included as covariates (gender × response association: $F(1,8591.6) = 0.10$; $p = .75$), and was independent of the *type* of target (type × gender × response association: $F(2,8594.0) = 0.17$; $p = .85$). The covariates were not significant, but nevertheless eliminated the gender congruency effect.

Experiment 3b (NON-ANIMATES): For the accuracy data, the analysis of the fixed effects indicated a general gender congruency effect (gender × response association: $\chi^2[df = 1] = 50.06$; $p < .001$), but no interaction with the *type* of target (type × gender × response association: $\chi^2[df = 2] = 4.83$; $p = .089$). Testing the congruency effect for the three noun categories revealed a significant effect for CONGRUENT ANIMATES ($M = 5.75$ %; $\chi^2[df = 1] = 63.41$; $p < .001$; $d_z = 1.22$), for the gendered NON-ANIMATES ($M = 2.25$ %; $\chi^2[df = 1] = 5.93$; $p = .007$; $d_z = 0.50$), and for the neuter NON-ANIMATES ($M = 2.92$ %; $\chi^2[df = 1] = 12.51$; $p < .001$; $d_z = 0.56$). The difference between gendered and neuter NON-ANIMATES was not significant (type × gender × response association: $\chi^2[df = 1] = 0.87$; $p = .350$). As in Experiment 1b (but different from 2b), the general

gender congruency effect observed in the initial analysis disappeared when the ratings were included as covariates (gender × response association: $\chi^2[df = 1] = 2.38$; $p = .123$), and was independent of the *type* of target (type × gender × response association: $\chi^2[df = 1] = 0.67$; $p = .413$). The covariates were not significant, but nevertheless eliminated the gender congruency effect.

For the reaction time data, the analysis of the fixed effects indicated a general gender congruency effect (gender × response association: $F(1,8680.7) = 18.61$; $p < .001$), but no interaction with the *type* of target (type × gender × response association: $F(2,8679.0) = 1.72$; $p = .18$). Testing the congruency effect for the three noun categories revealed a significant effect for CONGRUENT ANIMATES ($M = 16.45$ ms; $F(1,4234.1) = 22.48$; $p < .001$; $d_z = 0.62$) and for the gendered NON-ANIMATES ($M = 10.62$ ms; $F(1,2133.2) = 4.42$; $p = .02$; $d_z = 0.38$), but not for the neuter NON-ANIMATES ($M = 6.41$ ms; $F(1,2163.7) = 1.73$; $p = .085$; $d_z = 0.21$). Different from Experiments 1b and 2b, the general gender congruency effect observed in the initial analysis did not disappear completely when the ratings were included as covariates (gender × response association: $F(1,4375.5) = 5.52$; $p = .02$), but was independent of the *type* of target (type × gender × response association: $F(1,4375.3) = 0.85$; $p = .36$). The covariate *rating* × *response association* failed to reach statistical significance ($F(1,4376.4) = 3.41$; $p = .07$), whereas *rating* was significant ($F(1,34.7) = 4.75$; $p = .04$), but could not completely eliminate the gender congruency effect.

Taken together, the findings from the third pair of experiments with plural forms again replicated the strong and consistent congruency effect for CONGRUENT ANIMATES. For the target items, the effect occurred to a similar extent in accuracy for gendered and neuter nouns, both for GENERIC ANIMATES and NON-ANIMATES, which would be consistent with the connotation hypothesis. In reaction time, however, the effect occurred only for gendered nouns, while for neuter nouns, the effect pointed in the same direction but did not reach significance, suggesting at least an additional influence of grammatical gender. Except for one condition (Experiment 3b, RT),

the congruency effect disappeared when the explicit ratings were included in the analyses as covariate, again consistent with the connotation hypothesis.

Experiments 3a and 3b thus indicate that—even if the stimuli themselves lack grammatical gender and are not accompanied by a gender-indicating article—we can find instances of the gender congruency effect that are comparably strong for neuter nouns as for gendered nouns. These findings suggest that the referent's connotation with biological gender contributes a substantial proportion to the gender congruency effect.

Joint Analysis of the Six Experiments

So far, the analyses of the gendered and neuter target items of the six experiments (each with two measures: accuracy and RT) have provided a somewhat mixed picture regarding the source of the observed gender congruency effect. The findings from four measures (Exp. 1a and 2b, accuracy; Exp. 3a and 3b, RT) speak for the hypothesis that the effect is based on grammar only—or at least includes grammar as an explanatory variable—with a congruency effect occurring for gendered items but not for neuter items, or at least not to the same extent[5]. The findings from six other measures (Exp. 1b, accuracy; Exp. 2a, accuracy and RT; Exp. 2b, RT; and Exp. 3a and 3b, accuracy) speak for the hypothesis that the effect is based on connotations only, with a congruency effect of comparable size occurring both for gendered and neuter items. Additional support for the connotation hypothesis comes from the fact that the inclusion of the explicit association ratings as covariate eliminated the gender congruency effect in ten cases; only in Experiments 2b (accuracy) and 3b (RT) was the effect not eliminated completely.

---

[5] But note that in several of these cases, the associations evoked by the gendered nouns were slightly stronger than those evoked by the neuter nouns.

To ascertain whether the complete set of data lends support to a consistent explanation across all six experiments, we ran a joint analysis using (generalized) mixed linear models with two additional factors, type of *experiment* (pair 1 vs. pair 2 vs. pair 3) and type of *material* (GENERIC ANIMATES [Exp. Xa] vs. NON-ANIMATES [Exp. Xb] as targets). The analysis included a random intercept and random slopes for response association as a function of participants, and a random intercept for the items (without random slopes).

For the accuracy data (from $N = 232$ participants), we found a general gender congruency effect (gender × response association: $\chi^2[df = 1] = 398.11$; $p < .001$) that interacted with the *type* of target (type × gender × response association: $\chi^2[df = 2] = 68.52$; $p < .001$). We also found a five-way interaction indicating small differences between the six experiments for the different types of nouns (experiment × material × type × gender × response association: $\chi^2[df = 4] = 9.56$; $p = .048$). Testing the congruency effect for the three noun categories revealed a significant effect for CONGRUENT ANIMATES ($M = 7.82$ %; $\chi^2[df = 1] = 581.76$; $p < .001$; $d_z = 1.18$), for gendered target items ($M = 3.74$ %; $\chi^2[df = 1] = 79.71$; $p < .001$; $d_z = 0.54$), and for neuter target items ($M = 2.59$ %; $\chi^2[df = 1] = 41.98$; $p < .001$; $d_z = 0.42$). In none of these cases did the congruency effect significantly interact with the factors *experiment* or type of *material*[6]. The difference between gendered and neuter target items was not significant (type × gender × response association: $\chi^2[df = 1] = 2.46$; $p = .117$).

For the reaction time data (from $N = 230$ participants), we found a general gender congruency effect[7] (gender × response association: $\chi^2[df = 1] = 106.49$; $p < .001$) that interacted

---

[6] If at all, then the 5-way interaction observed in the overall analysis was driven by the neuter targets, although the corresponding interaction did not reach significance for this group of nouns (experiment × material × gender × response association: $\chi^2[df = 2] = 5.63$; $p = .060$).

[7] Due to limitations in computing power, we used likelihood ratio tests and report delta chi-square statistics also for the reaction time data instead of *F* tests with Kenward-Roger approximated degrees of freedom.

with the *type* of target (type × gender × response association: $\chi^2[df = 2] = 9.91$; $p = .007$), but did not interact with *experiment* or with the type of *material*. Testing the congruency effect for the three noun categories revealed a significant effect for CONGRUENT ANIMATES ($M = 14.75$ %; $\chi^2[df = 1] = 105.69$; $p < .001$; $d_z = 0.52$), for gendered target items ($M = 9.60$ %; $\chi^2[df = 1] = 24.40$; $p < .001$; $d_z = 0.30$), and for neuter target items ($M = 7.97$ %; $\chi^2[df = 1] = 15.85$; $p < .001$; $d_z = 0.24$). In none of these cases did the congruency effect significantly interact with the factors *experiment* or type of *material*. The difference between gendered and neuter target items was again not significant (type × gender × response association: $\chi^2[df = 1] = 0.57$; $p = .450$).

These joint analyses reveal a clear pattern. Across all six experiments, a strong and consistent congruency effect occurred for CONGRUENT ANIMATES and a smaller effect for gendered and neuter target items. The congruency effect for the two targeted noun categories did not differ in size, which is consistent with the connotation hypothesis.

General Discussion

The main goal of this study was to assess the extent to which the grammatical gender of a noun contributes to the gender congruency effect in affecting the conceptualization of its referent, thereby making a theoretical and empirical contribution to the debate on grammatical gender as a potential candidate for an effect of language on thought. Specifically, we aimed at providing the empirical basis for separating the influence of grammatical gender—which would constitute a genuine impact of language on thought—from a possibly deliberate usage of gender information by participants on the one hand, and from largely non-linguistic, conceptual connotations on the other hand.

In the following, we first critically discuss our findings in view of this goal and relate them to alternative accounts, also paying specific attention to the characteristics of our design. We then propose a possible mechanism for the emergence of the effect in an attempt to reconcile conflicting patterns of findings in the literature, before explicating some possible limitations of our study and drawing a tentative conclusion.

Evidence for Gender Congruency: Pros and Cons

As explained in the introduction, a gender congruency effect is reflected in associations of a noun's referent with the biological sex that is congruent with the grammatical gender of the noun. The current study measured such associations with an explicit method (using a variant of the voice assignment task, VAT) and an implicit method (using a variant of the *Extrinsic Affective Simon Task*, EAST). We also investigated the potential role of the gender-indicating definite article in the measurement of these associations.

In three pairs of experiments with gendered (i.e., masculine or feminine) versus neuter nouns referring to GENERIC ANIMATES and NON-ANIMATES as target items—nouns accompanied by a non-informative, randomly assigned article (in the first pair of experiments), singular nouns lacking the gender-indicating article (second pair), and plural nouns lacking gender itself (third pair)—we observed the following pattern for target categories: A stable gender congruency effect emerged for CONGRUENT ANIMATES (overall d_z = 1.18 for accuracy and 0.52 for RT) and a weaker, yet significant effect emerged for both GENERIC ANIMATES (d_z ranging between 0.23 and 0.61 for accuracy and between 0.03 and 0.33 for RT) and NON-ANIMATES (d_z ranging between 0.19 and 0.68 for accuracy and between 0.22 and 0.49 for RT) with largely comparable effect sizes. Crucially,

this pattern was the same for both singular (gendered) and plural (non-gendered) forms, and was the same for nouns both with masculine/feminine gender and with neuter gender.

