Review of Philosophy and Psychology (2026) 17:275–299
https://doi.org/10.1007/s13164-025-00794-9

Who's the Surgeon? Cognitive Implications of Gendered Language

Mara Floris[1] · **Federico Cella**[2] · **Camilla Borgna**[3,4]

Received: 14 April 2025 / Accepted: 3 November 2025 / Published online: 7 March 2026

Abstract

Psycholinguistic evidence shows that grammatical gender can shape mental representations, influencing how individuals interpret ambiguous scenarios. This study investigates whether varying the linguistic formulation of the "specialist riddle" affects the ability to recognize the surgeon as a woman. We tested four versions of the riddle on Italian high-school students ($n = 691$), manipulating the term used for "surgeon" to compare the masculine form, a control condition, a gender-neutral periphrasis, and an innovative gender-neutral form. Participants exposed to the masculine generic form perceived the riddle as more difficult and were markedly less likely to identify the surgeon as a woman compared to those exposed to the gender-neutral formulations, especially the innovative one. These findings illustrate how grammatical gender can influence mental representations and highlight the potential of alternative linguistic forms to mitigate the effects of the generic masculine or the "male as default" bias.

✉ Camilla Borgna
camilla.borgna@unito.it

Mara Floris
floris.mara@unisr.it

Federico Cella
federicocella@ub.edu

[1] Facoltà Di Filosofia, Università Vita-Salute San Raffaele, Via Olgettina, 58, 20132 Milan, Italy

[2] LOGOS, Universitat de Barcelona, Carrer de Montalegre, 6-8 08022 Barcelona, Spain

[3] Dipartimento Di Culture Politica E Società, Università Degli Studi Di Torino, Via Verdi, 8 - 10124 Turin, Italy

[4] Collegio Carlo Alberto, Piazza Arbarello 8, Turin, Italy

1 Introduction

In recent years, scholarly discussions have increasingly examined the cognitive effects of grammatical gender. Particular attention has been given to generic masculine terms in gendered languages—i.e., grammatically masculine terms used to refer to persons of any gender. Extensive empirical evidence from psycholinguistics shows that grammatically masculine forms, even when intended as generic, disproportionately activate male referents, leading to a male bias in mental representations[1] (Gabriel et al., 2018; Gygax et al. 2008, 2012, 2021; Stahlberg et al., 2007). A central issue concerns gender-fair language (GFL)—i.e., strategies to avoid the use of generic masculine terms—and professional titles (e.g., Sczesny et al. 2016; Vergoossen et al. 2020).

In this work, we analyze the relationship between the interpretation of the generic masculine, gender stereotypes, and gender-fair language in Italian, a language in which there are still few studies (cfr. Gygax et al. 2021; Sczesny et al. 2016). In the present study, we focus on the so-called "specialist riddle":

> "A father and his son have a car accident. The father dies. The son is rushed to the emergency room. The attending surgeon looks at the boy and says: 'I cannot operate. He is my son!'. How is this possible?".

This well-known riddle has captured popular culture's attention since at least the 1970s. There are multiple solutions to the riddle, but the most statistically plausible hypothesis—i.e., that the surgeon is the mother—is particularly hard to conceive.

The significance of the riddle extends beyond casual discussions, as it found its way into the academic literature when Moulton (1978) utilized it to highlight the complexities of seemingly gender-neutral terms in English, such as "surgeon". Decades later, the riddle became a matter of experimental investigation. Recent findings show that few participants answer that the surgeon is the mother (~30%) (e.g., Belle et al. 2021; Kollmayer et al. 2018). People often make diverse guesses, as reported in the literature (Morehouse et al. 2022; Belle et al. 2021): the surgeon could be the other parent in a homo-parental couple, or the boy could have been adopted; there is no shortage of solutions that appeal to the supernatural and complex fantasy stories. Yet the somewhat simpler hypothesis seems to be precluded: it is difficult to think of the mother of the boy as the surgeon, even though the vast majority of parents are heterosexual. One straightforward explanation is the presence of a strong stereotype associating the medical profession, and in particular surgery, with mas-

[1] The notion of *mental representation* is the subject of a complex and ongoing debate in philosophy of mind and cognitive science (see Pitt 2020; Fodor 1987; Dretske 1995). While some theories emphasize the computational and syntactic aspects of mental representations (Fodor 1975; Pylyshyn 1984), others highlight their phenomenological or intentional content (Block 1996; Chalmers 1996; Kriegel, 2013). In this article, we do not take a stance on these theoretical controversies. Rather, we use the term *mental representation* in a minimal and empirically grounded sense, to refer to the internal cognitive effects that correlate with linguistic input—specifically, the activation of gendered concepts in response to gendered language forms. Our use of the term is thus purely functional and intended to justify the interpretation of linguistic variation as producing distinct behavioral outcomes, in line with experimental findings from psycholinguistics.

culinity. Despite the increasing number of women in surgery (e.g., Linscheid et al. 2020) and gender-balanced representations in popular media (e.g., TV series such as *Grey's Anatomy*, *ER*, and *The Good Doctor*), this stereotype persists. An additional, not mutually exclusive, explanation is the presence of a gender bias that goes beyond occupational stereotypes. The so-called "male as default bias" is the empirically documented tendency to automatically assume that an unknown person's gender is male (e.g., Bailey et al., 2020; Bem 1993; Hamilton 1991; Lindsay 2023).[2]

To date, the specialist riddle has primarily been tested in English and with adult populations. We tested the classical version of the riddle, along with several variants using gender-fair language (GFL), on Italian-speaking adolescents ($n = 691$, age 14–19). Testing the original riddle with Italian speakers is crucial because, while personal nouns in English tend to be gender-neutral (with some exceptions for compound words like "housewife", "firemen", "policeman", etc.) Italian nouns are gendered, whether they refer to animate or inanimate entities. The Italian masculine term for "surgeon" is "chirurgo," which can be used generically—i.e., it can refer to surgeons of any gender. In contrast, the feminine form "chirurga" refers to female surgeons exclusively. Given this asymmetry, the use of the masculine generic term "chirurgo" may make the mental representation of the surgeon as a woman less accessible, thus reducing the likelihood of providing the "mother" answer. Furthermore, while previous research has mainly focused on the socially-selected group of university students, our sample is more socially representative (drawing on secondary schools, both academically and vocationally oriented). Although not fully representative of the general population of adolescents, this ensures a greater external validity for our findings.

Our study builds on this literature by investigating how linguistic form influences the accessibility of gender representations in Italian, a language that habitually employs the masculine form to refer to persons of unspecified or generic gender. To our knowledge, no previous research has directly examined whether using GFL in the formulation of the specialist riddle can modulate the *male-as-default* bias in such linguistic contexts. We focus on Italian because it represents a particularly revealing case: while the masculine generic ("chirurgo") is grammatically acceptable and frequently used, it often fails to be interpreted as genuinely neutral. The feminine form ("chirurga"), by contrast, is grammatically possible but still uncommon in everyday discourse, especially for high-prestige professions. This asymmetry makes Italian an ideal testing ground for understanding how grammatical and social conventions jointly shape gendered mental representations.

