Grammatical gender and linguistic relativity: A systematic review.

Steven Samuel

Geoff Cole

Madeline J. Eacott

Department of Psychology, University of Essex, U.K.

"This is a post-peer-review, pre-copyedit version of an article published in Psychonomic Bulletin & Review. The final authenticated version will soon be available online"

Word count: 12,778 (excluding abstract, acknowledgements, references and tables, including footnotes and captions)

Key words: grammatical gender; Whorf; Linguistic relativity; language and thought

Address for correspondence: Steven Samuel. University of Essex, Department of Psychology, Wivenhoe Park, CO4 3SQ. Email: ssamuea@essex.ac.uk

Abstract

Many languages assign nouns to a grammatical gender class, such that 'bed' might be assigned masculine gender in one language (e.g. Italian) but feminine gender in another (e.g. Spanish). In the context of research assessing the potential for language to influence thought (the linguistic relativity hypothesis), a number of scholars have investigated whether grammatical gender assignment 'rubs off' on concepts themselves, such that Italian speakers might conceptualise beds as more masculine than Spanish speakers. We systematically reviewed 43 pieces of empirical research examining grammatical gender and thought, which together tested 5,895 participants. We classified the findings in terms of their support for this hypothesis, and assessed the results against parameters previously identified as potentially influencing outcomes. Overall, we found that support was strongly task- and context-dependent, and rested heavily on outcomes that have clear and equally-viable alternative explanations. We also argue that it remains unclear whether grammatical gender is in fact a useful tool for investigating relativity.

1. Introduction

The Sapir-Whorf or linguistic relativity hypothesis—henceforth 'Relativity' (Whorf, 1956)—takes various forms, but at its heart it contends that the idiosyncrasies of the languages we speak influence the way we think about the world. In its strongest incarnation—linguistic determinism—thought is constrained by language, but few if any contemporary scholars take this view (Athanasopoulos, 2009). At the opposite extreme is the 'universalist' position, in which thought is said to be independent of language (e.g., Pinker, 1994; See Lucy, 2016, for a historical overview). Although there is no agreement as to *where* between these two opposing views the truth is situated, there is broad consensus that neither extreme is correct (Gleitman & Papafragou, 2013). In the middle are a variety of standpoints; some more limiting of the role of language than others. For example, thinking *for* language (i.e. thinking about speaking) might allow language to bias our attention towards the more linguistically-describable aspects of what we perceive (Slobin, 1996), but this does not mean that language alters underlying conceptual structures. Another standpoint is that language can influence our *judgments* about what we perceive, but not perception itself, which is cognitively impenetrable (see Firestone & Scholl, 2016for debate; Pylyshyn, 1984). For many, our perceptions are perhaps best described as modulated or biased by language (Athanasopoulos, 2006; Dolscheid, Shayan, Majid, & Casasanto, 2013; Gilbert, Regier, Kay, & Ivry, 2006, 2008; Lupyan, 2012). Overall, investigating the ways in which language does and does not relate to thought now appears to be the prevailing approach (Lucy, 2016; Lupyan, 2012; Thierry, 2016).

Evidence for language influencing thought and judgments about perceptions comes from various domains. For example, performance on colour discrimination/matching tasks has been shown to be more efficient when the stimuli have different labels in a participant's language relative to when they are subsumed under one term (e.g. Roberson, Pak, & Hanley,

2008; Winawer et al., 2007). Differences between languages have also been demonstrated to affect how people think about object relations (Park & Ziegler, 2014) and objects themselves (Imai & Gentner, 1997) as well as broader, more abstract concepts such as quantity (Athanasopoulos, 2006), time (Boroditsky, 2001; Casasanto et al., 2004) and motion (Athanasopoulos & Bylund, 2013). Some of these studies have generated vigorous debate, perhaps most notably in the area of colour perception (e.g., A. Brown, Lindsey, & Guckes, 2011; Franklin, Clifford, Williamson, & Davies, 2005; Witzel & Gegenfurtner, 2013). For some, aspects of grammar might provide a better tool by which Relativity can be explored than vision-focussed research; this is because grammar is unaffected by sensory input, obviating the need for language to 'breach' psychophysical barriers. Additionally, for the Relativity hypothesis to hold it should not be limited to category labels but extend to features of syntax too (Sato & Athanasopoulos, 2018).

One aspect of grammar that has received attention in recent years is grammatical gender (Bassetti & Nicoladis, 2016; Bender, Beller, & Klauer, 2018). Unlike English, which has a semantic (or conceptual) gender system, whereby the gender of a noun is dictated by its biological sex, many languages have a formal system by which all nouns are assigned to a grammatical gender category whether they have biological sex or not. For example, the English word 'bed' has no gender, and is referred to with the pronoun 'it', but in Italian (*il*$_m$ *letto*) takes the masculine gender. The consequence of a formal grammatical gender system is obligatory conformity or agreement with the syntactic rules of that class. This may involve marking for gender any definite or indefinite articles ('the', 'a/an'), plural markings, case markings, and other forms of agreement. Different languages also assign different genders to the same objects; for example, in contrast to Italian, 'bed' takes feminine grammatical gender in Spanish (*la*$_f$ *cama*). Such differences are not at all exceptional; grammatical gender assignment in a given language is largely arbitrary (Corbett, 1991; Foundalis, 2002), and

'escapes logic' (Boutonnet, Athanasopoulos, & Thierry, 2012). Indeed, as English and other languages with purely semantic gender show, a formal grammatical gender system is also unnecessary when it comes to the communicative function of a language.

The crucial exception to arbitrariness in grammatical gender assignment concerns words referring to entities, usually human, with actual *biological* sex. 'Woman' is feminine and 'man' masculine across German, French, Italian and Spanish, and it is from these associations, presumably[1], that grammatical 'gender' receives its name. Occasionally, the correlation between grammatical gender and biological sex is imperfect; witness German, which has a third, neuter gender category in addition to masculine and feminine, and which assigns this gender to the word for 'girl' (*das*$_n$ *Mädchen*). Nevertheless, even in German the vast majority of words referring to humans which have biological sex show a strong pattern of masculine with males and feminine with females. In this sense, languages with a formal gender system can be viewed as having a largely semantic gender system for human agents at least.

It is the combination of the arbitrariness of grammatical gender assignment and the link with biological sex that has attracted researchers to grammatical gender when examining Relativity (e.g. Bassetti, 2007; Sato & Athanasopoulos, 2018). The question asked in such studies is whether grammatical gender assignment 'rubs off' on the conceptual representations of inanimate objects that have no biological sex such that, for example, Italian speakers conceptualise beds as somehow more masculine than speakers of a language with no formal grammatical gender system, or than speakers *with* a gender system but for whom a different gender is assigned. Although historically seen as a 'quirk' of grammar, grammatical gender is therefore regarded to be in a particularly strong position to inform long-standing

[1] We return to this issue and discuss it in greater detail in the Discussion section.

philosophical, linguistic and psychological debates around the universality or otherwise of human thought (Phillips & Boroditsky, 2003).

A variety of tasks have been employed to understand whether grammatical gender does influence concepts. The most common has been the Voice Choice task (sixteen different publications using this task were discovered in this review), in which participants are asked to assign a male or female voice to objects, with many finding that the sex of the voice and grammatical gender of the target are indeed broadly consistent (e.g. Kurinski, Jambor, & Sera, 2016; Lambelet, 2016; Ramos & Roberson, 2011; Sera, Berge, & del Castillo Pintado, 1994; Sera et al., 2002). Similar results have been found when asking participants to assign a human name or a sex to an object ('Sex Assignment' tasks, e.g., Belacchi & Cubelli, 2012; Flaherty, 2001), and when asking participants to rate on a scale the similarity ('Similarity Task') between pictures of male and female humans and objects (Phillips & Boroditsky, 2003). In Object-Name Memory Association tasks participants are instructed to remember male and female names that substitute object names, such that 'chair' might now be 'Patricia', and results have sometimes shown that the ability to recall the human name is enhanced if it is congruent with the grammatical gender of the object in question (Boroditsky & Schmidt, 2000).

The tasks described above all involve judgments made without time pressure or measurement, but there is also some, albeit limited evidence in favour of grammatical gender influencing concepts from speeded-response tasks. Converging evidence from different measurement types is important to generate the clearest picture possible of any effect. For example, in the Extrinsic Affective Simon Task ('EAST'), participants respond to words using two keys; in one condition, each key is mapped to a colour, and in another each key is mapped to a sex (male or female). Bender, Beller, and Klauer (2016b) found that German speakers were faster to respond to the colour of a word if that word was a gendered,

personifiable noun (e.g., death, beauty) and the response key was also mapped to the sex implied by the target's grammatical gender. However, they also found that the effect was weak or non-existent for inanimate objects, and only held for personifiable nouns that had connotations of sex that were themselves congruent with the noun's grammatical gender (e.g., 'war' is masculine in connotation and gender, but 'Spring' is feminine in connotation but masculine in gender).

The role of gender or sex in the tasks so far described is clear, as participants are making judgments and responses with biological sex playing an obvious role in this process. Some researchers have preferred tasks where the salience of sex and gender can be reduced or made more orthogonal to the participant's conscious experience. For example, Konishi (1993) employed a Property Judgment paradigm, finding that German speakers rated masculine-gendered objects such as 'key', 'table' and 'beach' as more 'potent' (a trait associated with masculinity) than concepts with feminine gender; Spanish speakers for whom the same targets took the *feminine* gender rated the same items less potent. This type of task makes participants think about concepts without having to relate them to gender or sex directly; indeed, neither term needs to come up in the instructions to such experiments at all.

Other studies, often using the same or similar designs, as those described above have instead reported potentially important absences of evidence. For example, in a Properties Judgment task, Landor (2014) asked approximately 600 speakers of gendered languages to generate adjectives to describe inanimate objects, and then asked a different group of participants to assign a male or female voice to them. Cross-referencing these results back to the original stimuli, no evidence of property judgments consistent with grammatical gender was found. Absences of evidence have also been reported in Sex Assignment tasks (Nicoladis & Foursha-Stevenson, 2012), Similarity tasks (Degani, 2007), the EAST (Bender et al., 2018), and Priming-type tasks (e.g. Degani, 2007; Samuel, Roehr-Brackin, & Roberson,

2016). The importance of these findings is that they might delineate important constraints on the Relativity hypothesis which, as discussed earlier, would be a finding consistent with the majority of theoretical standpoints in the contemporary literature.

Further clarifications of such constraints might come from results that do not conform to a binary 'support or no-support' distinction. For example, a Property Judgment task by Semenuks, Phillips, Dalca, Kim, and Boroditsky (2017) found that grammatical gender influenced property judgments but only in an analysis that left out participants' first-choice adjectives, focussing on second- and third-choice adjectives alone. The authors suggest that the absence of an effect from the first adjective choice might explain the tendency for more negative findings from speeded-response tasks, but an alternative explanation is that participants fell back on grammatical gender as a strategy after they had made their initial and potentially most faithful descriptions (cf. Bender, Beller, & Klauer, 2016a).