To eliminate gender-checking strategies, three measures were implemented: the adoption of an implicit test, which allowed us to scrutinize associations between grammatical and biological gender without explicating this link; the presentation of randomly selected definite articles together with the target items (in Experiments 1a and 1b) and usage of plural forms that lack gender information altogether (in Experiments 3a and 3b); and the inclusion of neuter nouns. The similar patterns obtained across all of these attempts suggest that deliberate usage of grammatical gender can be disregarded as an explanation for the persistence of the observed effect in the current set of experiments.

The findings then seem to indicate that the German speakers participating in our study associate masculine nouns for animates and non-animates more strongly with male properties, and feminine nouns more strongly with female properties. This pattern is remarkable insofar as it seems to attest to a gender congruency effect both in an implicit task and for speakers of a language with three genders, for which such findings tend to be unstable or even absent (we will return to these peculiarities in the next section). Importantly, however, the same participants also associate the *neuter* nouns for animates and non-animates with either male or feminine properties, which, by definition, cannot be an instance of the gender congruency effect.

Neuter nouns were included in an attempt to assess the relative contribution of grammatical gender to the gender congruency effect, as compared to conceptual connotations. In selecting noun categories that share one of these features (sex-related connotations) but differ with regard to the other (gender class), we contrasted items for which grammatical gender and conceptual connotations may co-occur with items that, by definition, lack the gender of interest. Consequently, any effect observed for these neuter nouns would have to be driven by the non-linguistic,

conceptual connotations. We reasoned that this would allow us to tease apart a genuine impact of grammatical gender from an influence of the connotations at least for the neuter nouns, yet by inference also for the gendered nouns, namely by comparing effect sizes for the two classes. Our finding of an effect for the neuter nouns that is similar both in direction and size to that for the gendered nouns across the six experiments (N = 230/232) speaks *against* a direct impact of grammatical gender itself, pointing instead to a central role of connotations in generating the effect.

The Role of Connotations

As elaborated on in the introduction, connotations have been acknowledged, in theory, as a possibly powerful source of gender congruency (Bassetti, 2007, 2014; Beller et al., 2015; Flaherty, 2001; Guiora & Sagi, 1978; Kurinski & Sera, 2011). Whereas denotative meaning is part of a noun's content and hence should be shared by all speakers of a specific language, the connotative meaning may arise from a multitude of sources—including personal experiences, feelings, gender-related stereotypes, personified allegories, or cultural symbols—and hence varies across individual speakers (Cubelli et al., 2011; Nicoladis & Foursha-Stevenson, 2012).

To control for the influence of such connotations on the gender congruency effect, we measured the average sex-related associations for our items in a pretest, and we controlled for associations with the items in the participants of the experimental studies. Our pattern of findings indicates that connotations indeed contribute to the emergence of the gender congruency effect. This is supported by the apparent mediation of the effect by the respective strength of participants' connotations, measured in the assignment task. A similar dependence of the effect on the strength of the connotation was also observed in previous studies, in which weak or gender-incongruent connotations eliminated the effect (Bender et al., 2016a, 2016b).

More importantly, the observed effect for the class of *neuter* nouns cannot be accounted for by either a deliberate use of gender information or an immediate impact of grammatical gender, hence rendering connotations the only possible source for the effect. As the neuter nouns used in this study were matched with the gendered nouns on important variables, and yet still produced an effect of same size, we feel safe to infer that for the gendered nouns too, connotations play a crucial role in the emergence of the gender congruency effect.

Two objections to this account may be raised: (i) By controlling for, and subsequently discounting, connotations as a contribution to the gender congruency effect, we may run the risk of throwing out the baby with the bathwater; and (ii) the connotations themselves may be motivated by grammatical gender, which would then constitute a case not of immediate but of indirect impact of a grammatical property on conceptualization.

The first objection would be critical mainly for the gendered nouns for which grammatical gender and connotations may be conflated. This is one of the reasons why we consider previous studies, which exclusively used masculine and feminine nouns, to be non-conclusive. Here, however, the inclusion of neuter nouns allowed us to compare items for which gender may have contributed to the effect with items for which this cannot be the case. That the same effect was observed for both types of items encourages us to infer that it is generated by the same mechanism.

But what if the connotations themselves are motivated, at least indirectly, by grammatical gender? This second objection is justified insofar as especially allegorical portrayals of abstract terms often followed exactly this logic. As noted above, the Latin *libertas,* being feminine, invited the portrayal of the tutelary deity as goddess. One reason for picking a gender-congruent portrayal was certainly that this facilitated linguistic references to “her” (Cook, 2016). Other examples of this mechanism are fairy tale frogs that turn into princes or princesses, depending on the frog’s grammatical gender in the respective language (Bassetti, 2007; and see Segel & Boroditsky, 2011).

Although this appears to be the outcome of a deliberate decision indicating strategic usage of gender information, we cannot rule out that, in some cases, the choice of allegorical sex was triggered by grammatical gender in a less conscious manner, hence reflecting a case of linguistic relativity. Still, this historical possibility should not be taken as evidence for a cognitive mechanism in our present-day participants. Liberty kept both its feminine gender and its female connotations when turning from Latin into French. It is therefore not surprising that when designing the Statue of Liberty, the French sculptor Bartholdi depicted her as a woman. As a consequence, US Americans today strongly associate liberty with female properties, too, but only based on the conceptual connotations linked to a cultural symbol, since the noun itself is neuter. By contrast, the German noun for liberty (*Freiheit*) does *not* evoke any sex-related associations (Bender et al., 2016b)—although it happens to have feminine gender just like the Latin and French noun. In other words, while sex-related connotations may be rooted in grammatical gender historically, the mechanism driving a gender congruency effect in present-day participants is a conceptual rather than a grammatical one.

A second reason why the two mechanisms should be distinguished is that connotations themselves are generally only weakly correlated with the congruent grammatical gender. To address this, we conducted extensive pre-tests of the material with the aim of identifying suitable items. These tests revealed few systematic effects of gender: Across the set of items, most were neutral, some were gender-congruent, and fewer were gender-incongruent. Crucially, neuter nouns sometimes are among those that do evoke strong sex-related connotations in the absence of the respective gender. If the impact of grammatical gender on conceptualization were systematic and persistent (even if weak), however, it should produce an effect (even if small) beyond the connotations, leading, ceteris paribus, to a somewhat stronger gender congruency effect for the masculine/feminine nouns than for the neuter nouns.

Reconciling Conflicting Findings: A Possible Mechanism of the Gender Congruency Effect

Our findings are in line both with studies using the VAT, in demonstrating a gender congruency effect, and with studies using implicit tasks, in suggesting that an effect of grammatical gender proper is restricted to the lexical level and does not spill over onto the conceptual level. However, our findings also deviate from this previous research in three important ways. First, we did not find the difference in effects for (GENERIC) ANIMATES versus NON-ANIMATES reported elsewhere (Cook, 2016; Saalbach et al., 2012; Vigliocco et al., 2005). Second, the effect in our study emerged both in an implicit task and for speakers of a language with three genders, for which such findings tend to be unstable or even absent (Cubelli et al., 2011; Sera et al., 2002). And third, the effect emerged in the absence of the definite article, which earlier studies have shown to be instrumental for producing the effect in German (Imai et al., 2014; Konishi, 1994; Vigliocco et al., 2004). We believe that all of these peculiarities of our findings can be accounted for by the same mechanism, and that this mechanism is even able to accommodate the conflicting findings reported in the literature.

Let us begin with the issue of ANIMATES versus NON-ANIMATES. As we aimed at disentangling an influence of grammatical gender from conceptual connotations, we pre-selected our items so as to ensure that they evoke comparably strong sex-related associations across the gendered and neuter items. While the crucial question was whether gendered and neuter items would generate a gender congruency effect of comparable size, we had no reason to assume that this effect should differ in size between ANIMATES and NON-ANIMATES. On the contrary: If we were to find supportive evidence for the connotation hypothesis, we would have to predict that GENERIC ANIMATES and NON-ANIMATES should generate an effect of similar size, just as gendered and neuter nouns should. In a previous study, this was what we found: When comparing GENERIC ANIMATES

with and without strong associations and NON-ANIMATES with and without strong associations, we obtained a congruency effect for the items of both semantic categories with strong associations, but not for those without such associations (Bender et al., 2016a). And in the current study in which we only used items with strong associations, we again found a similarly strong congruency effect for both semantic categories.

The reason, then, why a difference between the two semantic categories emerged in other studies may be that previous research did not control for such associations and may have happened to utilize more (GENERIC) ANIMATE items with strong associations than NON-ANIMATE items, therefore obtaining an effect for the former but not for the latter. After all, ANIMATES are much more likely than NON-ANIMATES to figure in fairy tales or cartoons and to serve as allegories for human characteristics, virtues and vices, thereby also reflecting gender stereotypes to a greater extent. In other words, a random selection of ANIMATE nouns has a higher base rate of yielding items with strong associations than an equally random selection of NON-ANIMATE nouns.

The same mechanism accounts for the emergence of the effect in a three-gender language and in the absence of the definite article. So far, effects have been observed more reliably in two-gender than in three-gender languages (Sera et al., 2002; Vigliocco et al., 2005), and in the latter almost exclusively when the gender-indicating article was present (Imai et al., 2014; Konishi, 1994; Vigliocco et al., 2004). While our findings deviate from this pattern, it is only the items with strong associations—and more specifically with gender-congruent associations—that do generate the effect in German. When using items with weak (Bender et al., 2016a) or gender-incongruent associations (Bender et al., 2016b), the effect disappears. And, again, a random selection of nouns across semantic categories in German has a higher base rate of yielding items with weak or no associations than of yielding items with strong associations. It is possible that gender-congruent

associations are more frequent in two-gender than in three-gender languages, which would then explain the greater likelihood of observing the effect in the former than the latter.

Alternatively, the effect in two-gender languages may have been brought about by faster gender identification due to a more transparent gender marking. This would then be parallel to those studies with speakers of three-gender languages (notably German), in which the non-transparent gender assignment was compensated by presenting the gender-indicating definite article together with the noun. If no such article is presented (or if it is non-informative as in our Experiments 1a and 1b), only strong associations appear to be sufficiently powerful to still generate a gender congruency effect. This may also explain why, in our study, the effect could emerge in German even in the absence of the article.