In what follows (see Sect. 5), we present a preregistered experiment designed to test whether different GFL strategies—both conservative and innovative—affect participants' interpretation of the riddle, as well as their response times and perceived task difficulty. By including a large sample of adolescents from diverse educational backgrounds, our study also extends prior work conducted primarily with university students, thereby improving the external validity and sociolinguistic relevance of the findings. In the next sections, we provide an overview of previous research on the specialist riddle, how participants' background and linguistic factors affected the

[2] See below in the text, Sect. 1.5.

likelihood of providing the "mother" answers, and the problem of names of professions in grammatically gender languages.

1.1 The Specialist Riddle in the Experimental Literature

In the last two decades, the riddle has been a matter of empirical investigation, providing support for the intuition about the difficulty of envisioning the surgeon as the mother. Most studies were conducted on convenience samples of university students. In a series of experiments by Stoeger et al. (2004), the riddle was tested with a group of university students from different majors. Only 46 out of 144 (32.2%) participants answered that the surgeon could be the mother. With reversed genders (the mother dies, so the surgeon is likely the father), 80% of the participants solve the riddle. Reynolds et al. (2006) obtained similar results, though their study involved smaller samples as they tested 10 different versions of the riddle: of the total sample of 201 participants, only 21 were presented with the traditional version of the riddle. More recently, Belle et al. (2021) presented the riddle to a group of psychology students ($N = 297$) and found that 30% of the respondents identified the mother as a possible solution. Interestingly, they also found that 36% of the respondents answered that the boy could be the child of a same-sex couple, with the surgeon being the other father, despite the fact that same-sex couples are a minority. Morehouse et al. (2022) tested the riddle with a nationally representative sample of the general population from the US ($n = 3010$); in this less socially selected sample, only 18% of the respondents could provide the "mother" answer.

In all the above-mentioned studies, the riddle was tested in English, a language in which nouns typically lack gender. To the best of our knowledge, the only experiment testing the riddle in a language with gendered nouns was conducted by Kollmayer et al. (2018). In their study, the riddle was presented to a group of 517 native German speakers recruited online. Participants were randomly assigned to one of two conditions, with the treatment variable being the language used in a priming text they read before attempting the riddle: either standard German or a version adopting gender-fair language (GFL). The findings showed that individuals exposed to the GFL priming text were more likely to identify the specialist as the mother compared to those who read the text using masculine generics (44% vs. 33%).

A similar result was observed in English by Morehouse et al. (2022). In their first study on the classic riddle, which included 606 participants (mean age 37.83, 53% female, 76% White), over 70% of participants failed to identify the mother as the surgeon. In a second study, the term "son" was replaced with the gender-neutral kinship term "child," resulting in a substantial increase in the percentage of participants providing the "mother" solution. This finding highlights that even incidental exposure to gender-fair language can help reduce stereotypical associations. Interestingly, the gender-neutral term "child" was notably more effective at mitigating gender bias than the explicitly female-gendered term "daughter". The authors hypothesize that incidental exposure to a gender-neutral term is sufficient to disrupt gender biases.

2 Linguistic Aspects of the Specialist Riddle

Linguistic aspects in the formulation of the riddle may affect respondents' answers. As mentioned, the studies by Kollmeyer et al. (2018) and Morehouse et al. (2022) indicate that indirectly priming respondents with GFL written texts may facilitate the readiness to provide the most statistically plausible solution, i.e., that the surgeon is the mother. However, to our knowledge, no previous study has directly examined whether using gender-fair language (GFL) in the formulation of the riddle influences respondents' solutions in languages that conventionally employ the masculine form to refer to individuals of unspecified or generic gender.

2.1 Grammatical Gender Languages

English is a language where gender is typically not expressed in nouns: only pronouns are gendered, with a few exceptions among nouns like compound profession names, such as "fireman," "midwife," and "policeman," or kinship terms like "son/daughter" or "grandmother/grandfather." In contrast, many other European languages are grammatical gendered languages where nearly all nouns are gendered. In particular, Romance languages such as French, Spanish, Portuguese, and Italian operate on a binary gender system with feminine and masculine genders. In languages where nouns are gendered, the assignment is not always tied to the gender of the referent. This is evident not only with inanimate objects, which naturally lack gender, but also with many animal names, where the gender mark is independent of the actual gender of the animal. Regarding names referring to the roles of individuals, Italian names are typically conjugated based on the gender of the referent. For instance, "amica" denotes a female friend, while "amico" refers to a male friend. However, professional titles are a notable exception to this rule, as they are frequently used in the masculine form, particularly for prestigious professions—a phenomenon that has been the subject of significant debate (see Menegatti & Rubini 2017; Nitti 2021; Sulis & Gheno 2022).

In Italian, traditionally masculine and esteemed professional titles (e.g., doctor, lawyer, president) are often employed in the masculine form, even when the referent is a woman, especially for prestigious professions. This practice introduces a linguistic asymmetry: the masculine singular form can refer to a man, a woman, or an individual of unspecified gender, while the feminine form is exclusively used to denote a woman. This phenomenon, known as the 'generic masculine,' plays a significant role in shaping perceptions of gender within language. Furthermore, when applied to women, the feminine form of these titles is sometimes perceived as diminutive or less prestigious, prompting many women to prefer being addressed using the generic masculine (cf. Gheno, 2022; Pulijana & Stevens 2023). This asymmetry is also mirrored in the phenomenon of the so-called 'over-extended masculine,' wherein the masculine plural form is used to refer to mixed-gender groups. In this usage, the masculine can denote either a group composed entirely of men or a mixed-gender group, while the feminine plural is strictly reserved for groups consisting exclusively of women. Notably, the inclusion of even a single man in a predominantly female group automatically shifts the term to the masculine plural—a linguistic convention

not reciprocated in reverse (see Formato 2019). This structural bias reinforces male-centered language norms and diminishes the linguistic visibility of women in both individual and collective references.