Other studies find support for Relativity only for limited subsets of results rather than the for the full range of predictions. For example, in the study by Konishi (1993) already described, German and Spanish speakers both showed evidence of perceiving objects with masculine gender as more 'potent' than objects with feminine gender, with 'potent' rated as a masculine trait; however, participants also rated 'negative' as a masculine trait, but this property did not transfer to objects. Indeed, even the link with potency involved very small differences between masculine and feminine concepts. Similarly, in a Sex Assignment task, Nicoladis, Da Costa, and Foursha-Stevenson (2016) found that Russian-English bilingual pre-schoolers assigned male sex to masculine gender objects more frequently than they did to neuter gender objects, but not more frequently than they did to feminine gender objects; and in a Voice Choice task, decisions have been found to be consistent with masculine gender but not with feminine gender (Bassetti, 2007).

As mentioned earlier, the prevailing standpoint is that neither a strong view of Relativity nor a universalist view, provides an accurate account of the relationship between language and thought, and as a result the absences of evidence and the mixed results hint at precisely the systematic constraints most researchers seek to reveal. It is our view that a systematic review of the empirical literature would be well-placed to begin the process of drawing conclusions concerning the strength, extent and limitations of any such relationship.

Before describing the results of the review, we outline the six potential constraints or 'parameters' that we have extracted from discussions in the literature. The assessment of the influence of these parameters, together with the effects of the large variety of tasks used in this field, run through our review. Finally, we consider whether the literature supports the view that the effects of grammatical gender are primarily about language, or whether the effects attributed to Relativity might instead be explained by statistical co-occurrences and associations that are *carried* by language, but are not linguistic themselves. Such questions have the potential to go to the heart of contemporary thought about the role of language in cognition (Lupyan, 2012; Thierry, 2016).

1.1. The salience of gender/sex in the task.

In terms of methodology, many studies that have found effects of grammatical gender have tended to come from explicit judgments when gender and/or sex is a salient context in the experiment (e.g. Phillips & Boroditsky, 2003; Saalbach, Imai, & Schalk, 2012; Sera et al., 1994; Sera et al., 2002). Indeed, the most common task in the field, Voice Choice, asks participants to assign a biological sex to an object. Given the nature of the question (i.e., there is no objectively correct answer), the participant may seek some rationale for their choices rather than assign at random. Under such conditions, the chances of grammatical gender being consciously recruited are increased, potentially undermining a conceptual change

account. A number of scholars have raised this issue (e.g., Bender, Beller, & Klauer, 2011; Bender et al., 2018; Kousta, Vinson, & Vigliocco, 2008; Pavlidou & Alvanoudi, 2013; Ramos & Roberson, 2011), and Phillips and Boroditsky (2003) have suggested that it is difficult to resolve empirically. For instance, in study 2 of Sera et al. (1994), reference to 'masculine' and 'feminine' was removed from the instructions from the previous experiment, but participants were nevertheless still required to assign male or female voices. Even when the role of gender/sex is less prominent in a task, it is often still high. For instance, the EAST (Bender et al., 2016a, 2016b) maps all responses onto the same keys, meaning that even when participants respond according to colour they do so using a key that has also been assigned to a biological sex. This is a necessary facet of a design that relies on mapping two concepts to the same response in order to test for any effect of the overlap. Any strategic use of grammatical gender, or simply its recruitment via strong associations with biological sex in a task, might undermine the case for conceptual change account of results, and it is clearly important to understand the extent to which evidence in favour of Relativity might rely on such possibilities by means of comparing high versus low-gender/sex salience in tasks.

1.2. The salience of language in the task.

Another issue concerns the salience of language in the task, and whether the design instead indexes language processes as in 'thinking for speaking' (Slobin, 1996). Indeed, it was an early requirement of language and thought research that effects of language should be evident on non-linguistic tasks (R. Brown & Lenneberg, 1954). As with the salience of gender/sex, it is important to understand the extent to which the evidence for Relativity might rest upon the recruitment of grammatical gender through language processing rather than any underlying conceptual change. We therefore classified research as either high or low in terms of the salience of language in the task.

1.3. Testing participants in their gendered language.

For some, testing that occurs in participants' gendered language limits inferences to effects of grammatical gender on concepts within that language, rather than concepts themselves (Boroditsky & Schmidt, 2000; Slobin, 1996); if one's language means that 'bed' is masculine, 'bed' should be conceptualised as masculine not only when acting in the context of that particular language but also when acting in a non-gendered language, such as English. By definition, this argument suggests that research is best conducted on bilinguals. In favour of such a possibility, testing participants in a second, ungendered language context has also revealed influences of a first-language gender system (Boroditsky & Schmidt, 2000; Phillips & Boroditsky, 2003; Semenuks et al., 2017). In order to test this particular hypothesis, we classified experiments in terms of whether participants performed in their gendered or non-gendered language.

1.4. Two-gender versus three-gender languages.

Another issue concerns the precise nature of the grammatical gender system under investigation. For example, in Romance languages like Spanish, Italian, and French most nouns that refer to humans carry grammatical gender that is consistent with the target's biological sex. In German, however, the correlation is weaker, in part owing to its third, neuter gender, but also because German articles do not always differentiate between genders as a result of the German case system. This can result, for example, in even animates being labelled with a grammatical gender incongruent with their biological sex. The issue of two- versus three-gendered languages therefore concerns whether an influence might be strongest in speakers of languages with two gender classes that form a particularly 'tight fit' with semantic gender (see Saalbach et al., 2012; Sera et al., 1994; Vigliocco, Vinson, Paganelli, &

Dworzynski, 2005). Results with German speakers on Voice Choice tasks have indicated that grammatical gender does not influence decisions in the way that, for example, French or Spanish grammatical gender does (Sera et al., 2002). If this pattern was borne out in a broader review, it might suggest a statistical, correlative relationship between biological sex and grammatical gender in Relativity[2]. For Lucy (2016), the structural consequences of grammatical gender (case markings, adjective agreement, etc.) are too often overlooked, and might explain some inconsistencies in results. We classified experiments in terms of whether participants spoke a two- or three-gendered language[3].

1.5. Effects with animate and inanimate targets.

Another parameter concerns whether grammatical gender might influence the conceptualisations of animate but not inanimate targets (e.g., Vigliocco et al., 2005). For

[2] The potential for grammatical gender effects on object conceptualisation to be the result of statistical co-occurrences goes to the heart of a debate around whether grammatical gender is in fact a suitable tool for investigating Relativity at all, and it is one we turn to in detail later in this section and in the Discussion.

[3] Not all two-gendered languages divide nouns into masculine and feminine; some (like Dutch, Swedish, and some Norwegian dialects) instead divide nouns into 'gendered' and 'neuter' categories. To our knowledge only one study in our review included participants who spoke a language of this type: Bergensk, a language spoken in Norway (Beller, Brattebø, Lavik, Reigstad, & Bender, 2015). For the purposes of the review, we excluded the results of this study from our two- vs. three-gender comparison. The sample size (N=107) suggests that neither leaving this study in nor taking it out would have any meaningful effect on the patterns of results described.

example, participants who speak a gendered language (German) showed a greater willingness to erroneously endorse sex-specific statements about animals if those statements were consistent with the animals' grammatical gender than speakers of an ungendered language (Japanese), but the same was not true of inanimate targets (Imai, Schalk, Saalbach, & Okada, 2014). In a series of priming experiments, Bender et al. (2011) asked German speakers to decide the gender of a target word after they had seen either i) a definite article denoting gender (*der* for masculine and *die* for feminine); ii) the words *Mann* (Man) and *Frau* (Woman); iii) the symbols for male and female; or iv) pictograms of a man or woman. They found that the congruent linguistic article primes sped up judgments relative to incongruent trials, for both animate and inanimate targets. However, of the other primes only the Mann/Frau pictograms had an effect, and only on animate targets. The researchers concluded that the grammatical gender of objects does not appear to 'seep' into the semantic content of inanimate nouns. A broadly similar pattern was found in the Properties Judgment task by Semenuks et al. (2017). Results like these have led some scholars to suggest that grammatical gender is only relevant to conceptualisation when sex is a relevant property of the target (Ramos & Roberson, 2011; Vigliocco et al., 2005). If this is true, then Relativity is subject to an important constraint, namely that grammatical gender only interacts with targets which have biological sex in the first place. We classified experiments in terms of whether participants responded to animate or inanimate targets.

1.6. Stronger effects in adults than children.

Finally, studies with adults have been thought to provide more supporting evidence for Relativity than studies with children, and particularly very young children. For example, only 6 out of 18 Spanish-speaking 3-5 year-olds freely sorted pictures of inanimate objects

and male and female people into groups defined by grammatical gender (Martinez & Shatz, 1996), and only very limited effects of grammatical gender on object conceptualisation, or indeed no effect at all, have been found in other studies with young children (Bassetti, 2007; Nicoladis & Foursha-Stevenson, 2012; Sera et al., 2002). Weaker effects at younger ages are consistent with the possibility that it is *experience* with grammatical gender that leads to biological sex connotations with objects, but also with the possibility that it is metalinguistic knowledge of grammatical gender acquired through formal instruction rather than grammatical gender assignment itself that might explain some positive results. We classified studies in terms of whether participants were children (<18) or adults.

1.7. An existential question for Relativity: What is 'gendered' about grammatical gender?

A crucial question that underpins everything in this review concerns the nature of the grammatical gender itself. It has been pointed out that the 'gender' in grammatical gender is not intrinsic to language but is itself an arbitrary, human-made label (Bender et al., 2018). At some point in their lives speakers of gendered languages usually learn that the formal names for the categories are 'masculine' and 'feminine', but would they ever choose those names without formal instruction? As highlighted by Foundalis (2002) 'masculine' and 'feminine' are in fact poor predictors of the majority of nouns in their class, and even the relationship between the meanings of the words *gender* and *sex* are stronger for speakers of non-gendered languages, such as English; speakers of gendered languages would not usually use the translation equivalent for 'gender' in grammatical gender to refer to biological sex at all.

Grammatical gender has usually been perceived as a particularly useful tool to study Relativity because its assignment patterns are so arbitrary and have no psychological reality outside of language itself, but the same point could be levelled with more detrimental

consequences for the research paradigm at the titles of the categories themselves, which are metalinguistic labels (they *describe* something about language) detached from the use of the grammar itself. Grammatical gender therefore appears to suffers from an identity problem that other tools used to investigate Relativity do not. To illustrate, we might as a thought experiment substitute the labels 'masculine' and 'feminine' for either 'plant' and 'non-plant', 'sky' and 'earth', 'group one' and 'group two', 'x' and 'y', or any number of arbitrary labels without interfering with the performative use of a gendered language itself. We might predict that if we told a cohort of Spanish speakers to go about their day 'believing' that the titles of the classes had changed to 'x' and 'y', they could simply continue to refer to *la mesa* (the$_f$ table) and *el libro* (the$_m$ book) with no real-world consequences. If however we asked the same cohort to randomly shuffle their colour words mappings for a day such that they might need to refer to the Spanish word for 'green' with the Spanish word for 'blue', or to swap their spatial prepositions such that 'over' might become 'under', we ask them to violate the rules of language *use*, and we would expect there to be errors.