If we consider conceptual connotations as a key explanatory factor, we may also be able to reconcile the conflicting findings in previous research. One of the theoretical questions under debate was whether a gender congruency effect is confined to language processing and hence observable on the lexical level only, or whether it spills over onto the conceptual level. Apparently, gender information related to a noun is spontaneously activated whenever that noun is accessed, at least in speakers of two-gender languages (Boutonnet et al., 2012; Cubelli et al., 2005). This allows for an impact of grammatical gender in tasks that involve, or are open to, linguistic processing. That this impact is also confined to these contexts and hence to the lexical level is indicated particularly by those sets of studies employing *implicit measures* (Cubelli et al., 2011; Imai et al., 2014; Kousta et al., 2008; Ramos & Roberson, 2010; Vigliocco et al., 2004, 2005). Effects of gender beyond this level are therefore interpreted as a result of speakers utilizing the available gender information in a deliberate manner, tasks permitting.

This may have been the case in those studies that obtained a stable effect on the conceptual level by directly asking for participants' sex-related associations with referents of nouns in explicit

tasks such as the VAT: When pondering whether a bridge is more male or more female, an obvious strategy may be to draw on its grammatical gender (and studies requesting explanations from their participants report some confirmation for this assumption; e.g., Bassetti, 2007). The likelihood to find effects in two-gender languages would be reinforced by the fact that many of these languages are gender-loaded, that is, gender class is often marked on the noun. This marking not only increases the phonetic similarity of stimuli, but also permits strategic usage of gender information.

Alternatively, the effect measured in studies using the VAT and related explicit tasks may be a different effect to begin with, originating not from the grammatical gender of the noun, but from participants' conceptual connotations to its referent. While one of the sources for these connotations may be gender, as discussed above, the remaining non-linguistic sources supposedly have a much stronger impact. In directly tapping into the conceptual connotations, explicit studies therefore bypass the question of whether these may be proximally caused by grammatical gender. Implicit tasks such as similarity judgments, by contrast, are able to detect a congruency effect only if the grammatical gender has really had an immediate impact on conceptualization, and therefore a priori struggle more to obtain confirmatory results.

A possible reason why—when actually measuring connotations—a congruency-like effect can still emerge is that some connotations may indeed have been motivated by grammatical gender. As argued above, this is most likely the case for allegories, which are often selected deliberately so as to be gender-congruent, and most likely for ANIMATE items, which serve as allegories more frequently than NON-ANIMATE items. If, however, a random selection of ANIMATE nouns has a higher base rate of yielding items with strong gender-congruent associations than an equally random selection of NON-ANIMATE nouns, then previous studies utilizing a blending of animate (or other allegorically used nouns) and non-animate items would have happened to observe a congruency-like effect due to the greater likelihood of gender-congruent connotations in the former

(Flaherty, 2001; Nicoladis & Foursha-Stevenson, 2012; Sera et al., 1994, 2002). Likewise, those studies that separated animate and non-animate items would have happened to observe the effect for the former, but not for the latter (Cook, 2016; Saalbach et al., 2012; Vigliocco et al., 2005).

Taken together, the findings reported here and the considerations based on them suggest that, rather than the grammatical property of the words (i.e., their gender) affecting how we conceptualize their referents, it seems to be the conceptual connotations directly linked to these referents that generate the gender congruency effect. But what do these findings reveal about the relationship between language and thought more generally?

The Gender Congruency Effect and Linguistic Relativity

For grammatical gender to be a convincing instance of linguistic relativity, the referents of nouns should be conceptualized as somewhat more male or more female *by virtue of the gender class* to which the nouns belong. As we have argued so far, gender may have indeed influenced the sex-related connotations that people have for some gendered nouns, but this influence appears to be neither systematic nor persistent, and was occasionally brought about by strategic usage of gender information. If this is correct, then grammatical gender is likely not the most promising candidate for the influence of grammatical patterns on thought.

Still, our findings are in line with recent developments in the field of linguistic relativity. Compared to other domains for which linguistic relativity has been investigated, language-specific gender classes are much closer to the analogy than the codability end of the continuum (Lucy, 2016): Rather than carving different categories out of the same spectrum, they therefore create purely linguistic categories (Bassetti, 2007). Apart from some personal nouns such as “man”, “mother”, or “maid”, they simply do not reflect any real differences in the world: Not all mice are

female, simply because *die Maus* is feminine, and nor are the moon (*der Mond*) or the morning (*der Morgen*) in any conceivable manner male. In other words, even with language working as a spotlight (Wolff & Holmes, 2011), there is nothing to actually be spotted. And even if there were a difference to be spotted, boundaries between the classes would not be sufficiently fuzzy for a categorical system to provide conflicting information (cf. Cibelli et al., 2016).

A skeptical assessment of gender is therefore compatible with theoretical considerations in the field. In fact, even Lucy, in his supportive review on linguistic relativity, cautions against the prevailing approach taken in gender research, namely to recast the structural patterns caused by gender systems in terms of how they partition our familiar domains as "classifying objects according to the most salient values in our own language (natural sex)" (2016, p. 500). Not only are masculine and feminine gender *arbitrary classes* for the vast majority of nouns in formal gender languages such as German, they are also *arbitrary labels*. Still, these labels keep exerting a considerable impact on how people think about grammatical gender. It is partly for this reason that the hypothesis of gender classes as possibly affecting referent perception and conceptualization in native speakers has continued to be so inviting.

Possible Limitations of the Study

A substantial proportion of our participants were university students, all of whom had learned English in school, and some of whom may have learned a further foreign language. Although additional knowledge of (non-gendered) English is non-critical in this context (cf. Bassetti, 2014; Boroditsky et al., 2003; Ervin, 1962; Forbes et al., 2008; Kurinski & Sera, 2007), both a second gendered language and the meta-linguistic knowledge provided in the German educational system may increase awareness of the arbitrariness of the gender system. While it has become almost

impossible to recruit native speakers of German lacking this meta-linguistic awareness, the possibility should be taken into account that the effect might have been stronger with such participants.

Other limitations are related to the methods used. We combined two measures of gender-sex associations, one explicit (VAT) and one implicit (EAST). This allowed us to compare our results directly with studies using such measures, but not with studies adopting even more implicit measures (such as similarity judgments or substitution errors). More generally, implicit tasks are a suitable means to control for the deliberate usage of gender information by participants, but come at the cost of being less ecologically valid than tasks focusing on real-world applications. As a consequence of this methodological choice, together with the focus on bare nouns as stimuli, the findings from this study are harder to generalize to other contexts. Future research may therefore seek to investigate more systematically the dependence of the gender congruency effect on the methods employed.

An implication common to these potential limitations is that they may render our results more conservative in terms of effect size.

Conclusion

Gender classes are relevant primarily on the syntactic level, and are in fact so central to speech production that gender information is accessed automatically even in tasks that do not require it. This almost unavoidable availability of gender information poses challenges to the investigation of its subconscious effects, but renders it at least possible that a noun's gender may affect the conceptualization of its referent. Surprisingly, this effect seems to be rather small, if it exists at all—either because gender information simply does not spill over onto the conceptual

level, or because speakers even suppress the information as irrelevant. But even if an effect of grammatical gender on conceptualization emerges, it appears to be superimposed to a substantial degree by non-linguistic, conceptual connotations. If these connotations indeed bear the brunt of the gender congruency effect, as indicated by the current study, they deserve much more attention—both in study designs and the selection of material, and in investigations of their origins, composition, interactions, and modes of operation.

Acknowledgment

This work was supported in part by the Research Council of Norway through the SFF *Centre for Early Sapiens Behaviour* (SapienCE), project number 262618. We are grateful to Sarah Mannion de Hernandez for proofreading. Correspondence concerning this article should be addressed to A. Bender, Department of Psychosocial Science, University of Bergen, N-5020 Bergen, Norway. Electronic mail may be sent to Andrea.Bender@uib.no.

References

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition, 122,* 292-305.

Bassetti, B. (2007). Bilingualism and thought: Grammatical gender and concepts of objects in Italian-German bilingual children. *International Journal of Bilingualism, 11,* 251-273.

Bassetti, B. A. L. (2014). Is grammatical gender considered arbitrary or semantically motivated? Evidence from young adult monolinguals, second language learners, and early bilinguals. *British Journal of Psychology, 105,* 273-294.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *Lme4: linear mixed-effects models using eigen and s4. R package version 1.1-7*. [https://github.com/lme4/lme4/]

Beller, S., Brattebø, K. F., Lavik, K. O., Reigstad, R. D., & Bender, A. (2015). Culture or language: What drives effects of grammatical gender? *Cognitive Linguistics, 26,* 331-359.

Bender, A., & Beller, S. (2017). The power of 2: How an apparently irregular numeration system facilitates mental arithmetic. *Cognitive Science, 41,* 158–187.

Bender, A., Beller, S., & Klauer, K. C. (2011). Grammatical gender in German - a case for linguistic relativity? *Quarterly Journal of Experimental Psychology, 64,* 1821-1835.

Bender, A., Beller, S., & Klauer, K. C. (2016a). Crossing grammar and biology for gender categorizations: Investigating the gender congruency effect in generic nouns for animates. *Journal of Cognitive Psychology, 28*, 530-558

Bender, A., Beller, S., & Klauer, K. C. (2016b). Lady Liberty and Godfather Death as candidates for linguistic relativity? Scrutinizing the gender congruency effect on personified allegories with explicit and implicit measures. *Quarterly Journal of Experimental Psychology, 69,*

48-64

Boroditsky, L., & Gaby, A. (2010). Remembrances of times East: Absolute spatial representations of time in an Australian Aboriginal community. *Psychological Science, 21,* 1635-1639.

Boroditsky, L., & Schmidt, L. (2000). Sex, syntax, and semantics. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society* (pp. 42–47). New York, NY: Psychology Press.

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 investigation of language and thought* (pp. 61-79). Cambridge, MA: MIT Press.

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

Bowers, J. S., Vigliocco, G., Stadthagen, H., & Vinson, D. (1999). Distinguishing language from thought: Experimental evidence that syntax is lexically rather than conceptually represented. *Psychological Science, 10,* 310-315.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental Psychology, 58,* 412-424.

Cibelli, E., Xu, Y., Austerweil, J. L., Griffiths, T. L., & Regier, T. (2016). The Sapir-Whorf Hypothesis and probabilistic inference: Evidence from the domain of color. *PLoS ONE, 11(7),* e0158725.

Clark-Carter, D. (2004). *Quantitative psychological research: A student's handbook*. New York: Psychology Press.