2.2 Cognitive Effects of Grammatical Gendered Languages

Empirical evidence from psycholinguistics shows that grammatically masculine forms, even when intended as generic, disproportionately activate male referents, leading to a male bias in mental representations (Gabriel et al. 2018; Stahlberg et al., 2007). Studies show that after encountering generic masculine forms, participants are faster and more accurate when responding to male exemplars than female ones (e.g., Gabriel et al. 2008; Garnham & Yakovlev, 2015; Gygax et al. 2008, 2012). Similarly, when asked to estimate the proportion of men versus women in a group, participants consistently overestimate the presence of men (Hansen et al., 2016). These findings align with broader evidence indicating that masculine forms are more easily linked to male referents, regardless of the specific gender stereotypes associated with individual nouns (Gabriel & Mellenberger, 2004; Irmen & Kurovskaja 2010; Schärer 2008). Further research underscores the cognitive mechanisms that contribute to this male bias. Languages with grammatical gender categories reinforce associations between masculine forms and male referents, even when the intended meaning is gender-neutral (Gygax et al. 2021). For instance, when job titles are presented using generic masculine forms, both women (Hentschel et al. 2018; Stout & Dasgupta 2011) and children (Vervecken & Hannover 2015) report lower self-assigned fit to and interest in the same positions. These biases manifest in various decision-making tasks, including naming exemplars from occupational groups (Stahlberg et al., 2001), selecting pictures (Bailey & LaFrance, 2017), and determining first names (Vervecken et al., 2013). Across several languages—such as English, French, German, Norwegian, Russian, and Spanish—the use of generic masculine forms leads to a systematic male bias, hindering gender-neutral interpretations (Schneider & Hacker 1973; Hamilton, 1988).

The reason why this phenomenon is observed is still a matter of debate; we identified in the literature at least two theories, possibly concurrent:

(1) Grammatical gender influences the semantic representations of their referents, even when the referents have no inherent gender. This phenomenon is exemplified by studies conducted by Vigliocco and her research team over recent decades (see Vigliocco 1999; Vigliocco et al. 2005; Kousta et al. 2008). For instance, Vigliocco et al. (2005) investigated how grammatical gender influences meaning in Italian and German. The researchers conducted four experiments and found that in Italian, grammatical gender affects cognition for animals but not for artifacts, likely due to the association between noun gender and the biological sex of referents. This effect was not observed in German, where the correlation between grammatical gender and biological sex is less transparent. Their findings suggest that grammatical gender effects are more pronounced in languages with a clear correlation between noun gender and referent gender.

(2) As claimed by Gygax et al. (2021), the semantic ambiguity of grammatically masculine forms often encompasses multiple meanings, creating cognitive challenges for readers and speakers as they interpret language and construct mental representations of gender. This ambiguity frequently results in biased interpretations, with masculine forms predominantly being understood as male, even when they are intended to be gender-neutral or inclusive. The cognitive mechanisms underpinning these biases suggest that languages with grammatical gender categories inherently promote associations between grammatical gender and gender stereotypes, further reinforcing androcentric interpretations in contexts where gender is not explicitly specified.

2.3 Gender Roles and Stereotypes

Even in languages without grammatical gender, it may prove hard for people to evoke mental representations of women, as demonstrated by the literature on the "surgeon riddle", which is predominantly in English, i.e., a language where nouns generally do not carry a grammatical gender. Why is that the case? A common challenge in psycholinguistic studies on the reading of the generic masculine is the difficulty in experimentally separating the effects of the generic use of an expression from the stereotypes associated with gendered roles, such as professions. As measured by Misersky (2014), professions are strongly associated with gender stereotypes. In other words, professions that are stereotypically associated with masculinity tend to evoke male representations, and the reverse is true for professions associated with femininity, irrespective of the grammatical gender of the nouns and pronouns used, and even in the presence of GFL.

The experiments in Reynolds et al.'s (2006) study illustrate how stereotypes influence comprehension in riddle-like scenarios involving a range of professions beyond surgery. These experiments systematically tested variations of the classic "surgeon riddle," replacing "surgeon" with other stereotyped or gender-marked professions such as "nurse," "fireman," or "policewoman," as well as introducing neutral or less stereotypically gendered roles like "pediatrician." The results consistently revealed that readers automatically activate stereotypical gender associations tied to these professions, leading to entrenched mental representations that are resistant to contextual contradiction. For instance, when encountering "nurse," participants were more likely to assign female gender, while "fireman" or "policewoman" reinforced male and female expectations, respectively, causing difficulty in reconciling scenarios that violated these stereotypes. Notably, even in passages featuring a gender-neutral term like "pediatrician," readers often inferred male gender, demonstrating the pervasive influence of implicit biases. Attempts to mitigate these effects, such as introducing stereotype-breaking characters (e.g., a female porter), shifting the context to a non-clinical setting, or varying the gender of the parent–child pair in the narrative, failed to significantly reduce comprehension difficulties. These findings underscore that gender inferences are activated as soon as a profession is mentioned, becoming entrenched in readers' mental models and hindering their ability to integrate conflicting information later in the text.

2.4 GFL in Italian

In Italian and other languages with gendered nouns, the challenge of GFL is twofold. Firstly, the binary nature of the Italian gender system leads to consistent misgendering of individuals who do not identify with the traditional male–female dichotomy. Secondly, the use of generic masculine terms, such as masculine profession names for individuals of any gender, and the overextended masculine, referring to mixed-gender groups using the masculine plural, tends to evoke predominantly masculine mental representations (see Sect. 1.3.2). This limits the visibility of female and non-binary individuals.

In order to overcome these limits, in Italian, it is possible to adopt different GFL strategies. Following Rosola et al. (2023), we utilize a framework based on two orthogonal binary distinctions:

- *Visibility Strategies* employ distinct grammatical forms to explicitly highlight the gender(s) of the referent(s) or, if the gender is unknown, its possible options (e.g., feminine, masculine, non-binary). They are contrasted with *Obscuration Strategies*, which neutralize references to any gender by avoiding gender-specific expressions.

- *Conservative Strategies* adhere to the traditional grammatical structures of standard Italian. In contrast, *Innovative Strategies* introduce new linguistic forms and expressions into the language.

This framework results in a four-category classification, as outlined in Table 1.

In our experiment, we manipulated the word used for "surgeon" in Italian to examine its effect on participants' responses to the specialist riddle. Given that making the surgeon's female gender explicit would likely compromise the integrity of the riddle, and to ensure that the text remained as natural and fluent as possible, we opted for two obscuration strategies—one conservative and one innovative. We deliberately avoided using gender-splitting strategies (e.g., *chirurgo/chirurga*), which are common in written discourse but would have made the riddle highly unnatural and stylistically awkward in spoken Italian. This choice allowed us to preserve the pragmatic and linguistic plausibility of the riddle while still testing the effects of different gender-fair formulations. The conservative strategy uses the gender-neutral expression "persona che opera," (i.e., person who performs surgery), which is standard Ital-

Table 1 Based on Rosola et al. (2023), GFL strategies can be categorized along two axes: visibility vs. obscuration and conservative vs. innovative forms