If the labels 'masculine' and 'feminine' are in a sense historical accidents, it has consequences for the use of these categories in the Relativity paradigm because effects that have previously been attributed to conceptual change might in fact be the result of simple statistical co-occurrences and associations between two groups of labels: the entirely arbitrary, human-made labels of metalinguistic grammatical classes on the one hand, and the similarly arbitrary 'gender' assignment to noun labels on the other. This would be the equivalent, for example, of finding that English speakers conceive of the digits 3, 5, 7 and 9 as somehow 'weirder' that 2, 4, 6 and 8 because the former are arbitrarily labelled as 'odd'. While language would be the *vehicle* of any statistical associations, the outcome becomes trivial in the context of classic views of Relativity as engaging *conceptual* change.

What patterns would a statistical co-occurrence account predict? We would expect five out of the six parameters described to influence results. Firstly, speakers of a two-gendered language should show stronger effects than speakers of three-gendered languages because the relationship between human biological sex and grammatical gender is reinforced through greater repetition and a stronger gender/sex correlation in human animates, at least when compared to the three-gendered language most commonly used in the field (German). We would additionally predict that performing in a gendered language would give rise to such associations in a way that performing in a non-gendered language like English might not. We would predict that the more salient the role of both gender/sex and language in the experiment, the greater the opportunity for associations between biological sex and language to be engaged. Finally, we would expect the presence of animate targets to elicit biological sex information more than inanimate targets. A sixth possibility, albeit a theoretically more tenuous one, is that adults will have more strongly reinforced associations than children owing to their greater quantitative experience of language. Interestingly, *all* these parameter settings have already been cited in the literature as conducive to finding effects of grammatical gender, although these suggestions have not yet been supported by a systematic review. Assuming that the strong version of Relativity is wrong (as we do) it then becomes important to decide whether effects of grammatical gender are statistical and associative or an effect of language on the fabric of conceptual representation itself.

2. Review

2.1. The remit of the review.

This review includes i) *empirical* research with ii) *human* participants; iii) *real* languages and words (rather than languages and words invented for the purposes of

experimentation); iv) being either *published* or *unpublished*; and v) reported in *English*[4]. Studies were also required to test the influence of grammatical gender on at least some non-human targets (excluding, therefore, studies on grammatical gender and gender stereotyping of men and women).

Formal searches were conducted encompassing the years 1990 to 2018 using the search terms "grammatical" and "gender" together, once with "Whorf" and once with "relativity", in Web Of Science (all) and Google Scholar (first 5 pages). Additionally, we searched the EThOS PhD thesis bank using the terms "grammatical" and "gender" for the broadest possible range of results and the NDLTD thesis bank using "grammatical gender", once with "Whorf" and once with "relativity". All studies from a recent special issue of the International Journal of Bilingualism (volume 20/1) were included where relevant. To ensure maximal representation of relevant data and to minimise the potential for skewed results owing to potential publication biases (e.g. De Bruin, Treccani, & Della Sala, 2015), we also emailed the corresponding author of every study revealed in the first stage of the review to request any unpublished or in-press results. Finally, a call for data was also issued on the Research Gate website as part of a project linked to the present review. A further thirteen pieces of empirical research not turned up by these searches were added either because they were known to the authors or offered/recommended to the authors as part of this contact phase. The full list of items included and excluded from the review can be found in the Supplemental Online Materials (SOM1).

2.2. Classifications by task.

[4] Only one study discovered by the search procedure was eventually excluded for the absence of an English translation (see SOM1)

Owing to the heterogeneity of methodologies in the field, we opted to sort each individual experiment into eight task types: Voice Choice; Properties Judgments; EAST; Sex Assignment; Priming; Similarity Judgment; Association; Object-Name Memory Association. These eight task types were chosen because they were different enough in methodology to be considered in their own right. To illustrate, we felt that the closest pair was Voice Choice/Sex Assignment, since both involve explicit biological sex judgements to be made about targets. However, Voice Choice tasks require the participant to imagine an object *speaking*, which might recruit thinking about language in a way that assigning biological sex alone might not[5].

We classified every experiment according to the six parameters previously described as potentially important to the outcomes of research. We adopted the following approach to these classifications. Regarding Language Content, we asked whether language goes *in* to the task (for example, the stimuli are words), or *comes out* of the task (for example, choosing between words such as 'male' or 'female', thinking of adjectives in Property Judgment tasks). Where neither occurs, or where any role of language is judged to be highly orthogonal, that study is classified as Low Language Content. The same approach was taken to Gender/Sex Content (gender or sex information should not go in to the task or be part of the response or process leading to the response). The other parameters (Age, Language, Number of Gender Categories, and Target Type) were readily classifiable at face value. Full details of item classifications can be found in SOM2.

2.3. Classifications of results

[5] Additionally, there is at present no empirical evidence that participants assign the same biological sex to the same objects across both tasks.

Given that the results of tasks are sometimes not unambiguous in their support or otherwise of Relativity, each individual study was classified in terms of one of *three* outcomes: Support, Mixed Support, or No Support. Studies classified as offering *Support* showed an effect of grammatical gender on conceptualisations consistent with the hypothesis employed. A study was classified as offering *No Support* if there was an absence of evidence for Relativity that could be classified minimally as Mixed Support. Studies were classified as offering *Mixed Support* if they showed partial confirmation of an influence, such as an influence of one gender but not another (Bassetti, 2007), marginally significant effects (e.g. Semenuks et al., 2017), an influence in accuracy but not in response times (e.g. Bender et al., 2016a; Bender et al., 2016b), or an influence shown on one criterion but not another (e.g. Konishi, 1993). SOM2 lists these classifications for each experiment.

2.4. Adjustments for sample size

Given that sample sizes varied widely from study to study and condition to condition (from 7 to 924), we report results taking sample size into account. This has the obvious benefit of weighting the pattern of results in favour of those studies that are most likely to be highest-powered and less susceptible to spurious effects. Given that the same participants sometimes performed multiple conditions or analyses, we also allowed multiple data points in the review from the same participants. For example, this review allowed for separate data points for inanimate and animate targets from the same group. In such cases, separate classifications are necessary to ascertain the effect of different parameters, such that the evidence from animate targets might be classified as offering Support, but the results from inanimate targets classified as No Support. Full details can be found in SOM2.

This review adopts a methodological approach somewhere between a vote-count system (owing to the classifications of Support, Mixed Support or No Support) and a

statistical meta-analysis (owing to sample-size adjustments) (cf. Samuel, Roehr-Brackin, Pak, & Kim, 2018). Given the heterogeneity of research methods, languages tested, and so on, occasionally only very small clusters of studies could be considered to be using the same methods, and often these were studies that came from the same labs and sampled the same type of linguistic population. We therefore considered this approach the better way to provide an overview of the multiple ways in which the field has investigated the research question, and how different designs may culminate in different outcomes.

2.5. What is not *in the review.*

We excluded studies that looked not for relationships between targets and biological sex but rather for relationships between objects of the same grammatical gender versus objects of different grammatical gender (e.g., Almutrafi, 2015, exp. 2; Bobb & Mani, 2013; Boutonnet et al., 2012; Cubelli, Paolieri, Lotto, & Job, 2011; Kousta et al., 2008; Yorkston & De Mello, 2005). For example, it has been demonstrated that semantic category judgments about nouns belonging to the same grammatical gender are processed more quickly than nouns from different grammatical gender (Cubelli et al., 2011), and that grammatical gender information is processed in semantic similarity tasks even when irrelevant and undetectable by behavioural measures (Boutonnet et al., 2012). Although such studies offer support for the processing of grammatical gender information when it is apparently task-irrelevant (a useful prerequisite for tasks in this review), and even show that objects of the same grammatical gender are perceived to be more similar than objects which are not, this is not the same as demonstrating that objects are conceptualised as more *masculine* or *feminine* as a function of their gender assignment. Such results might be explained in terms of an effect of membership of the same grammatical category, independently of biological sex information (cf. Cook,

2016). As this review is entirely concerned with this specific relationship, such studies, though clearly interesting in their own right, were omitted.

There have also been studies assessing how and when grammatical gender is processed in language production and comprehension, including in bilinguals for whom the grammatical gender for the same object might be opposite in their two languages (Costa, Kovacic, Fedorenko, & Caramazza, 2003; Costa, Kovacic, Franck, & Caramazza, 2003). Again, such studies are not designed to look at whether object conceptualization has the potential to be influenced by the *biological* sex connotation of their grammatical gender assignment, and they were therefore excluded.

The review does not include one study which *does* pertain to grammatical gender and biological sex but where a meaningful understanding of the number of participants is not possible. This was the study by Segel and Boroditsky (2011), in which depictions of personifications and allegories in thousand or works of art were classified retrospectively in terms of their gender congruency. This was notionally classified as Support.

Finally, the remit of the review excludes studies that do not involve human participants but investigate the question of Relativity through the connectionist models (e.g., Dilkina, McClelland, & Boroditsky, 2007; Sera et al., 2002), or through training in artificial 'languages' (Eberhard, Heilman, & Scheutz, 2005; Phillips & Boroditsky, 2003, experiments 4 and 5; Sera et al., 2002, experiment 4) or nonsense words (Konishi, 1994; Vuksanovic, Bjekic, & Radivojevic, 2015). This is because we felt that we should limit our scope to behaviour more clearly grounded in the experience of real people with real grammatical gender categories.

3. Results

Overall, the initial search found 99 individual pieces of research, with a further 13 added that were known to the authors but were not revealed by the formal search. After removing those items which did not provide empirical data, the review included 43 individual pieces of research, one of which was unpublished (Nicoladis, unpublished), and three of which were doctoral or masters theses (Almutrafi, 2015; Degani, 2007; Landor, 2014). The remaining pieces of research were published journal articles, conference proceedings (always of the Cognitive Science Society) or book chapters. After subdividing this research by task type and condition, these pieces of research resulted in 158 lines of data (split by differences in conditions within experiments) which together surveyed 5,895 participants in total.

As described earlier, we then calculated the number of 'samples', which was 7,334. This number differs from the number of participants because it allows for the possibility that the same participant may have performed in multiple conditions. It is for this reason that the number of samples can be higher (but not lower) than the number of participants. We present all our results in the context of *samples*, rather than participants, in order to capture these important within-experiment differences.

3.1. Overall results.

Across the review as a whole, results from 32% of all samples were classified as offering Support for Relativity, 24% were classified as offering Mixed Support, and 43% No Support. With the exception of one particularly large study (Montefinese, Ambrosini, & Roivainen, 2018, N = 924 and N = 105, total N = 1029), there was no evidence that the results were driven by only a small cluster of highly-powered studies (see Figure 1). If this outlying study were removed, Support would be at 38%, Mixed Support at 28%, and No Support at 34%.