Clarke, M. A., Losoff, A., McCracken, M., & Still, J. (1981). Gender perception in Arabic and English. *Language Learning, 31,* 159–167.

Comrie, B. (1999). Grammatical gender systems: A linguist's assessment. *Journal of*

Psycholinguistic Research, 28, 457-466.

Cook, S. V. (2016). Gender matters: From L1 grammar to L2 semantics. *Bilingualism: Language and Cognition.* [online preprint]

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

Cubelli, R., Lotto, L., Paolieri, D., Girelli, M., & Job, R. (2005). Grammatical gender is selected in bare noun production: Evidence from the picture–word interference paradigm. *Journal of Memory and Language, 53,* 42-59.

Cubelli, R., Paolieri, D., Lotto, L., & Job, R. (2011). The effect of grammatical gender on object categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 449-460.

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology, 50,* 77-85.

Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychological Science, 24,* 613-621.

Ervin, S. M. (1962). The connotations of gender. *Word, 18(1-3),* 249-261.

Flaherty, M. (2001). How a language gender system creeps into perception. *Journal of Cross-Cultural Psychology, 32,* 18-31.

Forbes, J. N., Poulin-Dubois, D., Rivero, M. R., & Sera, M. D. (2008). Grammatical gender affects bilinguals' conceptual gender: Implications for linguistic relativity and decision making. *The Open Applied Linguistics Journal, 1,* 68-76.

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

Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition, 108,* 819-824.

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

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2008). Support for lateralization of the Whorf effect beyond the realm of color discrimination. *Brain & Language, 105,* 91-98.

Gomila, A. (2015). Language and thought: The neo-Whorfian hypothesis. In J. D. Wright (Ed.), *The international encyclopedia of the social & behavioral sciences* (2nd ed., pp. 293-299). Oxford: Elsevier.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464-1480.

Guiora, A., & Sagi, A. (1978). A cross-cultural study of symbolic meaning: Developmental aspects. *Language Learning, 28,* 381–386.

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

Haun, D. B. M., Rapold, C. J., Janzen, G., & Levinson, S. C. (2011). Plasticity of human spatial memory: Spatial language and cognition covary across cultures. *Cognition, 119,* 70-80.

Hofstätter, P. R. (1963). Über sprachliche Bestimmungsleistungen: Das Problem des grammatikalischen Geschlechts von Sonne und Mond. *Zeitschrift für Experimentelle und Angewandte Psychologie, 10,* 91–108.

Hohlfeld, A. (2006). Accessing grammatical gender in German: The impact of gender-marking regularities. *Applied Psycholinguistics, 27,* 127-142.

Imai, M., & Gentner, D. (1997). A cross-linguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition, 62,* 169-200.

Imai, M., Schalk, L., Saalbach, H., & Okada, H. (2014). All giraffes have female-specific

properties: Influence of grammatical gender on inferences about sex-specific properties in German speakers. *Cognitive Science, 38,* 514-536.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434-446.

Judd, C. M., Westfall, J., & Kenny D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103,* 54-69.

Koch, S. C., Zimmermann, F., & Garcia-Retamero, R. (2007). El sol – die Sonne: Hat das grammatische Geschlecht von Objekten Implikationen für deren semantischen Gehalt? *Psychologische Rundschau, 58,* 171-182

Konishi, T. (1993). The semantics of grammatical gender: A cross-cultural study. *Journal of Psycholinguistic Research, 22,* 519-534.

Konishi, T. (1994). The connotations of gender: A semantic differential study of German and Spanish. *Word, 45,* 317-327.

Köpcke, K.-M., & Zubin, D. A. (1983). Die kognitive Organisation der Genuszuweisung zu den einsilbigen Nomen der deutschen Gegenwartssprache. *Zeitschrift für germanistische Linguistik, 11,* 166-182.

Köpcke, K.-M., & Zubin, D. A. (1984). Sechs Prinzipien für die Genuszuweisung im Deutschen: Ein Beitrag zur natürlichen Klassifikation. *Linguistische Berichte, 93,* 26-50.

Kousta, S.-T., Vinson, D. P., & Vigliocco, G. (2008). Investigating linguistic relativity through bilingualism: The case of grammatical gender. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 843-858.

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

Cognition, 14, 203-220.

Kurinski, E., Jambor, E., & Sera, M. D. (2016). Spanish grammatical gender: Its effects on categorization in native Hungarian speakers. *International Journal of Bilingualism, 20,* 76-93.

Leinbach, M. D., Hort, B. E., & Fagot, B. I. (1997). Bears are for boys: Metaphorical associations in young children's gender stereotypes. *Cognitive Development, 12,* 107-130.

Lucy, J. A. (1992). *Language diversity and thought.* Cambridge: Cambridge University Press.

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

Lupyan, G. (2008). The conceptual grouping effect: Categories matter (and named categories matter more). *Cognition, 108,* 566–577.

Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science, 24,* 279–284.

Majid, A., Bowerman, M., Kita, S., Haun, D., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences, 8,* 108-114.

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

Mickan, A., Schiefke, M., & Stefanowitsch, A. (2014). Key is a llave is a Schlüssel: A failure to replicate an experiment from Boroditsky et al. 2003. In A. Stefanowitsch & S. Niemeier (Eds.), *Yearbook of the German Cognitive Linguistics Association* (pp. 39–50). Berlin: De Gruyter Mouton.

Mills, A. E. (1986). *The acquisition of gender: A study of English and German*. London: Springer.

Müller, O., & Hagoort, P. (2006). Access to lexical information in language comprehension:

Semantics before syntax. *Journal of Cognitive Neuroscience, 18,* 84–96.

Mullen, M. K. (1990). Children's classifications of nature and artifact pictures into female and male categories. *Sex Roles, 23,* 577–587.

Nicoladis, E., & Foursha-Stevenson, C. (2012). Language and culture effects on gender classification of objects. *Journal of Cross-Cultural Psychology, 43,* 1095–1109.

R Core Team. (2014). *R: a language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. [http://www.R-project.org/]

Ramos, S., & Roberson, D. (2010). What constrains grammatical gender effects on semantic judgements? Evidence from Portuguese. *Journal of Cognitive Psychology, 23,* 102–117.

Saalbach, H., Imai, M., & Schalk, L. (2012). Grammatical gender and inferences about biological properties in German-speaking children. *Cognitive Science, 36,* 1251-1267.

Schwichtenberg, B., & Schiller, N. O. (2004). Semantic gender assignment regularities in German. *Brain and Language, 90,* 326-337.

Segel, E., & Boroditsky, L. (2011). Grammar in art. *Frontiers in Psychology, 1:244,* 1–3.

Sera, M. D., Berge, C. A. H., & Castillo Pintado, J. (1994). Grammatical and conceptual forces in the attribution of gender by English and Spanish speakers. *Cognitive Development, 9,* 261-292.

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

Singmann, H. (2014). *Afex: analysis of factorial experiments. R package version 0.13-145*. [http://cran.r-project.org/package=afex]

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: Cambridge

University Press.

Slobin, D. I. (2003). Language and thought online: Cognitive consequences of linguistic relativity. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the investigation of language and thought* (pp. 157-191). Cambridge, MA: MIT Press.

Vigliocco, G., Vinson, D. P., Indefrey, P., Levelt, W. J. M., & Hellwig, F. (2004). The role of grammatical gender and semantics in German word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 483-497.

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,* 501-520.

Whorf, B. L. (1956). *Language, thought and reality.* Cambridge, MA: MIT Press.

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

Zubin, D. A., & Köpcke, K.-M. (1986). Gender and folk-taxonomy: The indexical relation between grammatical gender and lexical categorization. In C. Craik (Ed.), *Noun classes and categorizations* (pp. 139-180). Amsterdam: John Benjamins.

Appendix A: Material

For each experiment, all stimuli are listed according to the different item categories used. The values reported in brackets represent mean ratings of biological gender (sex), coded from clearly female (1) to clearly male (4). The first value originated from previous studies and was used to select the items for an experiment; the second value originated from the respective experiment and was obtained according to the instructions listed at the end of each section.

(I) Stimuli for Experiments 1a and 1b

The set of FIRST NAMES consisted of 40 masculine names used for male persons and 40 feminine names used for female persons. The set of CONGRUENT ANIMATES consisted of 20 pairs of masculine and feminine items. The set of target items contained GENERIC ANIMATES (Experiment 1a) and NON-ANIMATES (Experiment 1b), each consisting of 20 items with male associations (10 masculine and 10 neuter forms) and 20 items with female associations (10 feminine and 10 neuter forms).

FIRST NAMES, masculine, used for male persons (40 items)

Achim; Alfred; Bernhard; Boris; Christoph; David; Dieter; Erich; Georg; Gerhard; Heiko; Helmut; Harald; Herbert; Ingo; Jakob; Johann; Josef; Jürgen; Karl; Karsten; Klaus; Lothar; Lukas; Markus; Martin; Norbert; Paul; Peter; Stefan; Sven; Thomas; Theo; Timo; Ulrich; Uwe; Walter; Adam; Elvis; Heino

FIRST NAMES, feminine, used for female persons (40 items)

Anna; Astrid; Bärbel; Birgit; Christa; Dora; Edith; Esther; Gitte; Gerda; Heike; Helga; Hilde; Berta; Inge; Judith; Hannah; Jutta; Ingrid; Karla; Kirsten; Klara; Laura; Lisa; Marie; Martha; Nadja; Paula; Petra; Steffi; Sonja; Tanja; Thea; Tina; Ulla; Ute; Waltraud; Eva; Carmen; Milva

CONGRUENT ANIMATES, masculine, biologically male (20 items; ratings collected in Experiment 1a)

Mann (–; 3.97); Herr (–; 3.95); Opa (–; 3.97); Vater (–; 3.97); Papa (–; 3.97); Sohn (–; 3.92); Bruder (–; 3.95); Schwager (–; 3.89); Onkel (–; 3.97); Neffe (–; 3.86); Ritter (–; 3.95); Knappe (–; 3.78); Priester (–; 3.97); Mönch (–; 3.95); Knecht (–; 3.92); Teufel (–; 3.78); Hengst (–; 3.95); Stier (–; 3.86); Ochse (–; 3.81); Kater (–; 3.76)

CONGRUENT ANIMATES, feminine, biologically female (20 items; ratings collected in Experiment 1a)