	Visibility	**Obscuration**
Conservative	The female vs. male identity of the referent is made explicit with feminine vs. masculine grammatical forms E.g.: *chirurga o chirurgo; chirurgo/a*	Periphrases that circumvent gendered-specific expressions E.g.: *persona che opera; chi opera*
Innovative	The female, male, or non-binary identity of the referent is made explicit with linguistic innovations E.g.: *chirurga, chirurgo, o chirurg*; chirurga, chirurgo,* o *chirurgə*	Linguistic innovations that replace gendered-specific expressions E.g.: *chirurg*; chirurgə*

ian. The innovative strategy employs "chirurg*" (i.e., root of the word for surgeon, followed by *), a non-standard but increasingly common form in written documents, designed to avoid specifying the gender of the referent.[3]

3 Present Study

We conduct a preregistered[4] study to examine the difficulty of providing the "mother" solution to the specialist riddle in Italian and to test whether the use of gender-fair language (GFL) strategies can facilitate this solution. Our sample of participants comprises Italian-speaking adolescents in several North-Western Italian provinces ($N = 691$). Our choice to recruit teenagers was both pragmatic and theoretically motivated. On the practical side, the study relied on a convenience sample in schools where we were already conducting unrelated experiments. On the theoretical side, adolescence is widely regarded as a crucial period in which value-related (including gender) attitudes are formed and crystallized (Alwin & Krosnick 1991; Hendriks et al. 2024), making this population especially informative for investigating the development and malleability of gender-related biases. In addition, younger cohorts tend to endorse more egalitarian gender values on average, reflecting both compositional (cohort replacement) and period forces (Shu & Meagher 2018). Although not fully representative of the population of Italian adolescents, our sample is less socially selective than those of previous studies relying on university students (e.g., Stoeger et al. 2004; Belle et al. 2021), as it involves students attending high school with generalist, technical, or vocational curricula. The study was approved by the Ethical Committee of the University of Turin (Protocol No. 0183640 dated 15/03/2023—[UOR: S1000045—Classif. III/14]).

We developed four distinct variations of the specialist riddle, each constituting an experimental condition.

(1) In the '*Classical riddle' condition*, participants read an Italian translation of the classical riddle: the father of the boy dies in the car accident, and we referred to the surgeon who enters the operating room using the Italian generic masculine term "chirurgo". The most plausible solution to the riddle is that the surgeon is the mother.
(2) In the '*Parent-role reversal' condition*, participants read a version of the riddle where the gendered roles are reversed: the mother of the boy dies, the surgeon is still referred to as "chirurgo", but here the most plausible solution is that the surgeon is the father. The 'Parent-role reversal' condition served as a control condition to measure the baseline 'male bias' in imagining the surgeon when the specialist is described using a generic masculine term. Inclusion of this control condition adheres to established practices in the literature (Stoeger et al. 2004;

[3] We chose the innovative strategy based on the star (*) rather than the one based on the schwa (ə), because the latter is more recent and less common in the Italian language.

[4] We preregistered our research design and hypotheses before starting the data collection. See: https://aspredicted.org/K4K_CY7

Belle et al. 2021), enabling an assessment of the riddle's complexity when traditional gender roles are inverted.

(3) In the '*Conservative GFL' condition*, participants read a version of the riddle where the father dies, the most plausible answer is that the surgeon is the mother, and the surgeon is described using an obscuring and conservative GFL strategy, namely the gender-neutral periphrasis "persona che opera" (Person who operates). Despite featuring a feminine grammatical gender, "persona" is an epicene term, signifying a word that does not convey information about the gender of the referent.

(4) Finally, in the *'Innovative GFL' condition*, participants read a version of the riddle where the father dies, the most plausible answer is that the surgeon is the mother, and the surgeon is described using an obscuring but innovative GFL expression, namely "Chirurg*." Although this spelling is not (yet) accepted in standard Italian grammar, its usage in written texts has grown in the last decades.

The full text of the riddles tested in the four conditions is available in Annex A of the Supplementary Materials.

Our dependent variables were: (1) answer to an open-ended question about the identity of the surgeon (see below for details on the coding procedure); (2) time spent on the riddle-solving task; (3) self-perceived difficulty of the task (1–5 Likert scale). We measured the male bias in imagining the surgeon by comparing each dependent variable in the 'Classical riddle' condition with a baseline given by the version of the riddle in the 'Parent-role reversal' condition. We tested whether obscuration GFL strategies decreased the male bias across the three dependent variables by comparing the effect of the 'Classical riddle' condition with those of the 'Conservative GFL' condition and the 'Innovative GFL' condition.

We hypothesized that:

(H1) There is a male bias in solving the riddle across the three dependent variables: participants in the 'Classical riddle' condition are less likely to indicate that the surgeon is the opposite-sex parent, spend more time on the riddle, and perceive it as more difficult than participants in the 'Parent-role reversal' condition;

(H1 bis) The male bias in solving the riddle across the three dependent variables is stronger for adolescents who identify as males compared to those who identify as females;

(H2) Conservative obscuration GFL decreases the male bias: participants in the 'Conservative GFL' condition are more likely to provide the 'mother' solution, spend less time on the riddle, and perceive it as less difficult than participants in the 'Classical riddle' condition;

(H3) Innovative obscuration GFL decreases the male bias: participants in the 'Innovative GFL' condition are more likely to provide the 'mother' solution, spend less time on the riddle, and perceive it as less difficult than participants in the 'Classical riddle' condition.

Concerning the relative effectiveness of the conservative vs. innovative strategies, we had two alternative hypotheses. On the one hand, since the usage of the high-star is relatively uncommon, it may increase the difficulty of the text, and therefore it may pose additional challenges in answering the riddle:

(H4a): Innovative obscuration GFL is less effective than conservative obscuration GFL in decreasing the male bias.

On the other hand, the usage of the star may alert readers that gender identity matters in solving the riddle:

(H4b): Innovative obscuration GFL is more effective than conservative obscuration GFL in decreasing the male bias.

Besides manipulating the linguistic formulation of the riddle, we also explored whether participants' characteristics and previous experiences may predict their responses. Demographic factors, such as age, gender, and type of school attended (academic, technical, or vocational), were also measured to examine their potential impact on the participants' likelihood of providing the correct solution. Additional variables included familiarity with GFL forms (e.g., prior exposure to terms like "chirurg*"), personal experience with female physicians and with same-sex parenting, which were assessed to explore their moderating effects on participants' interpretations. To evaluate participants' gender attitudes, we included measures of feminist identity on a 1–5 Likert scale and the Modern Sexism Scale (continuous 1–5 scale) (Swim et al. 1995).