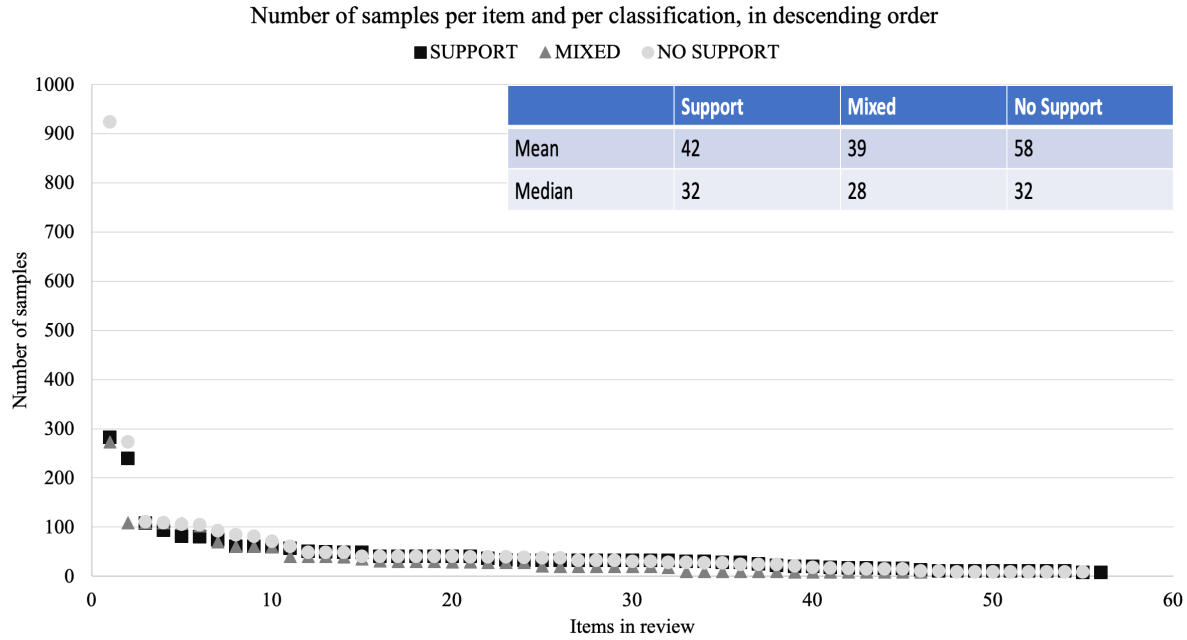

Figure 1. Number of samples (vertical axis) for each research item, or line of the review (horizontal axis). Mean and median samples are displayed in the text box. The outlier is Montefinese et al. (2018) with a Properties Judgment task (N = 924 Italian speakers).

In what follows, we first describe the results by task type. We then describe results as a function of task parameters. A full at-a-glance view of all the results by task type and by parameter can be found in Figures 2 and 3.

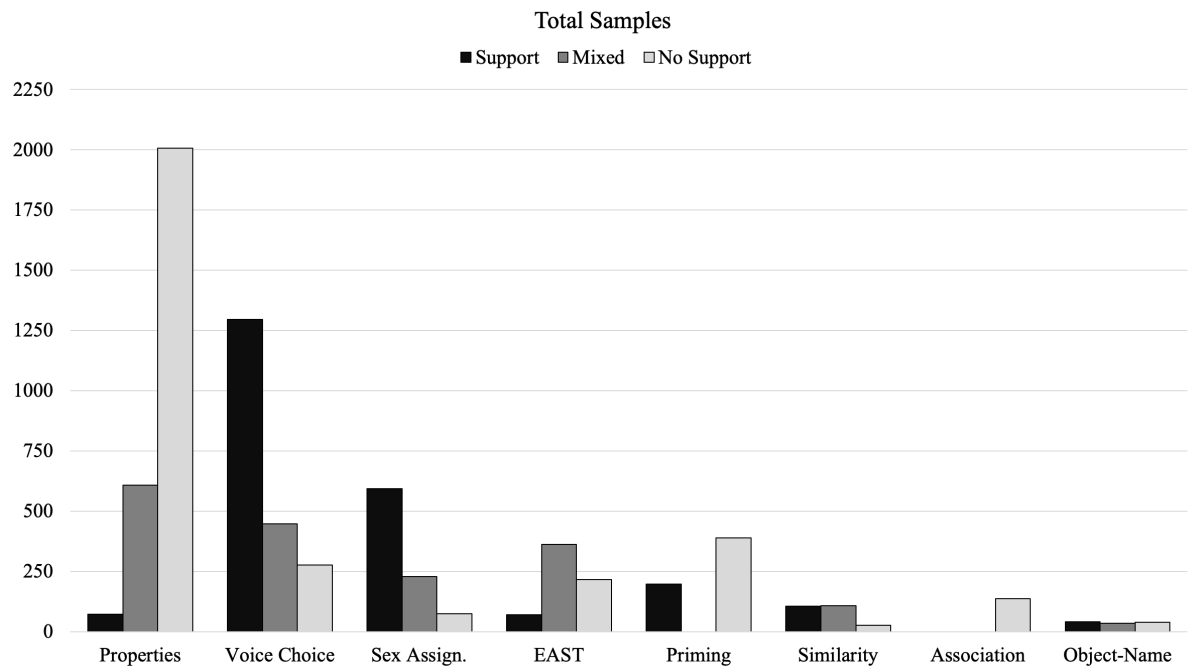

Figure 2. Classification of Support, Mixed Support and No Support by task type. Total number of samples is shown on the y-axis.

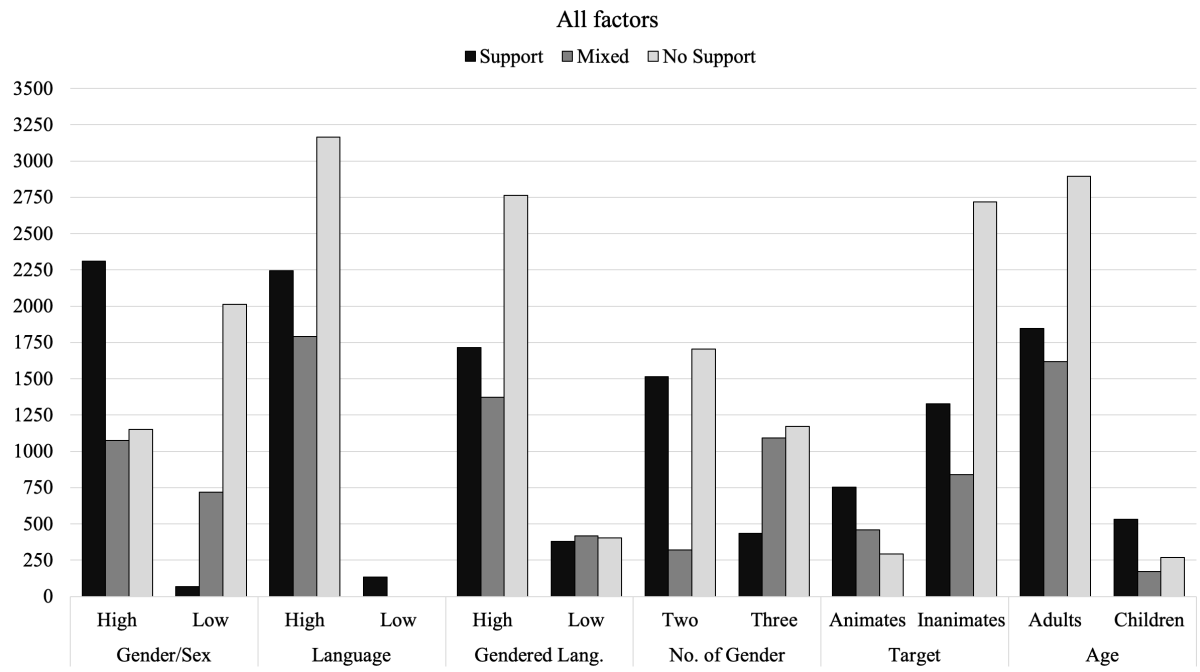

Figure 3. Classification of Support, Mixed Support and No Support according to task parameters. Total number of samples is shown on the y-axis.

3.2 Results by task type

Properties Judgment. Full results for the Properties Judgment task are displayed in Table 2. Properties Judgments made up 37% of all samples in the review, making it the most commonly-performed task in the literature (that is, it comes with the highest number of samples, rather than the highest number of uses in the literature). Only 3% of samples were classified as offering Support (Flaherty, 2001; Imai et al., 2014; Saalbach et al., 2012). A further 23% were classified as Mixed Support; reasons were finding that results were more consistent with grammatical gender in one group than another (that spoke a different language), but apparently not more so than chance itself (Haertlé, 2017), results limited to one property but not another despite evidence that both were linked to biological sex (Konishi, 1993), evidence to suggest an effect of the grammatical gender of a language the participants did not speak, with no direct comparison of this effect with the language they did speak (Sedlmeier, Tipandjan, & Jänchen, 2016), and effects limited to second- and third-choice adjectives but not first-choice adjectives (Semenuks et al., 2017). The remaining 75% of samples offered cases of No Support at all (Flaherty, 2001; Imai et al., 2014; Landor, 2014; Mickan, Schiefke, & Stefanowitsch, 2014; Montefinese et al., 2018; Semenuks et al., 2017). It should be noted that the study by Montefinese et al. (2018) represents an extreme outlier, with over 1000 participants in total. The results of this study were classified as No Support. However, even if the results of this study were removed, the rate of Support for Properties Judgment tasks would only rise to 4%. All Properties Judgment tasks were intrinsically High in Language Content, and all have so far been conducted with adult participants. Given the almost floor-level overall rate of Support, an examination of the effect of different factor parameters was not conducted.

Table 1. Number of samples according to parameter setting and results classification, for the Properties task. NB – samples are repeated multiple times across parameters, hence difference from bottom-line totals.

Parameter	Setting	Support	Mixed	No Support	TOTAL
Gender/Sex	High	33	0	65	98
	Low	40	609	1942	2591
Gend. Lang.	High	73	336	1734	2143
	Low	0	273	273	546
No. of Gend.	Two	40	60	1337	1437
	Three	33	276	397	706
Target	Animate	33	273	24	330
	Inanimate	0	120	1943	2063
Age	Adults	57	609	1967	2633
	Children	16	0	40	56
TOTAL		73	609	2007	2689

Voice Choice. Full results for the Voice Choice task are displayed in Table 2. The Voice Choice paradigm, while the most common experimental task for researchers, is the second most commonly-performed by participants in the field, accounting for 28% of all samples. Of these, 64% were classified as offering Support (Almutrafi, 2015; Athanasopoulos & Boutonnet, 2016; Beller et al., 2015; Bender et al., 2016a; Haertlé, 2017; Kurinski et al., 2016; Lambelet, 2016; Ramos & Roberson, 2011; Sera et al., 1994; Sera et al., 2002; Vernich, 2017; Vernich, Argus, & Kamandulytė-Merfeldienė, 2017). An additional 22% were classified as Mixed Support; these included results consistent with the hypothesis but limited to one of two genders (Bassetti, 2007), effects for limited subsets of targets (Beller et al., 2015; Bender et al., 2016a), and statistically marginal results (Bender et al., 2018). Mixed

results also included cases where data suggested that voice choices were not more consistent with grammatical gender than chance levels (Forbes, Poulin-Dubois, Rivero, & Sera, 2008; Sera et al., 2002), and effects limited to native speakers but not learners (including advanced learners[6]) of the same language (Kurinski & Sera, 2011). A minority of samples offered No Support at all (12%) (Bassetti, 2007; Bender et al., 2018; Forbes et al., 2008; Sera et al., 1994; Sera et al., 2002).