Frau (–; 1.03); Dame (–; 1.05); Oma (–; 1.03); Mutter (–; 1.03); Mama (–; 1.03); Tochter (–; 1.05); Schwester (–; 1.03); Braut (–; 1.03); Tante (–; 1.03); Nichte (–; 1.03); Jungfer (–; 1.11); Zofe (–; 1.19); Amme (–; 1.19); Nonne (–; 1.03); Magd (–; 1.14); Hexe (–; 1.05); Stute (–; 1.22); Kuh (–; 1.43); Sau (–; 2.05); Henne (–; 1.27)

Experiment 1a

GENERIC ANIMATES, masculine, male association (10 items)

Fasan (3.01; 2.95); Bussard (3.55; 3.51); Dachs (3.44; 3.30); Adler (3.53; 3.59); Hering (3.31; 3.22); Frosch (3.24; 3.30); Rabe (3.18; 3.32); Storch (2.87; 3.05); Igel (2.94; 3.03); Kranich (3.07; 3.24)

GENERIC ANIMATES, neuter, male association (10 items)

Krokodil (3.59; 3.38); Mammut (3.61; 3.49); Nashorn (3.58; 3.65); Ross (3.21; 3.11); Reptil (3.16; 3.22); Gnu (2.83; 2.97); Rind (2.95; 2.97); Schwein (2.84; 3.05); Kamel (2.82; 2.92); Wiesel (2.85; 2.92)

GENERIC ANIMATES, feminine, female association (10 items)

Ente (1.83; 1.62); Schnecke (1.85; 1.81); Muschel (1.81; 1.86); Amsel (1.83; 1.95); Drossel (1.80; 2.00); Ziege (1.87; 1.89); Hummel (1.82; 2.27); Gans (1.69; 1.65); Biene (1.58; 1.57); Eule (1.88; 2.24)

GENERIC ANIMATES, neuter, female association (10 items)

Kätzchen (1.44; 1.65); Huhn (1.48; 1.27); Reh (1.54; 1.59); Schaf (1.93; 2.24); Fohlen (1.80; 2.11); Zebra (2.02; 2.41); Zicklein (1.78; 1.81); Küken (1.83; 2.00); Känguru (1.97; 2.49); Lamm (1.91; 2.11)

Practice items:
Günter, Heinrich, Manfred, Rainer, Gudrun, Rita, Margit, Ruth (*FIRST NAMES*); Chef, Häuptling, Ärztin, Witwe (*CONGRUENT ANIMATES*); Mücke, Eichhörnchen, Karpfen, Karibu (*GENERIC ANIMATES*)

Experiment 1b

NON-ANIMATES, masculine, male association (10 items)

Degen (3.39; 3.48); Pflug (3.46; 3.50); Säbel (3.59; 3.60); Amboss (3.59; 3.58); Speer (3.60; 3.48); Traktor (3.60; 3.50); Hammer (3.61; 3.53); Knüppel (3.65; 3.55); Frack (3.68; 3.63); Bagger (3.70; 3.65)

NON-ANIMATES, neuter, male association (10 items)

Geweih (3.41; 3.33); Hemd (3.51; 3.50); Jackett (3.35; 3.45); Trikot (3.43; 3.25); Metall (3.43; 3.35); Benzin (3.43; 3.25); Turnier (3.38; 2.95); Werkzeug (3.54; 3.30); Schwert (3.73; 3.60); Gebrüll (3.41; 3.25)

NON-ANIMATES, feminine, female association (10 items)

Bluse (1.23; 1.18); Schminke (1.24; 1.25); Pille (1.28; 1.30); Brosche (1.41; 1.50); Seide (1.43; 1.55); Puppe (1.45; 1.45); Bürste (1.58; 1.45); Spange (1.62; 1.55); Schürze (1.63; 1.80); Wiege (1.74; 1.70)

NON-ANIMATES, neuter, female association (10 items)

Haar (1.54; 1.73); Kleid (1.19; 1.25); Kostüm (1.65; 1.85); Korsett (1.46; 1.30); Juwel (1.46; 1.53); Parfüm (1.41; 1.68); Ballett (1.30; 1.43); Täschchen (1.41; 1.28); Garn (1.70; 1.75); Mitleid (1.78; 2.03)

Practice items:
Günter, Heinrich, Manfred, Rainer, Gudrun, Rita, Margit, Ruth (*FIRST NAMES*); Chef, Häuptling, Ärztin, Witwe (*CONGRUENT ANIMATES*); Bohrer; Spindel; Gebäck; Schnitzel (*NON-ANIMATES*)

Rating task

Subsequent to the EAST, a rating task was implemented for all CONGRUENT ANIMATES and target items (GENERIC ANIMATES in Experiment 1a and NON-ANIMATES in Experiment 1b) used in the EAST. Here, we provide the German instructions.

Instructions for Experiment 1a

In einem Zeichentrickfilm sollen verschiedene Lebewesen als Akteure eingesetzt werden, die Sie im Folgenden aufgelistet finden. Bitte kreuzen Sie für jedes dieser Lebewesen an, ob dazu besser eine weibliche oder eine männliche Stimme passt. Falls Sie unsicher sind, können Sie Ihr Urteil ein wenig abschwächen („eher weiblich“ oder „eher männlich“).

Wenn Sie beispielsweise ein Meerschweinchen weiblich finden, dann kreuzen Sie bitte „weibliche Stimme“ an; wenn Sie ein Murmeltier eher männlich als weiblich finden, kreuzen Sie bitte „eher männliche Stimme“ an; und wenn Sie ein Flusspferd ziemlich männlich finden, kreuzen Sie bitte „männliche Stimme“ an.

Wichtig: Kreuzen Sie für jedes Lebewesen immer genau eine der vier Antwortoptionen an. Sie können Ihre Auswahl auch jederzeit korrigieren.

Instructions for Experiment 1b

Viele Begriffe und Dinge lösen stereotype Assoziationen an Männer oder Frauen aus: bei einem Boxhandschuh beispielsweise denken wir eher an einen Mann, bei einem Lippenstift eher an eine Frau. Überlegen Sie im Folgenden bitte für jeden der nacheinander gezeigten Begriffe, ob Sie diesen eher mit einer Frau oder einem Mann assoziieren. Seien Sie dabei bitte ganz spontan.

BEISPIEL: HALSBAND

Wenn Sie beispielsweise für den Begriff „Halsband“ eine starke Assoziation an eine Frau haben, kreuzen Sie bitte das linke Kästchen an („eindeutig Frau“); bei einer starken Assoziation an einen Mann das rechte („eindeutig Mann“). Wenn Sie nicht so sicher sind, kreuzen Sie eines der beiden mittleren Kästchen an („eher Frau“ oder „eher Mann“). Sie können Ihre Auswahl auch jederzeit korrigieren.

(II) Stimuli for Experiments 2a and 2b

The stimuli were the same as in Experiments 1a and 1b; they are listed here with the ratings obtained in Experiments 2a and 2b.

FIRST NAMES (80 items; same as in Experiments 1a and 1b)

CONGRUENT ANIMATES, masculine, biologically male (20 items; ratings collected in Experiment 2a)

Mann (–; 3.95); Herr (–; 3.95); Opa (–; 3.97); Vater (–; 3.95); Papa (–; 3.92); Sohn (–; 3.90); Bruder (–; 3.90); Schwager (–; 3.97); Onkel (–; 3.95); Neffe (–; 3.82); Ritter (–; 3.92); Knappe (–; 3.79); Priester (–; 3.95); Mönch (–; 3.92); Knecht (–; 3.90); Teufel (–; 3.69); Hengst (–; 3.90); Stier (–; 3.79); Ochse (–; 3.79); Kater (–; 3.90)

CONGRUENT ANIMATES, feminine, biologically female (20 items; ratings collected in Experiment 2a)

Frau (–; 1.05); Dame (–; 1.05); Oma (–; 1.03); Mutter (–; 1.13); Mama (–; 1.05); Tochter (–; 1.08); Schwester (–; 1.08); Braut (–; 1.08); Tante (–; 1.05); Nichte (–; 1.10); Jungfer (–; 1.18); Zofe (–; 1.10); Amme (–; 1.05); Nonne (–; 1.08); Magd (–; 1.10); Hexe (–; 1.10); Stute (–; 1.15); Kuh (–; 1.46); Sau (–; 2.05); Henne (–; 1.21)

Experiment 2a

GENERIC ANIMATES, masculine, male association (10 items)

Fasan (–; 3.05); Bussard (–; 3.44); Dachs (–; 3.46); Adler (–; 3.46); Hering (–; 3.44); Frosch (–; 3.28); Rabe (–; 3.38); Storch (–; 3.13); Igel (–; 3.03); Kranich (–; 3.18)

GENERIC ANIMATES, neuter, male association (10 items)

Krokodil (–; 3.44); Mammut (–; 3.44); Nashorn (–; 3.46); Ross (–; 3.28); Reptil (–; 3.18); Gnu (–; 2.97); Rind (–; 3.10); Schwein (–; 3.18); Kamel (–; 3.00); Wiesel (–; 3.05)

GENERIC ANIMATES, feminine, female association (10 items)

Ente (–; 1.79); Schnecke (–; 1.74); Muschel (–; 1.85); Amsel (–; 1.79); Drossel (–; 1.97); Ziege (–; 1.79); Hummel (–; 2.18); Gans (–; 1.64); Biene (–; 1.62); Eule (–; 2.00)

GENERIC ANIMATES, neuter, female association (10 items)

Kätzchen (–; 1.67); Huhn (–; 1.41); Reh (–; 1.56); Schaf (–; 2.08); Fohlen (–; 2.21); Zebra (–; 2.46); Zicklein (–; 1.92); Küken (–; 2.00); Känguru (–; 2.56); Lamm (–; 2.03)

Experiment 2b

NON-ANIMATES, masculine, male association (10 items)

Degen (–; 3.54); Pflug (–; 3.54); Säbel (–; 3.77); Amboss (–; 3.59); Speer (–; 3.67); Traktor (–; 3.56); Hammer (–; 3.62); Knüppel (–; 3.64); Frack (–; 3.77); Bagger (–; 3.69)

NON-ANIMATES, neuter, male association (10 items)

Geweih (–; 3.41); Hemd (–; 3.38); Jackett (–; 3.41); Trikot (–; 3.36); Metall (–; 3.56); Benzin (–; 3.41); Turnier (–; 3.10); Werkzeug (–; 3.51); Schwert (–; 3.64); Gebrüll (–; 3.51)

NON-ANIMATES, feminine, female association (10 items)

Bluse (–; 1.13); Schminke (–; 1.15); Pille (–; 1.28); Brosche (–; 1.33); Seide (–; 1.59); Puppe (–; 1.36); Bürste (–; 1.44); Spange (–; 1.54); Schürze (–; 1.62); Wiege (–; 1.62)

NON-ANIMATES, neuter, female association (10 items)

Haar (–; 1.69); Kleid (–; 1.13); Kostüm (–; 1.51); Korsett (–; 1.18); Juwel (–; 1.51); Parfüm (–; 1.51); Ballett (–; 1.41); Täschchen (–; 1.28); Garn (–; 1.64); Mitleid (–; 1.95)

(III) Stimuli for Experiments 3a and 3b

The FIRST NAMES were the same as in Experiment 1a. The set of CONGRUENT ANIMATES consisted of 20 pairs of masculine and feminine items in plural form. The set of target items contained GENERIC ANIMATES (Experiment 3a) and NON-ANIMATES (Experiment 3b), each consisting of 20 items with male associations (10 masculine and 10 neuter forms) and 20 items with female associations (10 feminine and 10 neuter forms).