4 Experimental Study

4.1 Methods and Participants

We recruited 1013 high schoolers from 11 high schools in North-Western Italy and completed the study in class on their mobile phones by answering an online survey created using Qualtrics. The schools were recruited on a voluntary basis, thanks to the active collaboration of the first author in a previous research project on an unrelated topic (scientific disinformation). Data collection was carried out by a member of the research team, who visited the participating schools in person. In accordance with the preregistered analysis plan, we excluded non-native Italian speakers ($N = 132$) and respondents who reported previous knowledge of the riddle ($N = 118$). Other participants were also excluded because they provided uninterpretable answers ($N = 72$). The final sample consisted of 691 students (288 male, 385 female, 11 undisclosed, 7 non-binary/third gender). Participants were randomly assigned to one of four conditions: 'Classical riddle'($N = 188$), 'Parent-role reversal' ($N = 151$), 'Conservative GFL' ($N = 173$), 'Innovative GFL' ($N = 179$). A researcher was always present to ensure that students did not talk to each other during the survey, thus preventing spillover effects. Parental consent forms were signed prior to data collection, and on the day of the experiment, underage students completed an online assent form, while adult students completed an online informed consent form. The privacy rights of human subjects were observed throughout the study.

4.2 Materials and Procedure

We created four different versions of the riddle. Each participant was randomly assigned to one of the four conditions described in Sect. 1.6: '*Classical riddle*' con-

dition, '*Parent-role reversal'* control condition, '*Conservative GFL'* condition, or '*Innovative GFL*' condition. After the riddle, participants were asked "How is it possible?" and could provide their answers using an open-ended format. We registered the time participants spent on this question as an indication of the time they needed to come up with a solution to the riddle. We then asked participants whether they had previous knowledge of the riddle. We also asked the participants in the 'Innovative GFL' condition whether they had ever seen the use of high stars at the end of words (*) before. Furthermore, we asked participants to indicate how difficult they perceived the riddle to be on a five-point Likert scale (1 = very easy; 5 = really hard).

Similarly to Belle et al. (2021), at the end of the study, we asked participants to provide additional information to explore whether specific factors were predictive of participants' responses. We asked participants for: (1) basic demographics: age, gender, and type of school attended; (2) gender attitudes: feminist identity on a 5-point Likert scale and the eight items of the Modern Sexism scale (Swim et al. 1995),[5] (3) familiarity with same-sex parents and female physicians (see Annex B for details on the questions). We also controlled whether participants with more familiarity with same-sex parents were more likely to indicate that the surgeon is a parent of the same sex as the one who died in the car accident.

The raw data for the study, and our coding scheme, are available online on the Open Science Framework: [https://osf.io/kj8fv/?view_only=1f169360ba474a8d99ffef885f050e5b].

The three authors manually[6] and independently coded the responses, with the primary goal to identify the statistically plausible answer (the surgeon is the opposite-sex parent). During the coding process, we inductively identified various wrong or less likely response patterns. We cross-checked codes and agreed on a consolidated set of frames. Dubious cases were discussed together by the three authors until a consensus was reached. The categories of responses created as part of the coding procedure are the following:

1) Opposite sex/gender parent;
2) Same-sex parent;
3) Biological father (e.g., Cheating, Stepfather, Adopted twins, Kidnapping);
4) Surgeon's mistake
5) Supernatural event/Hallucination/Dream;
6) Cannot be solved | "I don't know" answers | "Not possible" answers | Contested premise (the father in the car cannot be the boy's father) | "The first/second parent is not a real parent" answers;
7) Other forms of failure to identify the opposite-sex/gender parent;
8) Misunderstanding of the text | "It's obvious" answers | Trolling | Not relevant/intelligible answers | "Luck"/"randomness" answers | Uncodifiable | "The father is the grandfather" answers.

[5] The internal consistency of the resulting index is good (Cronbach alpha: 0.78).

[6] No artificial intelligence tools were used in the coding procedure.

Following a conservative approach, the participants who provided responses coded in category 8 were excluded from the final sample, since it was not univocal that they failed to identify the opposite-sex parent.

In the regression framework, we controlled for the following variables:

- Age in years
- Gender identity (male/female/non-binary or third gender/refuses to answer)
- School track attended (14 types of academic, technical, or vocational tracks)
- Having had an employed mother during childhood (yes/no)
- Index of familiarity with female doctors (continuous scale 0–1, based on the simple average of the following indicators: having had a female general practitioner, GP, in the past; currently having a female GP; having a female relative who works as a doctor; having seen a female surgeon in TV/web series)
- Index of familiarity with same-sex parenting (continuous scale 0–1, based on the simple average of the following indicators: having same-sex parents; having a friend or acquaintance with same-sex parents; having seen same-sex parents in TV/web series)
- Self-assessed feminist identity: single item on a 1–5 Likert scale
- Modern Sexism Scale (continuous 1–5 scale).

5 Results

5.1 Descriptive Comparisons across Experimental Conditions

Table 2 provides a general overview of our three dependent variables (probability to provide the statistically plausible answer; time spent on the answer; self-perceived difficulty of the riddle) across the four experimental conditions. Comparing the first two rows provides the first evidence on the degree of gender bias. It is evident that the probability of providing the most statistically plausible answer (i.e., the surgeon is the opposite-sex parent) is much greater in the 'Parent-role reversal' condition (91.4%) than in the 'Classical riddle' condition (18.6%).

Table 2 The table presents descriptive statistics for our three dependent variables by experimental condition: (i) percentage of participants who provided the most statistically plausible answer to the riddle (i.e., identifying the surgeon as the opposite-gender parent);(ii) average and standard deviation for the time taken to solve the riddle (measured in seconds); average and standard deviation for the self-perceived difficulty of the task (measured on a 5-point Likert scale, where 1 = "Very Easy" and 5 = "Very Difficult")

Experimental Condition	Percentage of most likely answer	Average timing (std. dev.) [sec]	Average self-perceived difficulty (std. dev.) [scale 1 to 5]
Classical riddle	18.6%	116.73 (77.76)	3.19 (1.14)
Parent-role reversal	91.4%	120.91 (77.95)	2.75 (1.32)
Conservative GFL	37.6%	109.80 (82.72)	2.93 (1.18)
Innovative GFL	52.5%	106.10 (61.65)	2.83 (1.14)

In other words, it is much easier to imagine the surgeon as a male than as a female. The participants also perceived the classical riddle as more difficult (3.19) than the control one (2.75). However, the time spent to provide an answer is slightly higher in the 'Classical riddle' condition (116.73 s) than in the 'Parent-role reversal' condition (120.91). Compared to the 'Classical riddle' condition, the conservative GFL strategy increases the percentage of most-likely answers (37.6%) and decreases the time spent answering (109.8 s) as well as the self-perceived difficulty (2.93). The innovative GFL strategy does it to an even larger extent across the three dependent variables (52.5%, 106.10 s, 2.83 respectively).[7]

5.1.1 Multivariate Regressions

In order to formally test our hypotheses, we resort to multivariate regressions,[8] which also allowed us to control for possible differences between treated and controls generated by random variability and therefore achieve a greater precision of the estimates. Specifically, we estimated ordinary-least-square (OLS) regressions for self-perceived difficulty and duration, and logistic regressions for the probability of providing the most plausible solution. For the logistic regressions, we report results as Average Marginal Effects (AMEs). We run three series of models. In the first one, we only control for basic socio-demographics (respondent's gender and age) and school track. In the second one, we additionally control for having had a working mum during childhood (yes or no), familiarity with female doctors (scale 0–1, mean 0.56, sd: 0.20), and familiarity with homoparental families (scale 0–1, mean 0.32, sd: 0.17). In the third one, we additionally control for the Modern Sexism scale (scale 1–5, mean 3.6, sd: 0.70) scale and for feminist identity (scale 1–5, mean 3.3, sd 1.18).