All Voice Choice tasks were classified as having a High Gender content and High Language Content. A greater rate of support for Relativity was found when participants spoke a language with two genders (69%) instead of three (24%). There was also more support from adult samples (69%) than children (39%). To illustrate these comparisons, the majority of No Support cases came from participants who were 5-6 year-olds (Sera et al., 1994; Sera et al., 2002), 9 year-olds, (Bassetti, 2007), or adult speakers of German, a three-gendered language (Bender et al., 2018). In contrast, there was no evidence that voice choices were more consistent with grammatical gender when applied to animate (38%) than inanimate objects (68%), and a lower rate of Support was found when assigning voices in one's gendered language (56%) relative to one's ungendered language (73%). Some of these results might be skewed by the relative dearth of child samples, samples who performed the task in a low

[6] This study included 102 participants of varying levels of Spanish. Any effect of grammatical gender in the advanced learners in this experiment was confounded with the usually more reliable natural/artificial distinction (Mullen, 1990). Note that if this item were included as Support owing to the 26 native-speaking participants' performance alone, it would only move these 26 samples (the number of native Spanish speakers) from Mixed Support to Support.

gendered-language context, with a three-gendered language background, or with animate targets.

Table 2. Number of samples according to parameter setting and results classification, for the Voice Choice task. NB – samples are repeated multiple times across parameters, hence difference from bottom-line totals.

Parameter	Setting	Support	Mixed	No Support	TOTAL
Gend. Lang.	High	837	415	245	1497
	Low	177	32	32	241
No. of Gend.	Two	828	187	180	1195
	Three	79	153	97	329
Target	Animate	107	99	76	282
	Inanimate	1189	348	201	1738
Age	Adults	1160	372	140	1672
	Children	136	75	137	348
TOTAL		1296	447	277	2020

Sex Assignment. Full results for the Sex Assignment task are displayed in Table 3. Sex Assignment tasks are almost identical to Voice Choice tasks in that they are explicit assignments of male or female sex to animate and inanimate objects. Largely thanks to the study by Belacchi and Cubelli (2012), representing 412 samples (46% of all samples with this paradigm), Sex Assignment makes up 12% of samples in the review.

Overall, 66% of samples were classified as offering Support (Belacchi & Cubelli, 2012; Flaherty, 2001; Pavlidou & Alvanoudi, 2018; Sera et al., 1994), 26% Mixed Support (Bender et al., 2016b; Nicoladis, unpublished; Nicoladis et al., 2016; Nicoladis & Foursha-Stevenson, 2012; Pavlidou & Alvanoudi, 2013), and 8% No Support (Flaherty, 2001; Nicoladis & Foursha-Stevenson, 2012).

Sex Assignment tasks are intrinsically classified as High in Gender/Sex content. Although they need not also have a High Language content, they were judged High for every study in this review. For animate targets the rate of Support was 100%, higher than for inanimate targets (33%). The rate of Support in children (62%) was slightly higher than in adults (69%). The rate of Support was 75% when participants performed in their gendered language but zero in an ungendered language. Finally, the rate of Support from two-gendered languages was high (83%), but from three-gendered languages was low (23%). Again, some of these comparisons (with the probable exception of Age) might be skewed by imbalances in the number of samples that were classified as performing under each parameter setting

Table 3. Number of samples according to parameter setting and results classification, for the Sex Assignment task. NB – samples are repeated multiple times across parameters, hence difference from bottom-line totals.

Parameter	Setting	Support	Mixed	No Support	TOTAL
Gend. Lang.	High	594	151	48	793
	Low	0	78	27	105
No. of Gend.	Two	538	38	75	651
	Three	56	191	0	247
Target	Animate	412	0	0	412
	Inanimate	30	61	0	91
Age	Adults	214	131	0	345
	Children	380	98	75	553
TOTAL		594	229	75	898

EAST. Full results for the EAST are displayed in Table 4. The EAST is a unique case in this review because it has only been used by one core group of researchers (Bender et al., 2016a, 2016b, 2018), and only ever with adult speakers of German, which is a three-gender language. It comprises 9% of samples. There is a good case that it might constitute a priming

task, but given its singularity we felt it was best considered as a category of task in its own right. Overall, only 11% of samples were classified as offering Support (Bender et al., 2016a, 2018), while 56% were classified as offering Mixed Support (Bender et al., 2016a, 2016b, 2018) and 33% No Support (Bender et al., 2016a, 2018).

Variation in results by parameter should be interpreted in the context of the low overall rate of Support. The EAST uses a two-key response method, with one clearly mapped to 'male' and one to 'female', hence Gender Context is always High, and since the stimuli are always words, Language Context is also High. In fact, the only possible parameter comparison that can be made with the EAST concerns inanimate and animate targets., which showed similar levels of Support (10% vs. 11% respectively).

Table 4. Number of samples according to parameter setting and results classification, for the EAST.

Parameter	Setting	Support	Mixed	No Support	TOTAL
Target	Animate	30	89	146	265
	Inanimate	40	274	70	384
TOTAL		70	363	216	649

Priming. Full results for Priming tasks are displayed in Table 5. There are only five different pieces of research with priming experiments in the literature, comprising 8% of samples in the review. Care must be taken when attempting to interpret parameter patterns from only a handful of studies where settings can be entirely confounded with individual papers. Overall, Priming offered a 34% Support rate (Bender et al., 2011; Sato & Athanasopoulos, 2018) and 66% No Support rate (Bender et al., 2011; Degani, 2007; Mickan et al., 2014; Samuel et al., 2016). There were no cases of Mixed classifications. Given the

small overall numbers of Support (198 samples in total), coming from only two papers, comparisons were unlikely to reveal any reliable patterns.

Table 5. Number of samples according to parameter setting and results classification, for the Priming tasks. NB – samples are repeated multiple times across parameters, hence difference from bottom-line totals.

Parameter	Setting	Support	Mixed	No Support	TOTAL
Language	High	170	0	389	559
	Low	28	0	0	28
Gender/Sex	High	170	0	317	487
	Low	28	0	72	100
Gend. Lang.	High	142	0	317	459
	Low	56	0	72	128
No. of Gend.	Two	56	0	67	123
	Three	142	0	306	448
Target	Animate	170	0	48	218
	Inanimate	28	0	341	369
TOTAL		198	0	389	587

Similarity Judgment. Full results for Similarity tasks are displayed in Table 6. Similarity Judgment tasks comprised only 3% of all samples. The pattern of results indicating 44% Support (Phillips & Boroditsky, 2003), 45% Mixed Support (Sedlmeier et al., 2016), and 11% No Support (Degani, 2007) is entirely confounded with the three individual pieces of research to use the task. The small number of samples with this task make it difficult to draw meaningful conclusions as to what might lead to differences in results.

Table 6. Number of samples according to parameter setting and results classification, for the Similarity tasks. NB – samples are repeated multiple times across parameters, hence difference from bottom-line totals.

Parameter	Setting	Support	Mixed	No Support	TOTAL
Language	High	0	108	27	135
	Low	105	0	0	105
Gender/Sex	High	105	0	27	132
	Low	0	108	0	108
Gend. Lang.	High	0	108	27	135
	Low	105	0	0	105
No. of Gend.	Two	29	0	27	56
	Three	40	108	0	148
TOTAL		105	108	27	240

Association. Full results for Similarity tasks are displayed in Table 7. Only two pieces of research have employed an Association paradigm (Bender et al., 2018; Martinez & Shatz, 1996), which together comprise only 2% of samples in the review. All the studies with this paradigm were classified as No Support. Gender content and Language content were always High, and participants were always tested on inanimate targets and in a gendered-language context.

Table 7. Number of samples according to parameter setting and results classification, for the Association tasks. NB – samples are repeated multiple times across parameters, hence difference from bottom-line totals.

Parameter	Setting	Support	Mixed	No Support	TOTAL
No. of Gend.	Two	0	0	18	18
	Three	0	0	119	119
Age	Adults	0	0	119	119
	Children	0	0	18	18
TOTAL		0	0	137	137

Object-Name Memory Association. Full results for Object-Name Association tasks are displayed in Table 8. Comprising just under 2% of samples in the review, Object-Name Memory Association paradigms form the smallest task-type classification in this review. The task has been used in three separate pieces of research. Of the samples, 36% came under Support (Boroditsky & Schmidt, 2000), 31% Mixed Support (Kaushanskaya & Smith, 2016), and 33% No Support (Pavlidou & Alvanoudi, 2013). Note that these differences are split entirely by publication. All these studies were classified as having a High Gender content and High Language content. All were conducted with adult participants, and all included inanimate targets.

Table 8. Number of samples according to parameter setting and results classification, for the Object-Name Association tasks. NB – samples are repeated multiple times across parameters, hence difference from bottom-line totals.

Parameter	Setting	Support	Mixed	No Support	TOTAL
No. of Gend.	Two	25	35	0	60
	Three	16	0	38	54
Gend. Lang.	High	0	0	38	38
	Low	41	35	0	76
TOTAL		41	35	38	114

3.3. Results by task parameters

The distribution of samples displayed in Figure 3 points to a number of imbalances in the literature to date; samples in this review were typically involved in experiments with a high language content (98%). Samples were usually adults (87%), performing in their gendered language (83%), with inanimate targets (76%), with high sex/gender salience in the task (62%). Samples also usually spoke a language with two gender categories (57%). In other words, the average experiment incorporates five out of six of the parameter settings that are usually considered most conducive to results in support for Relativity (inanimate targets being the exception).

Changes in the rate of Support as a function of task parameters are displayed in Table 9. In the following sections we describe comparisons where it is possible to isolate one category from another. A study that puts speakers of two-gendered and three-gendered languages into the same group, for example, are excluded entirely rather than added to both categories.

Table 9. Summary of shifts in the rate of Support as a function of task parameters. The rightmost column displays the outcomes in terms of the indicators of a statistical association

account rather than a Relativity account. NB – 'Language' was almost entirely (98%) set at High.

Parameter	Setting	Support	Statistical Association Account
Gender/Sex	High	51%	(+49% Support) = *Consistent*
	Low	2%	
Language	High	N/A	*N/A*
	Low	N/A	
Gend. Lang.	High	29%	(-3% Support) = *Not consistent*
	Low	32%	
No. of Gend.	Two	43%	(+27% Support) = *Consistent*
	Three	16%	
Target	Animate	50%	(+23% Support) = *Consistent*
	Inanimate	27%	
Age	Adults	29%	(-26% Support) *N/A*
	Children	55%	

Gender/Sex content. Consistent with previous views of the literature, as well as with a statistical association account of grammatical gender effects, studies with a High gender/sex content showed a higher rate of support (51%) than studies with Low gender/sex content (2%). The Voice Choice, Sex Assignment, EAST, Association, and Object-Name association task types were always classified as High. Samples classified as Low came almost entirely from the 2689 (96%) who performed Properties Judgment tasks, which came with only a 3% Support rate. However, only 98 samples from this paradigm were classified as *High*, rendering any more detailed comparisons unreliable.

Language content. Almost all research was classified as High in Language content (98%). The almost complete absence of research classified as Low in Language Content makes any attempt to draw conclusions about this parameter liable to mislead. Only the

Priming and Similarity studies by Sato and Athanasopoulos (2018) and Phillips and Boroditsky (2003) respectively, both of which were classified as Support, were classified as Low in Language content. We return to this issue in our Discussion.