FIRST NAMES (80 items; same as in Experiments 1a and 1b [and 2a and 2b, respectively])

CONGRUENT ANIMATES, masculine, biologically male (20 items; ratings collected in Experiment 3a)

Männer (–; 4.00); Herren (–; 3.98); Väter (–; 3.95); Söhne (–; 3.88); Brüder (–; 3.93); Neffen (–; 3.85); Knappen (–; 3.80); Mönche (–; 3.90); Knechte (–; 3.95); Burschen (–; 3.85); Helden (–; 3.33); Knaben (–; 3.75); Gatten (–; 3.80); Päpste (–; 3.90); Fürsten (–; 3.90); Schurken (–; 3.73); Zwerge (–; 3.55); Stiere (–; 3.93); Ochsen (–; 3.90); Hähne (–; 3.60)

CONGRUENT ANIMATES, feminine, biologically female (20 items; ratings collected in Experiment 3a)

Frauen (–; 1.00); Damen (–; 1.03); Mütter (–; 1.03); Töchter (–; 1.03); Schwestern (–; 1.03); Nichten (–; 1.08); Jungfern (–; 1.13); Nonnen (–; 1.03); Mägde (–; 1.08); Dirnen (–; 1.35); Bräute (–; 1.00); Tanten (–; 1.05); Witwen (–; 1.05); Hexen (–; 1.13); Zofen (–; 1.10); Feen (–; 1.13); Nymphen (–; 1.30); Kühe (–; 1.48); Säue (–; 2.18); Hennen (–; 1.13)

Experiment 3a

GENERIC ANIMATES, masculine, male association (10 items)

Störche (2.87; 2.73); Bussarde (3.47; 3.38); Dachse (3.40; 3.23); Raben (3.07; 3.38); Frösche (3.10; 3.13); Habichte (3.27; 3.30); Falken (3.36; 3.40); Hirsche (3.73; 3.65); Löwen (3.70; 3.70); Wölfe (3.77; 3.58)

GENERIC ANIMATES, neuter, male association (10 items)

Krokodile (3.63; 3.43); Mammuts (3.43; 3.45); Nashörner (3.40; 3.43); Rösser (3.17; 2.98); Reptile (3.10; 3.18); Gnus (3.07; 3.03); Rinder (2.93; 3.18); Schweine (2.84; 3.10); Kamele (2.87; 2.80); Alpakas (2.77; 2.40)

GENERIC ANIMATES, feminine, female association (10 items)

Bienen (1.58; 1.65); Elstern (1.67; 1.85); Gazellen (1.47; 1.78); Giraffen (1.77; 2.00); Hummeln (1.82; 2.18); Gänse (1.69; 1.63); Meisen (1.77; 2.03); Muscheln (1.81; 2.15); Enten (1.87; 2.10); Tauben (1.77; 1.83)

GENERIC ANIMATES, neuter, female association (10 items)

Hühner (1.57; 1.28); Rehe (1.77; 1.60); Schafe (1.83; 2.13); Zebras (2.02; 2.43); Kängurus (2.03; 2.45); Lämmer (2.10; 2.00); Kitze (1.93; 2.05); Ponys (–; 1.93); Lamas (2.20; 2.58); Kälber (2.07; 2.23)

Practice items: Günter, Heinrich, Manfred, Rainer, Gudrun, Rita, Margit, Ruth (*FIRST NAMES*); Bauern, Hengste, Elfen, Zicken (*CONGRUENT ANIMATES*); Schwalben, Pferde, Hunde, Hermeline (*GENERIC ANIMATES*)

Experiment 3b

NON-ANIMATES, masculine, male association (10 items)

Bögen (3.24; 2.98); Pflüge (3.46; 3.43); Traktoren (3.60; 3.55); Ambosse (3.59; 3.55); Speere (3.60; 3.50); Bärte (3.90; 3.80); Kräne (3.59; 3.40); Muskeln (3.31; 3.30); Pfeile (3.34; 3.23); Dolche (3.34; 3.45)

NON-ANIMATES, neuter, male association (10 items)

Geweihe (3.41; 3.33); Hemden (3.51; 3.45); Jacketts (3.35; 3.45); Trikots (3.43; 3.20); Metalle (3.43; 3.40); Turniere (3.38; 3.05); Rohre (3.32; 3.35); Schwerter (3.73; 3.65); Autos (3.46; 3.28); Beile (3.51; 3.50)

NON-ANIMATES, feminine, female association (10 items)

Blusen (1.23; 1.23); Hüften (1.59; 1.55); Pillen (1.28; 1.70); Broschen (1.41; 1.48); Kerzen (1.86; 1.88); Puppen (1.45; 1.53); Windeln (1.97; 1.90); Spangen (1.62; 1.65); Schürzen (1.63; 1.60); Torten (1.79; 1.88)

NON-ANIMATES, neuter, female association (10 items)

Haare (1.54; 2.05); Kleider (1.19; 1.53); Kostüme (1.65; 1.90); Korsetts (1.46; 1.25); Juwelen (1.46; 1.68); Rezepte (1.81; 1.80); Ballette (1.30; 1.63); Gefühle (1.60; 1.78); Cafés (1.70; 2.08); Tücher (1.62; 1.78)

Practice items: Günter, Heinrich, Manfred, Rainer, Gudrun, Rita, Margit, Ruth (*FIRST NAMES*); Bauern, Hengste, Elfen, Zicken (*CONGRUENT ANIMATES*); Nadeln, Beete, Schlipse, Fässer (*NON-ANIMATES*)

Rating task

Subsequent to the EAST, a rating task was implemented for all CONGRUENT ANIMATES and target items (GENERIC ANIMATES in Experiment 3a and NON-ANIMATES in Experiment 3b) used in the EAST. Here, we provide the German instructions.

Instructions for Experiment 3a

In einem Zeichentrickfilm sollen verschiedene Gruppen von Lebewesen, Menschen und Tiere, als Akteure eingesetzt werden, die Sie im Folgenden aufgelistet finden. Bitte kreuzen Sie für jede dieser Gruppen von Lebewesen an, ob dazu besser weibliche oder männliche Stimmen passen. Falls Sie unsicher sind, können Sie Ihr Urteil ein wenig abschwächen („eher weiblich“ oder „eher männlich“).

Wenn Sie beispielsweise eine Gruppe von Meerschweinchen weiblich finden, dann kreuzen Sie bitte „weibliche Stimmen“ an; wenn Sie Murmeltiere eher männlich als weiblich finden, kreuzen Sie bitte „eher männliche Stimmen“ an; und wenn Sie Flusspferde ziemlich männlich finden, kreuzen Sie bitte „männliche Stimmen“ an.

Wichtig: Kreuzen Sie für jede Gruppe von Lebewesen immer genau eine der vier Antwortoptionen an. Sie können Ihre Auswahl auch jederzeit korrigieren. Probieren Sie es am Beispiel „Meerschweinchen“ unten einmal aus!

Instructions for Experiment 3b

Viele Begriffe und Dinge lösen stereotype Assoziationen an Männer oder Frauen aus: Bei Boxhandschuhen beispielsweise denken wir eher an Männer, bei Lippenstiften eher an Frauen. Überlegen Sie im Folgenden

bitte für jeden der nacheinander gezeigten Begriffe, ob Sie diesen eher mit Frauen oder Männern assoziieren. Seien Sie dabei bitte ganz spontan.

BEISPIEL: HALSBÄNDER

Wenn Sie beispielsweise für den Begriff „Halsbänder“ eine starke Assoziation an Frauen haben, kreuzen Sie bitte das linke Kästchen an („eindeutig Frauen“); bei einer starken Assoziation an Männer das rechte („eindeutig Männer“). Wenn Sie nicht so sicher sind, kreuzen Sie eines der beiden mittleren Kästchen an („eher Frauen“ oder „eher Männer“). Sie können Ihre Auswahl auch jederzeit korrigieren.

Appendix B: Random effects structures

In order to determine which random effects structure to assume, we used generalized mixed linear models with random effects for *participants* and *items* for accuracy data, and mixed linear models with random effects for *participants* and *items* for the reaction time data (Bender, Beller, & Klauer, 2016a, 2016b). The analyses were conducted in the statistical programming language R (R Core Team, 2014) using the packages *lme4* (Bates, Maechler, Bolker, & Walker, 2014) and *afex* (Singmann, 2014).

Model comparisons were performed in a two-step procedure: In the first step, we fitted a full model (am1, tm1, ar1, or tr1[8]) with a maximal random effects structure; that is, with random intercepts and random error slopes for the factors *type, gender, response association,* and all interactions thereof for participants, and random intercepts and random slopes for the factor *response association* for items. This model was contrasted with a "null" model (am2, tm2, ar2, or tr2) with only random intercepts for participants and items (without random slopes). If no difference was found, the null model was accepted as the final model, from which the fixed effects were then calculated.

Otherwise, we tested which random slopes as a function of participants or items were needed. Inspecting the variance estimated for the different random slopes, we selected random slopes in a stepwise fashion that appeared to be associated with the largest variances, beginning with the random slope that drew the most variance upon itself. In a second step, we therefore compared a "reduced" model with these random slopes (am3, tm3, ar3, or tr3) with both the full model and the null model, in order to check whether the reduced model explains the data as well as the full model (no difference compared to am1/tm1/ar1/tr1), and whether the additional random error components are necessary as compared to the null model (significant difference compared to am2/tm2/ar2/tr2). In that case, the reduced model was taken as the final model, from which the fixed effects were calculated (see Jaeger, 2008). In all cases, this second step was sufficient, and a third step, choosing the random slope with the next highest variance estimate, was not necessary. In all cases, the random slopes, where necessary, related to the factor *response association* as a function of participants, meaning that there is evidence overall for individual differences in the accuracy and reaction time associated with whether the required response involved the left or the right key.