Our main parameters of interest are the coefficients associated to the 'Parent-role reversal' condition (to test H1, existence of a male bias), to the 'Conservative GFL' condition (to test H2, conservative GFL decreases the male bias), and to the 'Innovative GFL' condition (to test H3, innovative GFL decreases the male bias), as compared to the reference category 'Classical riddle'. Figures 1 to 3 graphically display the estimates across the three regression models and for our three dependent variables: probability od solving the riddle, time spent on the answer, and self-perceived difficulty. More specifically, Fig. 1 displays the Average Marginal Effects (AMEs) estimated after the logistic regression and Figs. 2 and 3 report the coefficients of the OLS regressions. Annex D in the Supplementary material reports the full results of the regression analyses.

Figure 1 reveals the existence of a very large male bias in the probability of providing the most statistically plausible answer: across the three regression models, respondents exposed to the classical riddle are 73 percentage points (p.p.) more likely to identify the surgeon as the mother than those exposed to the 'Parent-role reversal' condition to identify the surgeon as the father. In contrast, the time spent on the riddle

[7] Note that although the use of high-star (*) is an innovative linguistic form, 91.6% of participants in the 'Innovative GFL' condition reported having seen the symbol before.

[8] Annex C in the Supplementary material reports ANOVA results for the two continuous dependent variables, consistent with the regression results presented in the main paper.

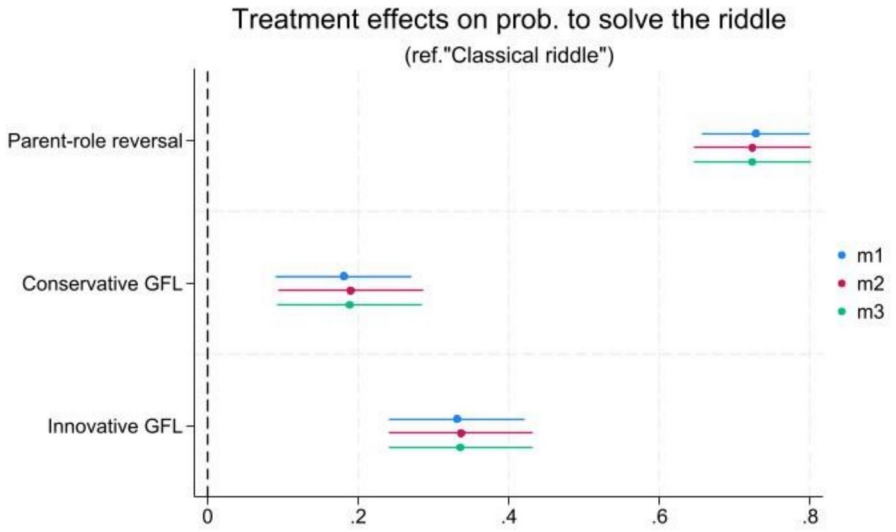

Fig. 1 The figure illustrates the Average Marginal Effects (AME) of the experimental treatments ('Parent-role reversal' condition, Conservative GFL, and Innovative GFL) of the logistic regression on the likelihood of providing the most plausible answer to the riddle compared to the baseline 'Classical riddle' condition (dashed zero line), with 95% confidence intervals. Controls in m1: respondent's gender, age, and school track. Controls in m2: m1 + working mum during childhood, scale for familiarity with female doctors, and scale for familiarity with homoparental families. Controls in m3: m2 + Modern Sexism scale and feminist identity

is similar between the two conditions (5 s difference, not statistically significant), as visible from Fig. 2. Turning to Fig. 3, respondents perceive the classical riddle as more difficult to solve than the riddle where the gender roles are reversed, with a 0.5 points difference (d = 0.4, statistically significant at 1%).

In an additional set of models, run for each dependent variable, we interact the experimental conditions with the gender identity of the respondent in order to test H1 bis (stronger male bias for respondents who identify as males, as opposed to those who identify as females). The students who identify as non-binary are too few to include as a separate category. However, as robustness checks, we run multiple models where students who classify as non-binary are excluded, merged with females, and merged with males. Across all models, the interaction term is not statistically significant at 10%, leading us to reject H1 bis.[9]

Turning to the hypotheses concerning the potential role of GFL in mitigating the male bias (H2 and H3), we find that respondents exposed to the 'Conservative GFL' condition were significantly more likely to provide the most plausible solution than participants in the 'Classical riddle' condition (18 p.p. difference, stat. sig. at 1%) and perceived the riddle as less difficult (0.3 points difference, d = 0.25, statistically significant at 1%). However, we did not detect any statistically significant differences across the two GFL conditions ('Conservative' vs. 'Innovative'). Similarly, the likelihood of the most plausible solution was higher under the 'Innovative GFL' condition

[9] We have shared on OSF the code to ensure full replicability of all statistical models.

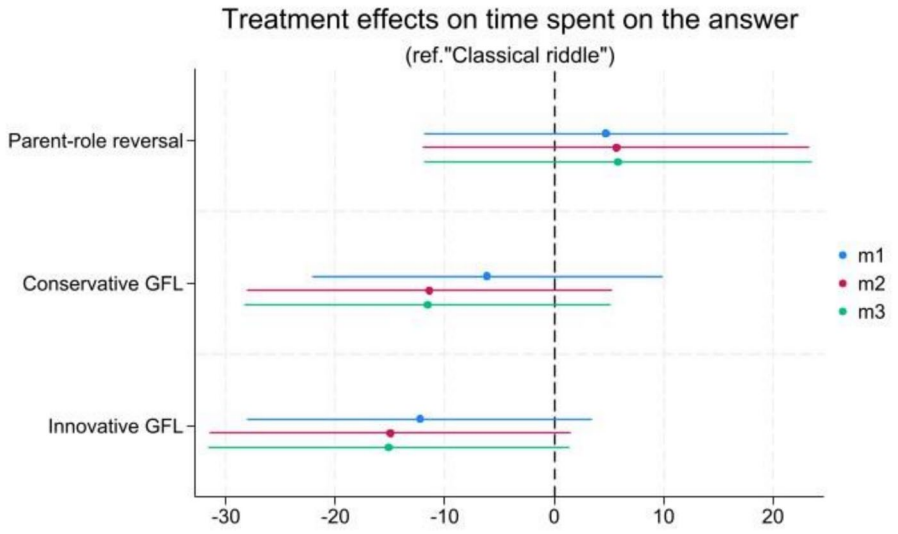

Fig. 2 The figure displays the coefficients of the experimental treatments ('Parent-role reversal' condition, Conservative GFL, and Innovative GFL) of the OLS regression on the time taken to solve the riddle (measured in seconds) compared to the baseline 'Classical riddle' condition (dashed zero line), with 95% confidence intervals. Controls in m1: respondent's gender, age and school track. Controls in m2: m1 + working mum during childhood, scale for familiarity with female doctors, and scale for familiarity with homoparental families. Controls in m3: m2 + Modern Sexism scale and feminist identity