Gendered vs. Ungendered language. A slightly lower rate of Support was found for studies performed in a gendered language (29%) than ungendered language (32%). This is not consistent with a statistical association account. When we compare performance in gendered and ungendered languages at the within-task level, we see that Support is higher in a gendered language context than an ungendered language context in the Sex Assignment task (75% vs. 0%) and the Properties tasks, albeit the latter with very low rates (3% vs. 0%). Support is slightly higher in the *ungendered* language for Voice Choice (73% vs. 56%) and Priming tasks (44% vs. 31%). It is also higher for Similarity tasks (100% vs. 0%), and Object-Name tasks (54% vs. 0%). Although 83% of all samples performed in a gendered language context, meaning comparisons are based on imbalanced sample sizes, the pattern of results within tasks suggest that there is no clear support for the hypothesis that grammatical gender is more likely to influence thought when participants perform in a gendered language.

Two-gender vs. Three-gender languages. The review found higher rates of Support from studies with two-gender languages (43%) than three-gender languages (16%). This outcomes is consistent with a statistical association account. Broken down by task type, Support from two-gendered languages over three-gendered languages came from the Sex Assignment tasks (83% vs 23%), Voice Choice tasks (69% vs. 24%), Similarity tasks (52% vs. 27%), Priming tasks (46% vs. 32%), and Object-Name Association tasks (42% vs. 30%). Only Properties tasks (3% vs. 5%) reversed this pattern, albeit with negligible Support rates in each category.

Animate vs. Inanimate targets. Consistent with a statistical association account, studies with animate targets showed a higher rate of Support (50%) than studies with inanimate targets (27%). Broken down by task, this pattern was true of Sex Assignment tasks (100% vs. 33%), Priming tasks (78% vs, 8%), and Properties tasks (10% vs. 0%). Results from the EAST were almost matched (11% vs. 10%). The reverse pattern was found for Voice Choice tasks (38% vs. 68%). Note that the great bulk of the positive results from *inanimate* targets comes from the Voice Choice task (90% of samples).

Adults vs. Children. In apparent contrast with some views expressed in the literature, research with children (55%) found a higher rate of Support than research with adults (29%). However, this result is strongly weighted by task type. Overall, 87% of samples came from adult participants, and the great bulk of the data from children comes from Voice Choice and Sex Assignment tasks (92%). Given that the tasks that children performed provided the highest rates of Support, and those performed almost exclusively by adults provide the lowest (e.g., Properties Judgments), it is difficult to know whether this outcome is the result of an age-related difference or a task-related difference. Looking at Age-related performance at the level of the individual task, there is some evidence that adults show a greater influence of grammatical gender than children; we see that the rate of Support is 30% higher in adults than children in Voice Choice tasks (69% vs. 39%), although it is 7% lower in adults Sex Assignment tasks (62% vs. 69%).

3.4. Other patterns

Almost half of all the data in the review (40%) came from the Voice Choice and Sex Assignment tasks, the two paradigms that make the clearest demand on participants to

consider targets in terms of biological sex. Since the potential for the strategic use of grammatical gender under such circumstances has been one of the most frequent issues brought up in the literature, we compared the rate of Support from the review as a whole with the results of these two paradigms included or excluded. Overall Support across the review drops from 32% to only 11% in their absence, and No Support rises from 43% to 64%. In other words, when all the data are included approximately one in three samples in the review provides Support; when the data from Voice Choice and Sex Assignment are excluded this rate drops to one in ten.

4. Discussion

At its broadest, the review finds that the evidence for an influence of grammatical gender on conceptualisations is highly task- and context-dependent. We found that the Voice Choice and Sex Assignment tasks formed the backbone of Support for Relativity; when they are removed, the Support rate drops to only 11%. With them included, about a third (32%) of the data was classified as Support, relative to a No Support rate of 43% and a Mixed Support rate of 24%. If we consider the possibility that publication biases mean that fewer null results make it to publication, it may be the case that even this Support rate is an overestimate.

The review provides support for a number of important constraints on the Relativity hypothesis. For example, the rate of Support is higher when the Gender Content of a task is High rather than Low, suggesting any influence might be at least partly contingent on the opportunity to strategically call upon grammatical gender. Results are also more likely to be classified as Support when participants are processing animate rather than inanimate targets, which also suggests that language might be partly contingent on the immediacy of the overlap between grammatical gender and biological sex. This finding argues against a singular, uniform effect of gender category on all its members. Finally, results were more likely to

offer Support when the gendered language has two gender categories rather than three; a finding that is inconsistent with a straightforward account of grammatical gender classification per se influencing the conceptualisations of objects. The review initially appeared to reveal one misconception concerning Age; there was actually a higher rate of Support from samples of children than adults. Upon closer inspection, this outcome was closely bound to the fact that children performed those tasks that most consistently produced positive results for Relativity, namely Voice Choice and Sex Assignment. However, not all the predicted biases were supported. We found no evidence that Support was more common when participants performed a task in a gendered-language context, which is what is predicted by thinking-for-speaking accounts. The only parameter that could not be meaningfully assessed at all concerned the salience of otherwise of language in the task, an issue that we return to later in this Discussion.

What these parameter-related comparisons reveal is that much of the positive evidence for Relativity comes from tasks and conditions that are particularly susceptible to alternative, strategy-based explanations. Overall, we therefore take the results of this review as imposing quite powerful constraints on the Relativity hypothesis as seen through the lens of grammatical gender. Nevertheless, this conclusion itself comes with a caveat; we feel the review also points to a significant weakness in much of the relevant research's ability to speak to the issue of Relativity with clarity. This means that future research could either cement this rather negative conclusion, or *overturn* it through stronger designs that are less susceptible to confounds. As a result, we suggest that our review provides an *interim* rather than final pattern of results.

We focus our discussion on those areas that we feel need addressing in future work, and make suggestions as to some experiments to deal with them. First, we describe why we feel much of the data speak only weakly to the question of Relativity.

4.1. How do people solve the tasks?

Some have held that for Relativity to be supported there should be a reasonable expectation that participants did not engage grammatical gender information strategically. This is most clearly an option in tasks where judgments are about sex and are explicit, such as Voice Choice and Sex Assignment, but possibly also Similarity, Association and Object-Name memory Associations. It is therefore important to distinguish between two means of arriving at decisions in many of the tasks in this review; one as a result of conceptual change as usually hypothesised in Relativity accounts, and another through metalinguistic knowledge. Metalinguistic knowledge refers to the influence of the knowledge of a formal property of language, such as grammatical gender, on judgments about objects. While both processes are interesting, it is Relativity that research in this review was designed to investigate. The problem is that the Voice Choice and Sex Assignment tasks upon which the bulk of the Support for Relativity apparently rests cannot tell the two apart.

It might be argued that researchers can simply ask participants how they performed the task, and therefore be in a position to rule out a metalinguistic strategy as a result. Interestingly, the cases in which participants have been asked how they came to their decisions have thrown up mixed results (e.g., Almutrafi, 2015; Kurinski & Sera, 2011; Sato & Athanasopoulos, 2018). Particularly convincing evidence for a *conscious* metalinguistic strategy account of Voice Choice performance comes from a study in which 25 out of 30 participants later admitted to using grammatical gender to guide their responses (Almutrafi, 2015). However, although it has often been assumed to be so, there is also no a priori reason that the use of metalinguistic strategy need be a conscious process at all. Regardless of how it might occur, the use of metalinguistic knowledge undermines a reading of results as the outcome of conceptual change.

Some researchers have pointed out that participants rarely if ever respond in a manner that is 100% consistent with biological sex. This seems to argue against a metalinguistic strategy (conscious or otherwise) account. However, we cannot know if participants had more than one strategy, some more universal (such as masculine for artefacts, feminine for natural kinds: Mullen, 1990), and some more idiosyncratic or personal (see for example participants' justifications for their choices in Kurinski & Sera, 2011), with different strategies bringing to bear at different times throughout the task. The absence of a consistent, 100% effect of grammatical gender on task performance therefore does not preclude the possibility that metalinguistic strategies might account for some of the effects that were found.

Is there evidence to support one or other account from the results of the review? Here again we run into the logical problem that we cannot tell processes apart. For example, the finding that more support came from speakers of two-gendered languages than three-gendered languages might be seen as weakening metalinguistic strategy accounts, because there is no reason to believe Spanish speakers should prefer such strategies over German speakers, for example. However, the availability or attractiveness of metalinguistic strategies might on the other hand be enhanced where there is a neater one-to-one mapping of grammatical gender category and biological sex.

Another potential criticism of a metalinguistic account rather than Relativity account is that in taking the former view one subscribes to an intrinsically negative, prejudicial default that effectively renders Relativity empirically impossible to support. Essentially, giving a metalinguistic alternative explanation equal weight might set the bar for actual conceptual change accounts impossibly high. However, there exists the scope to tighten experimental design to guard against alternative, ‘killjoy’ explanations; this review suggests that, for the most part, when these measures are in place the likelihood of finding a positive result decline.

In our view, the most convincing evidence of the potential for metalinguistic strategy accounts comes from the results of Property Judgment tasks. Since this task type does not incorporate a sex/gender prompt it keeps such information at arm's length. The very low rate of only 3% Support from this task type might therefore reflect the absence of such strategies in performance.

Overall, the Voice Choice and Sex Assignment paradigms are the most susceptible to alternative explanations and further research using these tasks without serious modification is unlikely to reveal more about Relativity. Since these tasks are the most common in the field, and similar objections can be raised against other task types like Similarity, Object-Name Memory Associations and Associations, the pool of information from which we can make the most meaningful inferences about the research question is likely to be small. It is for this reason that we feel the case for grammatical gender influencing concepts is currently difficult to weigh up, and awaits future research (see also section 4.3 below).

4.2. Language on language or language on concepts?

The process of judging whether a task incorporates language is a difficult one. For example, in a Voice Choice or Sex Assignment task participants are sometimes only required to produce a single word: 'male', 'female', 'boy', etc. They might only need to circle a letter M or F on a sheet of paper. Does this constitute a language process that might inadvertently recruit grammatical gender itself? What of the role of the instructions of the task, which are linguistic, in formulating a linguistic process to arrive at a response? For almost all the tasks in the field, even language processing of the more conspicuous kind is unavoidable, such as when making judgments based on linguistic stimuli in the EAST, or thinking of adjectives to describe pictures in Properties judgments. For some, linguistic relativity research using solely behavioural measures (response times and related patterns) is *always* susceptible to linguistic

processes (e.g., Gleitman & Papafragou, 2013), and it is for this reason that some now advocate primarily neurophysiological approaches (Thierry, 2016).

The argument that language needs to be controlled in Relativity research is usually attributed to the thinking for speaking argument (Slobin, 1996), which was originally based on the idea that languages require speakers to attend to certain aspects of a scene, such as temporal and spatial details, depending on what information their language required (see also Slobin, 2003). Slobin later also conceived of 'thinking for comprehending', by which the languages we speak also influence the way that we think about what we *comprehend* (Slobin, 2003). Such a view could mean that tasks which present participants with words will also be subject to the restrictions of thinking for speaking, as might tasks that use words about gender or sex in their instructions. This would likely encompass almost all the tasks in this review.