(I) Experiment 1a

(a) Analysis without covariates

Accuracy

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
am1	93	4785.5	5439.8	-2299.7	4599.5			
am2^{final}	14	4714.6	4813.1	-2343.3	4686.6	87.09	79	.250

Reaction time

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tm1	94	90071	90718	-44942	89883			
tm2	15	90028	90132	-44999	89998	115.12	79	.005

→ tm3: random slope for response association as a function of participants.

[8] Abbreviations: accuracy data (*a*), reaction time data (*t*), main analysis (*m*), and re-analysis (*r*).

tm1	94	90071	90718	-44942	89883			
tm3^{final}	17	89995	90112	-44980	89961	77.68	77	.4568
tm2	15	90028	90132	-44999	89998			
tm3^{final}	17	89995	90112	-44980	89961	37.43	2	<.001

(b) Re-analysis with explicit ratings as covariate

Accuracy

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
ar1	95	4783.8	5452.2	-2296.9	4593.8			
ar2^{final}	16	4712.7	4825.3	-2340.3	4680.7	86.94	79	.253

Reaction time

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tr1	96	90073	90733	-44940	89881			
tr2	17	90032	90148	-44999	89998	116.98	79	=.004

→ tr3: random slope for response association as a function of participants.

tr1	96	90073	90733	-44940	89881			
tr3^{final}	19	89998	90129	-44980	89960	79.47	77	.4011
tr2	17	90032	90148	-44999	89998			
tr3^{final}	19	89998	90129	-44980	89960	37.51	2	<.001

(II) Experiment 1b

(a) Analysis without covariates

Accuracy

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
am1	93	4861.1	5525.5	-2337.5	4675.1			
am2^{final}	14	4763.5	4863.5	-2367.8	4735.5	60.46	79	.9401

Reaction time

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tm1	94	104185	104847	-51999	103997			
tm2	15	104221	104326	-52095	104191	193.33	79	<.001

→ tm3: random slope for response association as a function of participants.

tm1	94	104185	104847	-51999	103997			
tm3^{final}	17	104089	104209	-52028	104055	57.86	77	.9493
tm2	15	104221	104326	-52095	104191			
tm3^{final}	17	104089	104209	-52028	104055	135.47	2	<.001

(b) Re-analysis with explicit ratings as covariate

Accuracy

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
ar1	49	2332.7	2648.8	-1117.3	2234.7			
ar2^{final}	12	2283.4	2360.8	-1129.7	2259.4	24.73	37	.9388

Reaction time

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tr1	50	52606	52924	-26253	52506			
tr2	13	52620	52702	-26297	52594	87.57	37	<.001

→ tr3: random slope for response association as a function of participants.

tr1	50	52606	52924	-26253	52506			
tr3^{final}	15	52555	52650	-26263	52525	18.62	35	.9894
tr2	13	52620	52702	-26297	52594			
tr3^{final}	15	52555	52650	-26263	52525	68.95	2	<.001

(III) Experiment 2a

(a) Analysis without covariates

Accuracy

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
am1	93	5074.0	5738.4	-2444.0	4888.0			
am2^{final}	14	4991.6	5091.6	-2481.8	4963.6	75.61	79	.5873

Reaction time

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tm1	94	102632	103292	-51222	102444			
tm2	15	102659	102764	-51314	102629	184.74	79	<.001

→ tm3: random slope for response association as a function of participants.

tm1	94	102632	103292	-51222	102444			
tm3^{final}	17	102553	102672	-51259	102519	74.56	77	.5576
tm2	15	102659	102764	-51314	102629			
tm3^{final}	17	102553	102672	-51259	102519	110.18	2	<.001

(b) Re-analysis with explicit ratings as covariate

Accuracy

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
ar1	95	5069.7	5748.4	-2439.8	4879.7			
ar2^{final}	12	2144.8	2222.2	-1060.4	2120.8	76.45	79	.5602

Reaction time

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tr1	96	102635	103309	-51221	102443			
tr2	17	102661	102781	-51314	102627	184.29	79	<.001

→ tr3: random slope for response association as a function of participants.

tr1	96	102635	103309	-51221	102443			
$tr3^{final}$	19	102555	102688	-51259	102517	74.05	77	.5742
tr2	17	102661	102781	-51314	102627			
$tr3^{final}$	19	102555	102688	-51259	102517	110.24	2	<.001

(IV) Experiment 2b

(a) Analysis without covariates

Accuracy

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
am1	93	4653.9	5318.3	-2234.0	4467.9			
$am2^{final}$	14	4576.8	4676.8	-2274.4	4548.8	80.86	79	.4207

Reaction time

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tm1	94	99889	100548	-49850	99701			
tm2	15	99872	99977	-49921	99842	140.98	79	<.001

→ tm3: random slope for response association as a function of participants.

tm1	94	99889	100548	-49850	99701			
$tm3^{final}$	17	99772	99891	-49869	99738	36.80	77	1.000
tm2	15	99872	99977	-49921	99842			
$tm3^{final}$	17	99772	99891	-49869	99738	104.19	2	<.001

(b) Re-analysis with explicit ratings as covariate

Accuracy

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
ar1	49	2186.8	2502.9	-1044.4	2088.8			
$ar2^{final}$	12	2144.8	2222.2	-1060.4	2120.8	32.02	37	.7013

Reaction time

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tr1	50	50432	50748	-25166	50332			
tr2	13	50438	50520	-25206	50412	79.97	37	<.001

→ tr3: random slope for response association as a function of participants.

tr1	50	50432	50748	-25166	50332			
tr3^{final}	15	50386	50481	-25178	50356	24.52	35	.9071
tr2	13	50438	50520	-25206	50412			
tr3^{final}	15	50386	50481	-25178	50356	55.45	2	<.001

(V) Experiment 3a

(a) Analysis without covariates

Accuracy

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
am1	93	4663.8	5330.6	-2238.9	4477.8			
am2^{final}	14	4582.6	4683.0	-2277.3	4554.6	76.83	79	.5482

Reaction time

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tm1	94	107267	107932	-53540	107079			
tm2	15	107318	107424	-53644	107288	208.72	79	<.001

→ tm3: random slope for response association as a function of participants.

tm1	94	107267	107932	-53540	107079			
tm3^{final}	17	107163	107283	-53564	107129	49.76	77	.9932
tm2	15	107318	107424	-53644	107288			
tm3^{final}	17	107163	107283	-53564	107129	158.97	2	<.001

(b) Re-analysis with explicit ratings as covariate

Accuracy

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
ar1	95	4657.3	5338.4	-2233.7	4467.3			
ar2^{final}	16	4575.2	4689.9	-2271.6	4543.2	75.88	79	.5785

Reaction time

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tr1	96	107268	107947	-53538	107076			
tr2	17	107319	107439	-53642	107285	208.91	79	<.001

→ tr3: random slope for response association as a function of participants.

tr1	96	107268	107947	-53538	107076			
tr3^{final}	19	107163	107297	-53563	107125	49.10	77	.9945
tr2	17	107319	107439	-53642	107285			
tr3^{final}	19	107163	107297	-53563	107125	159.83	2	<.001

(VI) Experiment 3b

(a) Analysis without covariates

Accuracy

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
am1	93	4231.6	4898.3	-2022.8	4045.6			
am2^{final}	14	4129.4	4229.7	-2050.7	4101.4	55.78	79	.978

Reaction time

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tm1	94	108932	109598	-54372	108744			
tm2	15	108928	109035	-54449	108898	154.03	79	<.001

→ tm3: random slope for response association as a function of participants.

tm1	94	108932	109598	-54372	108744			
tm3^{final}	17	108828	108948	-54397	108794	49.30	77	.9941
tm2	15	108928	109035	-54449	108898			
tm3^{final}	17	108828	108948	-54397	108794	104.73	2	<.001

(b) Re-analysis with explicit ratings as covariate

Accuracy

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
ar1	49	1907.4	2224.7	-904.70	1809.4			
ar2^{final}	12	1852.3	1930.0	-914.14	1828.3	18.86	37	.9941

Reaction time

Model	*df*	*AIC*	*BIC*	*loglik*	*deviance*	$\Delta\chi^2$	Δdf	*p*
tr1	50	55372	55693	-27636	55272			
tr2	13	55368	55451	-27671	55342	69.72	37	<.001

→ tr3: random slope for response association as a function of participants.

tr1	50	55372	55693	-27636	55272			
tr3^{final}	15	55327	55423	-27649	55297	24.45	35	.9089
tr2	13	55368	55451	-27671	55342			
tr3^{final}	15	55327	55423	-27649	55297	45.27	2	<.001

Appendix C: EAST-effects

			Incongruent		*Congruent*		*EAST-effect*	
Experiment		*Item group*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Exp. 1a:	*accuracy*	Congruent animates	81.98	13.97	92.88	10.60	10.67	8.41
		Gendered generics	85.95	16.33	91.26	12.51	4.29	6.98
		Neuter generics	89.28	13.06	90.99	11.35	1.52	6.63
	reaction time	Congruent animates	507.14	79.90	494.45	72.76	12.69	37.37
		Gendered generics	507.95	72.17	500.84	85.59	7.12	38.39
		Neuter generics	505.00	72.96	503.71	77.73	1.29	45.43
Exp. 1b:	*accuracy*	Congruent animates	88.46	7.81	95.30	3.66	6.84	7.29
		Gendered non-animates	91.20	6.19	94.53	4.68	3.33	6.84
		Neuter non-animates	91.71	6.39	95.56	4.35	3.85	6.82
	reaction time	Congruent animates	488.46	65.94	478.58	53.92	9.89	28.70
		Gendered non-animates	490.42	57.10	483.52	64.93	6.90	31.11
		Neuter non-animates	487.01	56.05	476.48	60.36	10.53	30.51
Exp. 2a:	*accuracy*	Congruent animates	86.28	8.54	94.83	5.54	8.55	5.63
		Gendered generics	90.77	8.63	95.04	5.07	4.27	7.25
		Neuter generics	90.26	8.86	93.93	5.51	3.68	6.06
	reaction time	Congruent animates	502.72	80.95	483.12	68.29	19.60	31.27
		Gendered generics	496.67	90.07	484.20	68.92	12.47	38.32
		Neuter generics	497.84	78.67	488.83	74.91	9.01	33.80
Exp. 2b:	*accuracy*	Congruent animates	87.78	8.08	95.85	4.32	8.08	6.79
		Gendered non-animates	90.68	7.92	96.07	4.71	5.38	7.90
		Neuter non-animates	93.25	6.60	94.19	5.66	0.94	5.07
	reaction time	Congruent animates	488.21	75.66	470.58	73.95	17.64	16.74
		Gendered non-animates	486.60	74.86	475.12	67.57	11.49	26.92
		Neuter non-animates	490.47	78.40	476.50	70.68	13.96	28.24
Exp. 3a:	*accuracy*	Congruent animates	88.63	6.08	96.00	3.56	7.38	5.83
		Gendered generics	92.33	7.25	95.33	5.16	3.00	7.27
		Neuter generics	92.75	7.77	95.25	4.46	2.50	6.99
	reaction time	Congruent animates	492.92	70.71	480.84	69.68	12.07	25.89
		Gendered generics	492.60	74.49	483.85	69.42	8.76	31.75
		Neuter generics	488.31	71.35	482.33	63.36	5.98	27.44
Exp. 3b:	*accuracy*	Congruent animates	90.08	6.43	95.83	4.29	5.75	4.70
		Gendered non-animates	93.33	4.95	95.58	4.36	2.25	4.49
		Neuter non-animates	94.08	6.29	97.00	4.26	2.92	5.24
	reaction time	Congruent animates	509.73	72.78	493.29	64.59	16.45	26.39
		Gendered non-animates	508.93	71.06	498.31	69.09	10.62	27.59
		Neuter non-animates	503.85	72.07	497.44	62.71	6.41	29.92