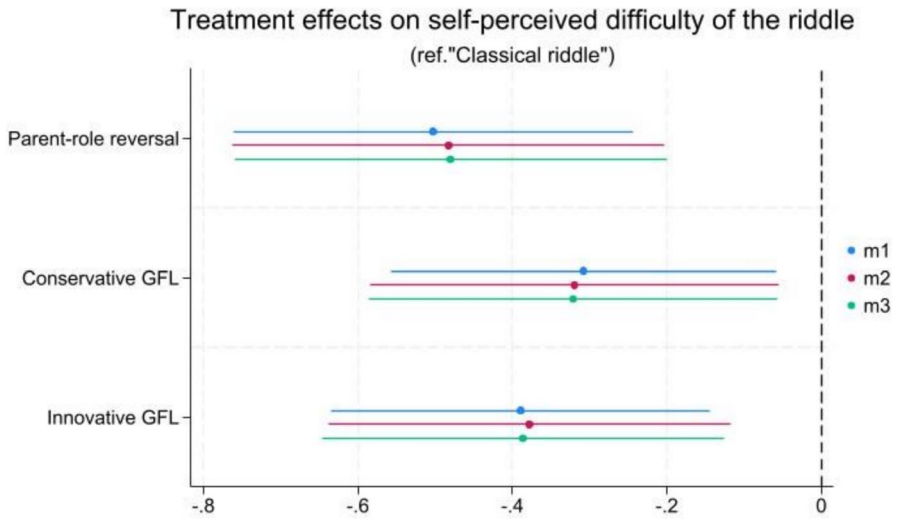

Fig. 3 The figure illustrates the coefficients of the experimental treatments ('Parent-role reversal' condition, Conservative GFL, and Innovative GFL) of the OLS regressions on self-perceived difficulty (measured on a 5-point Likert scale) compared to the baseline 'Classical riddle' condition (dashed zero line), with 95% confidence intervals. Controls in m1: respondent's gender, age, and school track. Controls in m2: m1 + working mum during childhood, scale for familiarity with female doctors, and scale for familiarity with homoparental families. Controls in m3: m2 + Modern Sexism scale and feminist identity

than under the 'Classical riddle' condition (33 p.p. difference, stat. sig. at 1%), and participants found it easier to answer (0.4 points difference, d = 0.3, stat. sig. at 1%), but the time spent on the question was similar.

Hence, the multivariate hypothesis testing supports H1, H2, and H3 on two dependent variables over three. The apparently contradictory results concerning timing may be explained by the peculiarity of the 'Parent-role reversal' condition: the story where the mother dies in the car accident and the surgeon is the father may have been too straightforward, leaving participants puzzled by the question "How is it possible?". This could have led them to spend additional time on the task, attempting to identify hidden complexities or overanalyzing the task. Consequently, the increased response time in the 'Parent-role reversal' condition may not reflect the cognitive effort related to stereotype-driven difficulties but rather the confusion stemming from the mismatch between the narrative clarity and the nature of the question. This interpretive issue may have offset the expected timing differences between conditions, resulting in a non-significant finding for this variable.

Finally, to test our alternative hypotheses concerning the relative effectiveness of innovative vs. conservative obscuration GFL strategies, we rerun the regressions by shifting the reference category to 'Innovative GFL'. The results indicate that participants exposed to the 'Conservative GFL' condition are less likely to provide the most plausible solution (15 p.p. difference, stat. sig. at 1 or 5% levels depending on the model specifications).[10] In contrast, we did not detect any significant difference across the two GFL conditions in terms of perceived self-difficulty and time taken to answer.

5.2 Exploratory Analyses

We explored the heterogeneity of the Male Bias (relative to H1) and of the two GFL alternatives (relative to H2 and H3) across participants' socio-demographic characteristics and attitudes. To this scope, we ran additional regression models where we interacted the experimental conditions with: (i) age; (ii) having had a working mother during childhood; (iii) index of familiarity with female doctors; (iv) feminist identity; (v) Modern Sexism scale. (See above for the details on variable construction). None of these analyses provides evidence of statistically significant differences in either of our dependent variables, with a single exception: the interaction between the 'Conservative GFL' condition and working mother in the probability of solving the riddle and in the self-perceived difficulty.[11] In other words, the (obscurative) conservative strategy resonated particularly among participants whose mother was working during their childhood, allowing them to access the mental representation of the woman surgeon more often. However, since this interaction was not detected for the 'Innovative

[10] Note that the confidence intervals reported in Figs. 1–3 refer to the main hypothesis testing (each condition against the 'Classical riddle' condition). Hence, the overlap of confidence intervals of the two GFL conditions does not reflect a lack of statistically significant difference between the two, which we tested in a different regression model with 'Innovative GFL' as reference category.

[11] We have shared on OSF the code to ensure full replicability of all statistical models.

GFL' condition, we cannot exclude that the statistically significant difference may also be a false positive.

6 General Discussion

The present study examined how grammatical gender in Italian can shape mental representations, particularly in the context of the "specialist riddle," where the most plausible solution is that the surgeon is the boy's mother. By manipulating the linguistic formulation of "surgeon," we aimed to determine whether alternative expressions could reduce the persistent male bias observed in similar scenarios (Stoeger et al. 2004; Kollmayer et al. 2018; Belle et al. 2021; Morehouse et al. 2022).

Our results show a pronounced gender bias among Italian adolescents when interpreting the generic masculine "chirurgo" (surgeon), a bias that persists even in a younger demographic that might be expected to hold more egalitarian views (Shu & Meagher 2018): compared to 91.4% of respondents who provide the statistically plausible answer in the 'Parent-role reversal' condition, only 18.6% did so in the 'Classical riddle' condition.[12]

This finding is strikingly similar to rates reported in English-speaking contexts (Belle et al. 2021; Morehouse et al. 2022), suggesting that the bias transcends linguistic boundaries and highlights the persistent stereotype that surgeons are men. Importantly, the introduction of GFL strategies significantly improved participants' ability to propose that the surgeon is the mother. Compared to the 18.6% in the 'Classical riddle' condition, the conservative periphrasis in the 'Conservative GFL' condition nearly doubled the likelihood of identifying the surgeon as the mother (37.6%). The innovative form Chirurg* increased this further to 52.5%. Thus, both GFL conditions substantially disrupted the default male interpretation.