In the tasks in this review, participants did not need to produce the actual nouns for the target items themselves in their response. There are some data from the review that we can bring to bear on this question, though it is not conclusive. The thinking for speaking theory predicts that effects of language on thought might not extend to performance outside of the language in which such effects are sourced. Translated into grammatical gender research, effects should therefore be strongest when performing in a gendered than in a non-gendered language context. This was not the case, though only by a very subtle margin (29% to 32%). However, given the fact that 83% of samples performed in a gendered language context, it is also possible that further data from research employing an ungendered context might lead to a change in this outcome, either in favour of or against thinking for speaking.

It is difficult to know where to draw the line between High and Low language content. We classified all but two papers (Phillips & Boroditsky, 2003; Sato & Athanasopoulos, 2018), involving only 133 samples in all, as Low in language content. It could therefore be argued that almost the entirety of the research in the review could be testing for an influence

of language on language. We ourselves do not make this claim; this is almost certainly too strong a conclusion to draw given the heterogeneity of task designs. In its broadest sense, the philosophical debate around the involvement of language in behaviour is beyond the remit of this review. More practically, however, we believe it difficult to argue that the tasks described in this review *vary* enough in their language content to allow meaningful comparisons along this dimension.

4.3. Practical suggestions for future research

We divide our suggestions for future research into two sections, in order of importance. Firstly, we point out that the results of the review support the possibility that there may be a fundamental flaw in the use of grammatical gender as a tool to speak to the question of Relativity at all. Secondly, we make the case that if grammatical gender *can* provide an insight, then future tasks would benefit from an overhaul in order to better control for alternative explanations.

4.3.1. Returning to the question of: what is 'gendered' about grammatical gender?

The results of the review make the case that effects of grammatical gender are for the most part predicted by parameter settings that would be consistent with a statistical association account, at least as well as by a conceptual change view of Relativity. This is because most settings that promote the association between biological sex and grammatical gender enhance the probability that effects are found.

That Relativity is scaffolded by associations *between* language and thought, rather than language *as* thought, is a view that is partially consistent with contemporary thought in the field, such as the label-feedback hypothesis (Lupyan, 2012) and its offshoot the structural feedback hypothesis (Sato & Athanasopoulos, 2018). These theories contend that labels or

grammatical information hone attention to associated features which in turn feed back down to lower-level processes in a feedback loop. These effects can be up-regulated or down-regulated by the salience of the relevant linguistic information in the task. The results of the review, as well as results from studies in which participants are briefly trained in invented or real languages and come to behave in line with those languages (e.g., Boroditsky, 2001; Casasanto, 2010; Phillips & Boroditsky, 2003), suggest that this is indeed the case. However, where a statistical association account departs from these accounts is that the latter allow for some degree of change at the conceptual level but a statistical association account does not. A statistical association account would predict that if an Italian speaker is processing the target 'bed' in the context of gender/sex, for example, then the concept of masculinity might receive activation by an association rather than by any lasting conceptual rub-off. This would be similar, for example, to the statistical association between the concepts of sunshine and ice-cream; we would be unlikely to conceive of sunshine as being *similar* to ice cream. Put simply, a statistical association account need not require conceptual change, especially *long-lived* change, to occur at all, and would therefore be incompatible with the spirit of Relativity in any of its theoretical incarnations.

As we stated in our Introduction, this review is not in a position to make such a distinction between Relativity and its alternatives, in part because more data is required, but also because our review did not find enough *unambiguous* support for Relativity to discriminate between possibilities. For example, it is difficult to assess positive results from Voice Choice and Sex Assignment in the light of the hierarchical taxonomy of Relativity accounts by Wolff and Holmes (2011), which ranges from the strongest form of Relativity (thought *is* language) through to subtler effects such as 'language as spotlight', when a non-Relativistic account is at least as likely an explanation for the results from such tasks. Instead, our review is in a position to weigh up the size of the problem in relating much grammatical

gender research to Relativity at all, in any of its forms. Of the five principal parameter settings that a statistical association account would predict, one (language content) was impossible to draw meaningful inferences from; three (target type, number of gender categories, salience of sex/gender) resulted in higher overall rates of Support, and only one (gendered language context) was equivocal. It is perhaps important to note that this latter parameter did not run powerfully in the *opposite* direction to what a statistical association account would predict; there was only a -3% Support rate difference, far smaller than the next smallest difference of +23%, which was instead in favour of the account. The sixth parameter—Age—was also equivocal, but less important for the account in any case. We therefore interpret these results as *framing and underlining the case* for a statistical association account, by which arbitrary labels for grammatical classes interact with arbitrary assignments of nouns to those classes, under conditions that facilitate their association. This is not to imply that we prefer such an account, or to rule out the possibility that multiple factors might have simultaneous and additive effects. It does imply, however, that grammatical gender is presently a foggier lens through which to inspect the case for Relativity than the domains of categorical perception, space or time, to name a few.

As a first step, it would be useful to establish whether 'masculine' and 'feminine' are psychologically privileged 'attractor' concepts that impose their status on other objects in their grammatical class, or whether these metalinguistic labels are themselves arbitrary. This is important because it would help researchers to understand whether the idea of a relationship between grammatical 'gender' and biological sex has any psychological reality. If it does, then it becomes more likely that members of a class are in some sense imbued with this conceptual relationship, and the case for conceptual change accounts would be enhanced.

There is already a study that suggests a method by which to test this. In one experiment not included in this review because it involved invented languages, native English

speakers were taught ‘Gumbuzi’, an artificial language with two artificial grammatical ‘gender’ groups labelled ‘soupative’ and ‘oosative’ (Phillips & Boroditsky, 2003). Participants were taught ten items in each group, six of which were inanimate objects, the remaining four humans who were either all female or all male. After learning which items were assigned to which category, participants rated the similarity of human-object pairs both within and across the two groups. The results found that pairs from the same group were rated as more similar than items from different groups, leading the authors to conclude that there can be a causative (i.e. *learned*) relationship between grammatical gender and people’s conceptualisations of objects.

Since 40% of all the items in a group were humans of the same biological sex, the groups were strongly biased towards sex/gender regardless of the labels ‘soupative’ and ‘oosative’. It is also likely that the participants were aware of such things as grammatical gender categories through formal second language instruction in schools, knowledge of which they may have applied to the task. Additionally, the labels ‘oosative’ and ‘soupative’ fail to actually describe *any* of the items within each group; real grammatical gender categories are at least partially correlated with the biological sex of its members. Nevertheless, this study provides a template for a future study that might teach one group of participants that groups are called masculine and feminine, and another group that the groups are named after another and *equally-represented* natural kind. If the arbitrary labelling of the classifiers themselves drives performance, then the results of similarity ratings should pattern in line with other labels at least as much as they would with masculine and feminine. This would suggest that any influence of grammatical gender is a human-made one that is independent of linguistic structure and lacking in psychological reality, undermining the notion of a conceptual relationship between grammatical gender and biological sex, and in turn favouring a statistical association relationship.

4.3.2. Dulling Occam's razor.

If grammatical gender is not merely a cultural label, then we follow Ramos and Roberson (2011) and others who suggest that studies be conducted that aim to restrict both gender and language to as oblique a role as possible. Property Judgment tasks do the former very well, but language remains fundamental, and in any case the evidence from these tasks is to date overwhelmingly negative. Instead, the Priming tasks by Sato and Athanasopoulos (2018) would seem strong candidates for future investigations. In their first experiment, French-English participants were found to be slower to indicate whether two objects were associated with a male or female face when the grammatical gender of those objects was incongruent with the biological sex of the person. In a second experiment participants matched one of two trait words (e.g., 'charming', 'realistic') to a now *genderless* face after being primed with a pair of objects, such as a tie and a spade. Results again pointed to an influence of the grammatical gender of the objects in French-English bilinguals' choices. These studies, while not eliminating biological sex and language altogether, keeps some distance between these and participants' actual responses because associations come from the task irrelevant grammatical gender of objects that participants are presented with earlier. Future work might find a way of making this distance greater still, and include direct statistical comparisons between speakers of a gendered language and speakers of a non-gendered language in order to establish more clearly that any effects are attributable to grammatical gender specifically[7].

[7] These studies also tested Monolingual English speakers but the pertinent analyses were within-group.

4.4. Limitations

This review represents, to the best of our knowledge, the first systematic attempt to assess the literature in a quantitative manner. The heterogeneity of methods and their uneven representation in the literature presented us with a difficult decision; to group together research of different types, or to provide a finer-grained picture. We took the view that it was better in a first review to provide a nuanced picture that takes into account differences between, for example, instructing participants to assign a voice to an object or a sex to an object. This does have the drawback of making it harder to make reliable inferences based on less well-used tasks, in particular Object-Name memory associations, Association tasks, and Similarity tasks. On the other hand, it makes it easier for later work to be incorporated into future reviews.

A more difficult and subjective issue concerns the interpretation of results classified as offering Mixed Support. The argument for the inclusion of this category is to our minds quite compelling. If we take for example the finding that training in Spanish improves the rate at which Voice Choices are consistent with grammatical gender but nevertheless fails to raise this rate above chance (Kurinski & Sera, 2011), neither an entirely cautious approach (i.e. this result finds no support for Relativity) nor an endorsement (Voice Choices are consistent with grammatical gender) naturally follow, and the need for a third category becomes clear. In order to present as objective a view as possible, we have for the most part focussed on the rate of Support in the first instance, and the rate of No Support second, and Mixed Support only where it is necessary, such as in conditions where there is little data either side of this middle category.

5. Conclusion

In conclusion, our review found that support for an influence of grammatical gender on concepts is strongly task- and context-dependent. Support also comes for the most part from tasks that are susceptible to clear alternative explanations. Perhaps most importantly, it needs to be empirically established that grammatical gender itself is not a cultural label but a concept with psychological reality before any influence can be reasonably attributed to truly *linguistic* processes.

Disclosure of interest

The authors report no conflicts of interest.

Open practices statement

Details of the review and the search mechanisms can be found in the supplemental online materials (SOM1 and SOM2).

References

Almutrafi, F. (2015). Language and cognition: effects of grammatical gender on the categorisation of objects. *Unpublished Doctoral Thesis*.

Athanasopoulos, P. (2006). Effects of the grammatical representation of number on cognition in bilinguals. *Bilingualism: Language and Cognition, 9*(1), 89-96.

Athanasopoulos, P. (2009). Cognitive representation of colour in bilinguals: The case of Greek blues. *Bilingualism: Language and Cognition, 12*(1), 83-95.

Athanasopoulos, P., & Boutonnet, B. (2016). Learning grammatical gender in a second language changes categorization of inanimate objects: Replications and new evidence from English learners of L2 French. In R. Alonso (Ed.), *Cross-linguistic influence in second language acquisition.* (pp. 173-192). Bristol, UK.: Multilingual Matters.

Athanasopoulos, P., & Bylund, E. (2013). Does grammatical aspect affect motion event cognition? A cross-linguistic comparison of English and Swedish speakers. *Cognitive Science, 37*(2), 286-309.

Bassetti, B. (2007). Bilingualism and thought: Grammatical gender and concepts of objects in Italian-German bilingual children. *International Journal of Bilingualism, 11*(3), 251-273.