Figure Captions

Figure 1. Four example items of the EAST showing a stimulus on the screen with category labels (second row) and the decision to be made (third row), adapted from Bender and colleagues (2016a).

Figure 2. Results of Experiments 1a and 1b in terms of accuracy and reaction time (mean values and standard deviations are reported in Appendix C).

Figure 3. Results of Experiments 2a and 2b in terms of accuracy and reaction time (mean values and standard deviations are reported in Appendix C).

Figure 4. Results of Experiments 3a and 3b in terms of accuracy and reaction time (mean values and standard deviations are reported in Appendix C).

Figure 1

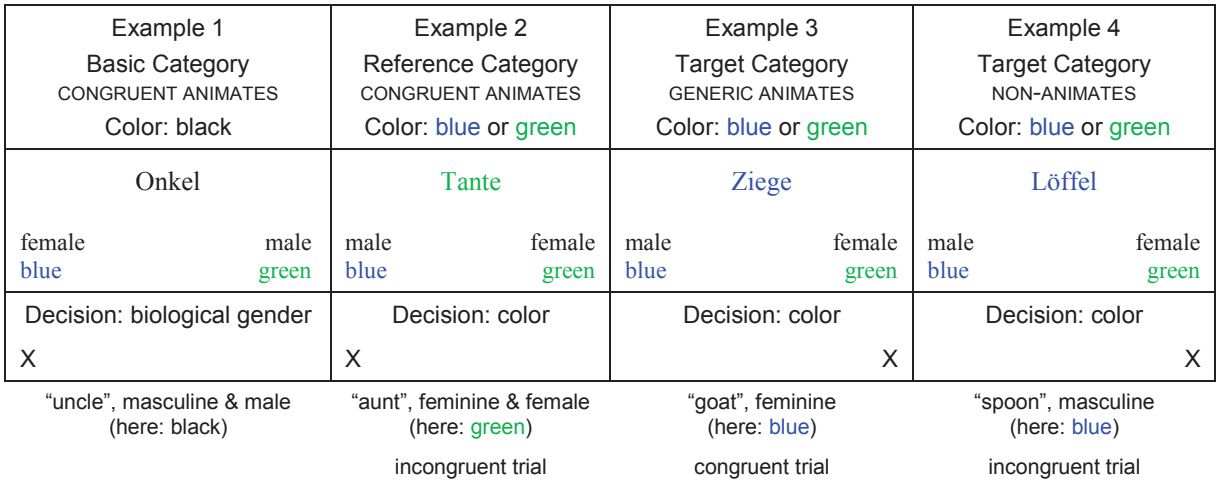

Figure 1. Four example items of the EAST showing a stimulus on the screen with category labels (second row) and the decision to be made (third row), adapted from Bender and colleagues (2016a).

Figure 2

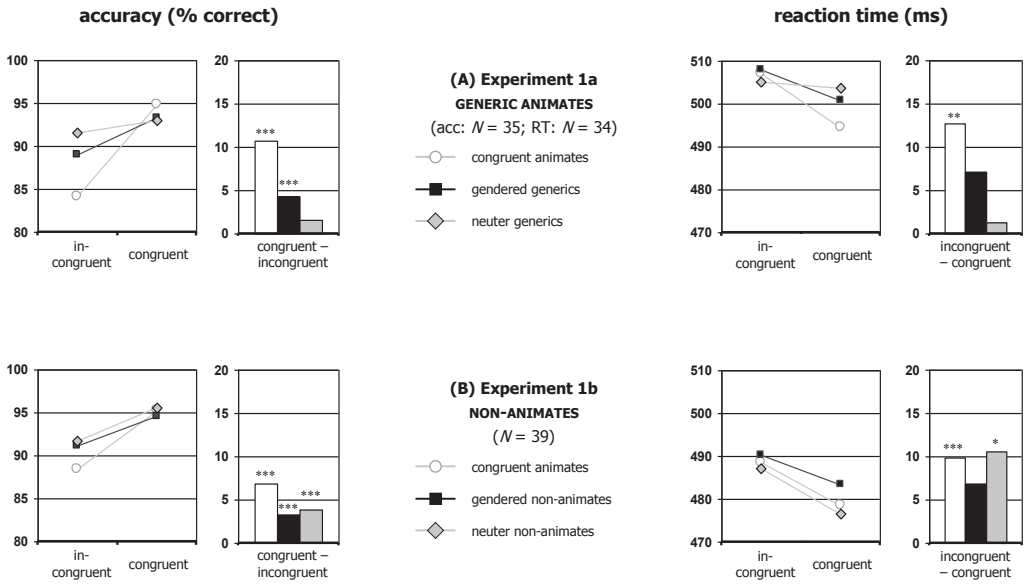

Figure 2. Results of Experiments 1a and 1b in terms of accuracy and reaction time (mean values and standard deviations are reported in Appendix C).

Figure 3

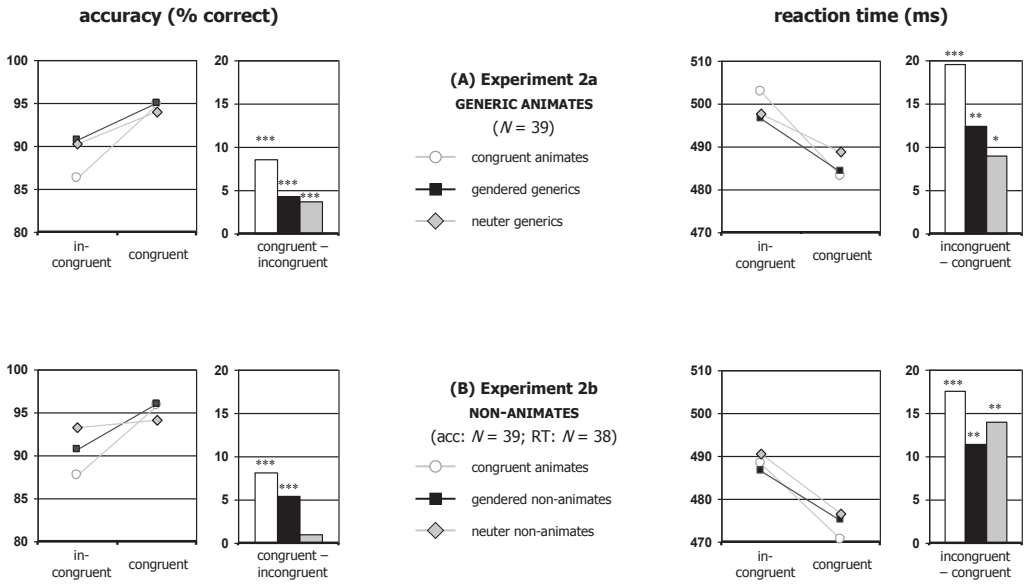

Figure 3. Results of Experiments 2a and 2b in terms of accuracy and reaction time (mean values and standard deviations are reported in Appendix C).

Figure 4

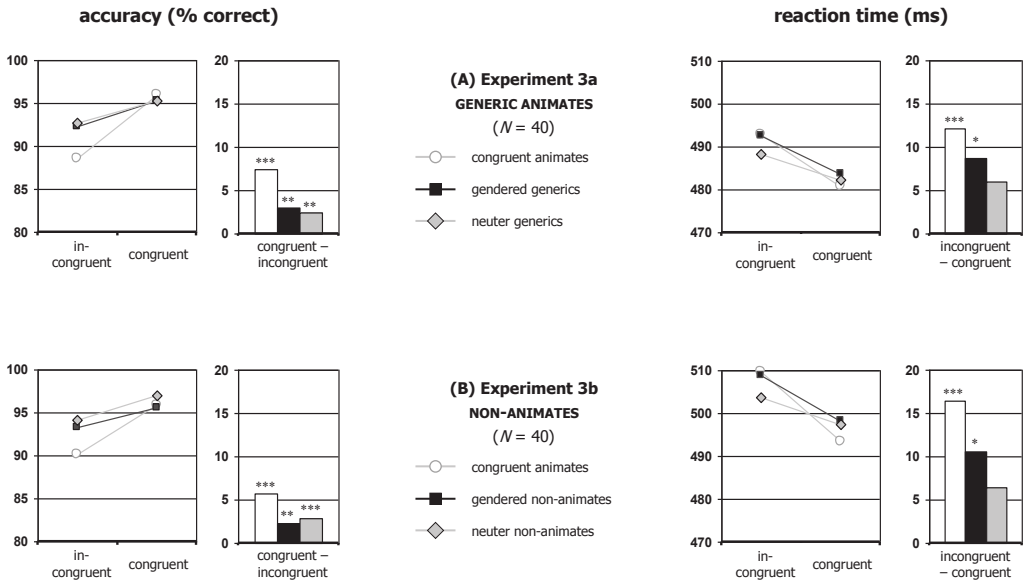

Figure 4. Results of Experiments 3a and 3b in terms of accuracy and reaction time (mean values and standard deviations are reported in Appendix C).