Bassetti, B., & Nicoladis, E. (2016). Research on grammatical gender and thought in early and emergent bilinguals. *International Journal of Bilingualism, 20*(1), 3-16.

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

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*(2), 331-359.

Bender, A., Beller, S., & Klauer, K. C. (2011). Grammatical gender in German: A case for linguistic relativity? *The Quarterly Journal of Experimental Psychology, 64*(9), 1821-1835.

Bender, A., Beller, S., & Klauer, K. C. (2016a). Crossing grammar and biology for gender categorisations: Investigating the gender congruency effect in generic nouns for animates. *Journal of Cognitive Psychology, 28*(5), 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. *The Quarterly Journal of Experimental Psychology, 69*(1), 48-64.

Bender, A., Beller, S., & Klauer, K. C. (2018). Gender congruency from a neutral point of view: The roles of gender classes and conceptual connotations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(10), 1580-1608.

Bobb, S. C., & Mani, N. (2013). Categorizing with gender: Does implicit grammatical gender affect semantic processing in 24-month-old toddlers? *Journal of experimental child psychology, 115*(2), 297-308.

Boroditsky, L. (2001). Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive psychology, 43*(1), 1-22.

Boroditsky, L., & Schmidt, L. A. (2000). Sex, Syntax, and Semantics. *Proceedings of the Annual Meeting of the Cognitive Science Society, 22*(Retrieved from https://escholarship.org/uc/item/0jt9w8zf).

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

Brown, A., Lindsey, D. T., & Guckes, K. M. (2011). Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. *Journal of vision, 11*(12), 1-21.

Brown, R., & Lenneberg, E. (1954). A study in language and cognition. *The Journal of Abnormal and Social Psychology, 49*(3), 454-462.

Casasanto, D. (2010). Space for thinking. In V. Evans & P. Chilton (Eds.), *Language, cognition and space: The state of the art and new directions* (pp. 453-478). London: Equinox Publishing.

Casasanto, D., Boroditsky, L., Phillips, W., Greene, J., Goswami, S., Bocanegra-Thiel, S., . . . Gil, D. (2004). *How deep are effects of language on thought? Time estimation in speakers of English, Indonesian, Greek, and Spanish.* Paper presented at the Proceedings of the 26th Annual Conference of the Cognitive Science Society.

Cook, S. V. (2016). Gender matters: From L1 grammar to L2 semantics. *Bilingualism: Language and Cognition*, 1-19.

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

Costa, A., Kovacic, D., Fedorenko, E., & Caramazza, A. (2003). The gender congruency effect and the selection of freestanding and bound morphemes: evidence from croatian. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1270-1282.

Costa, A., Kovacic, D., Franck, J., & Caramazza, A. (2003). On the autonomy of the grammatical gender systems of the two languages of a bilingual. *Bilingualism: Language and Cognition, 6*(3), 181-200.

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*(2), 449-460.

De Bruin, A., Treccani, B., & Della Sala, S. (2015). Cognitive advantage in bilingualism: An example of publication bias? *Psychological Science, 26*(1), 99-107.

Degani, T. (2007). The semantic role of gender: Grammatical and biological gender match effects in English and Spanish. *Unpublished Masters Thesis*.

Dilkina, K., McClelland, J. L., & Boroditsky, L. (2007). *How language affects thought in a connectionist model.* Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychological Science, 24*(5), 613-621.

Eberhard, K. M., Heilman, M., & Scheutz, M. (2005). *An empirical and computational test of linguistic relativity.* Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and brain sciences, 39*, E229.

Flaherty, M. (2001). How a language gender system creeps into perception. *Journal of Cross-Cultural Psychology, 32*(1), 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.* Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

Franklin, A., Clifford, A., Williamson, E., & Davies, I. (2005). Color term knowledge does not affect categorical perception of color in toddlers. *Journal of experimental child psychology, 90*(2), 114-141.

Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences, 103*(2), 489-494.

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 and language, 105*(2), 91-98.

Gleitman, L., & Papafragou, A. (2013). Relations Between Language and Thought. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology* (pp. 504-523). New York: Oxford University Press.

Haertlé, I. (2017). Does Grammatical Gender Influence Perception? A Study of Polish and French Speakers. *Psychology of Language and Communication, 21*(1), 386-407.

Imai, M., & Gentner, D. (1997). A cross-linguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition, 62*(2), 169-200.

Imai, M., Schalk, L., Saalbach, H., & Okada, H. (2014). All giraffes have female-specific properties: Influence of grammatical gender on deductive reasoning about sex-specific properties in German speakers. *Cognitive Science, 38*(3), 514-536.

Kaushanskaya, M., & Smith, S. (2016). Do grammatical–gender distinctions learned in the second language influence native-language lexical processing? *International Journal of Bilingualism, 20*(1), 30-39.

Konishi, T. (1993). The semantics of grammatical gender: A cross-cultural study. *Journal of Psycholinguistic Research, 22*(5), 519-534.

Konishi, T. (1994). The connotations of gender: A semantic differential study of German and Spanish. *Word, 45*(3), 317-327.

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*(4), 843-858.

Kurinski, E., Jambor, E., & Sera, M. D. (2016). Spanish grammatical gender: Its effects on categorization in native Hungarian speakers. *International Journal of Bilingualism, 20*(1), 76-93.

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

Lambelet, A. (2016). Second grammatical gender system and grammatical gender-linked connotations in adult emergent bilinguals with French as a second language. *International Journal of Bilingualism, 20*(1), 62-75.

Landor, R. (2014). *Grammatical Categories and Cognition across Five Languages: The Case of Grammatical Gender and its Potential Effects on the Conceptualisation of Objects.* Thesis (PhD Doctorate).–Griffith University, Brisbane.

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

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis. *Frontiers in psychology, 3*, 54.

Martinez, I. M., & Shatz, M. (1996). Linguistic influences on categorization in preschool children: A crosslinguistic study. *Journal of Child Language, 23*(3), 529-545.

Mickan, A., Schiefke, M., & Stefanowitsch, A. (2014). Key is a llave is a Schlussel: A failure to replicate an experiment from Boroditsky et al 2003. *Yearbook of the German Cognitive Linguistics Association, 2*(1), 39-50.

Montefinese, M., Ambrosini, E., & Roivainen, E. (2018). No grammatical gender effect on affective ratings: evidence from Italian and German languages. *Cognition and Emotion*, 1-7.

Mullen, M. K. (1990). Children's classifications of nature and artifact pictures into female and male categories. *Sex roles, 23*(9-10), 577-587.

Nicoladis, E. (unpublished). Unpublished data.

Nicoladis, E., Da Costa, N., & Foursha-Stevenson, C. (2016). Discourse relativity in Russian-English bilingual preschoolers' classification of objects by gender. *International Journal of Bilingualism, 20*(1), 17-29.

Nicoladis, E., & Foursha-Stevenson, C. (2012). Language and culture effects on gender classification of objects. *Journal of Cross-Cultural Psychology, 43*(7), 1095-1109.

Park, H. I., & Ziegler, N. (2014). Cognitive shift in the bilingual mind: Spatial concepts in Korean–English bilinguals. *Bilingualism: Language and Cognition, 17*(2), 410-430.

Pavlidou, T.-S., & Alvanoudi, A. (2013). Grammatical gender and cognition. In N. Lavidas, T. Alexio, & A. Sougari (Eds.), *Major Trends in Theoretical and Applied Linguistics 2* (Vol. 2, pp. 109-124). London: Versita.

Pavlidou, T.-S., & Alvanoudi, A. (2018). Conceptualizing the world as 'female'or 'male': Further remarks on grammatical gender and speakers' cognition. In N. Topintzi, N. Lavidas, & M. Moumtzi (Eds.), *Selected Papers on Theoretical and Applied Linguistics from ISTAL23*. Thessaloniki: School of English, Aristotle University of Thessaloniki.

Phillips, W., & Boroditsky, L. (2003). *Can quirks of grammar affect the way you think? Grammatical gender and object concepts.* Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

Pinker, S. (1994). The language instinct. New York, NY, US: William Morrow & Co.

Pylyshyn, Z. W. (1984). *Computation and cognition*: MIT press Cambridge, MA.

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

Roberson, D., Pak, H., & Hanley, J. R. (2008). Categorical perception of colour in the left and right visual field is verbally mediated: Evidence from Korean. *Cognition, 107*(2), 752-762.

Saalbach, H., Imai, M., & Schalk, L. (2012). Grammatical gender and inferences about biological properties in German-speaking children. *Cognitive Science, 36*(7), 1251-1267.

Samuel, S., Roehr-Brackin, K., & Roberson, D. (2016). 'She says, he says': Does the sex of an instructor interact with the grammatical gender of targets in a perspective-taking task? *International Journal of Bilingualism, 20*(1), 40-61.

Samuel, S., Roehr-Brackin, K., Pak, H., & Kim, H. (2018). Cultural Effects Rather Than a Bilingual Advantage in Cognition: A Review and an Empirical Study. *Cognitive Science, 42*(7), 2313-2341.

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

Sedlmeier, P., Tipandjan, A., & Jänchen, A. (2016). How Persistent are Grammatical Gender Effects? The Case of German and Tamil. *Journal of Psycholinguistic Research, 45*(2), 317-336.

Segel, E., & Boroditsky, L. (2011). Grammar in art. *Frontiers in psychology, 1*, 244.

Semenuks, A., Phillips, W., Dalca, I., Kim, C., & Boroditsky, L. (2017). *Effects of Grammatical Gender on Object Description.* Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

Sera, M. D., Berge, C. A., & del Castillo Pintado, J. (1994). Grammatical and conceptual forces in the attribution of gender by English and Spanish speakers. *Cognitive development, 9*(3), 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*(3), 377-397.

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 study of language and thought* (pp. 157-192). Cambridge: MIT Press.

Thierry, G. (2016). Neurolinguistic relativity: how language flexes human perception and cognition. *Language Learning, 66*(3), 690-713.

Vernich, L. (2017). Does learning a foreign language affect object categorization in native speakers of a language with grammatical gender? The case of Lithuanian speakers learning three languages with different types of gender systems (Italian, Russian and German). *International Journal of Bilingualism, 23*(2), 417-436.

Vernich, L., Argus, R., & Kamandulytė-Merfeldienė, L. (2017). Extending research on the influence of grammatical gender on object classification: a cross-linguistic study comparing Estonian, Italian and Lithuanian native speakers. *Eesti Rakenduslingvistika Ühingu aastaraamat, 13*, 223-240.

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

Vuksanovic, J., Bjekic, J., & Radivojevic, N. (2015). Grammatical Gender and Mental Representation of Object: The Case of Musical Instruments. *Journal of Psycholinguistic Research, 44*(4), 383-397.

Whorf, B. L. (1956). Language, Thought and Reality. Selected Writing: of Benjamín Lee Whorf. *Cambridge, MA.: The MIT Press*.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences, 104*(19), 7780-7785.

Witzel, C., & Gegenfurtner, K. R. (2013). Categorical sensitivity to color differences. *Journal of vision, 13*(7), 1-33.

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

Yorkston, E., & De Mello, G. E. (2005). Linguistic gender marking and categorization. *Journal of Consumer Research, 32*(2), 224-234.