Our data highlight how the use of the generic masculine ("chirurgo") constrains readers' ability to imagine a female surgeon, thereby perpetuating male-dominated interpretations of professional roles. While previous research in English showed that people struggle to identify the surgeon as female when confronted with "surgeon" (Belle et al. 2021; Morehouse et al. 2022), our study adds the specificity that, in a grammatically gendered language like Italian, seeing the masculine form "chirurgo" may strengthen masculine representations even more. This is plausible if we assume that the English word "surgeon", being grammatically neutral, might be cognitively processed as similar to the Conservative GFL condition.

As discussed in Sect. 1.3.2, two main theoretical explanations help account for this phenomenon: (1) the masculine form "chirurgo" may prime a male-associated mental representation—an effect similar to that described in studies on grammatical gender, where even inanimate referents can evoke masculine or feminine characteristics (Vigliocco et al. 2005); (2) following Gygax et al. (2021), a grammatically masculine

[12] It may be argued that answering to the riddle in the 'Parent-role reversal' condition is obvious and that the share of respondents who provide the statistically plausible answer ("the father") should approximate 100%. However, across all conditions, respondents may fail to answer correctly due to misunderstanding and lack of attention. Moreover, when faced with the 'Parent-role reversal' condition, respondents may be puzzled by obviousness of the riddle as specified in the 7.2.

word like "chirurgo" can ambiguously mean either a male surgeon or a surgeon of unspecified gender. The cognitive complexity involved in managing this dual interpretation tends to result in a preference for interpreting the term as masculine rather than gender-neutral. The effects observed depend on the difficulty of disambiguating the case where masculine is used in a proper sense (e.g., when referring to a male surgeon) and those where masculine is used in a generic way (e.g., when referring to a surgeon whose gender is unknown).

Despite the significant improvement in the GFL conditions, with 37.6% in the 'Conservative GFL' condition and 52.5% in the 'Innovative GFL' condition, these percentages still fall short of the 91.4% achieved in the 'Parent-role reversal' condition version of the riddle, where the surgeon is clearly the father following the mother's death. This discrepancy can be explained by two primary factors.

Beyond occupational stereotypes tied to specific professions, a broader gender bias exists wherein masculinity is implicitly treated as the default or prototypical human representation—a phenomenon known as the *male-as-default bias* or *people=male bias*. This androcentric tendency has been extensively documented across disciplines, showing how cultural norms and linguistic conventions reinforce the assumption of maleness in generic contexts. For instance, studies have shown that when participants are asked to imagine an unspecified "person" or "someone," they overwhelmingly envision men (Hamilton 1991, 2006). This bias extends even to ostensibly neutral entities: everyday items such as "slices of bread" can elicit male associations, underscoring the pervasiveness of androcentrism in cognition (Wardle et al. 2022, 2023). Experimental and corpus-based evidence further supports this view. Hamilton (1991) showed that men are more likely than women to be perceived as representative of the category "person," while Merritt and Kok (1995) found that contextual cues only marginally mitigate this bias. Large-scale corpus analyses reveal that the concept of "people" is disproportionately associated with men (Bailey et al. 2022). Moreover, psycholinguistic studies show that generically intended masculine forms (e.g., "he") automatically activate male imagery during sentence processing (Redl et al. 2021), and even visually neutral figures are often assigned male pronouns and narratives (Lindsay 2023). Collectively, this evidence illustrates that *male-as-default* is a deeply rooted cognitive bias that persists across linguistic, visual, and conceptual domains.

The second explanatory factor concerns entrenched occupational stereotypes, particularly the strong cultural association between surgery and men. Such stereotypes are resilient and may not be fully disrupted by brief linguistic manipulations, especially in contexts where they are continuously reinforced by social and professional norms (Linscheid et al. 2020). Our research design cannot disentangle whether the residual bias stems primarily from *male-as-default* reasoning or from the occupational stereotype linking surgery to men. It is plausible that both mechanisms operate in tandem, jointly sustaining the persistence of gendered representations even under gender-fair linguistic conditions.

In line with Belle et al. (2021), we did not find substantial interactions between riddle-solving outcomes and the demographic or background variables we measured (e.g., school track, familiarity with female physicians, or homoparental families). Previous research on the specialist riddle has explored whether individual demographic, experiential, or attitudinal factors predict the likelihood of providing the

"mother" answer, but findings remain mixed. Belle et al. (2021), for instance, conducted a regression analysis including variables such as having an employed mother, exposure to female physicians, self-identification as a feminist or political liberal, and scores on the Modern Sexism scale (Swim et al. 1995). None of these factors significantly predicted the ability to recognize that the surgeon might be a woman, except for participants' own gender—those identifying as female were more likely to provide the "mother" solution.

By contrast, Morehouse et al. (2022) tested the robustness of the surgeon = male stereotype across seven studies with a nationally representative U.S. sample. They found that while gender and political ideology exerted smaller effects, riddle-solving rates varied significantly across age, race, education, and income. Specifically, stereotypical responses increased with age, and White participants were substantially more likely to offer the "mother" response than Black or Latinx participants. Lower levels of education and income were generally associated with higher rates of stereotypical responding. The discrepancy between Belle et al.'s and Morehouse et al.'s results likely reflects sample characteristics: Belle et al. relied on a small, relatively homogeneous sample of university students ($n = 152$), whereas Morehouse et al. analyzed a large, demographically diverse population. Importantly, even among subgroups associated with reduced stereotypic responding in Morehouse et al. (2022), the rate of "mother" responses remained low—for instance, only 22% among participants with a bachelor's degree or higher, and nearly 70% of those with a close female relative working as a doctor or surgeon still failed to provide the "mother" solution.

In light of these findings, our null results regarding demographic predictors are not surprising. Our sample was relatively homogeneous in key respects (e.g., region, age bracket), and we did not collect parental education data, which could have offered finer-grained insight into socioeconomic background. Future research should replicate and extend these analyses by including more diverse samples and contextual variables. Nonetheless, our partial control for socioeconomic background through school track revealed no significant effects. This is consistent with Pabst and Kollmayer (2023), who found that using gender-fair forms—in their study, the gender star—did not reduce text comprehensibility or aesthetic appeal, even among participants without an academic background. Similarly, our results suggest that both "persona che opera" and "chirurg*" are comprehensible to adolescents across educational tracks, indicating that gender-fair language innovations do not necessarily impose a cognitive burden or reduce clarity.

Interestingly, while participants in the 'Classical riddle' and 'Parent-role reversal' conditions spent similar amounts of time on the riddle, the perceived difficulty was higher in the 'Classical riddle' condition. One explanation for these seemingly contradictory findings is that encountering a grammatical form that triggers cognitive dissonance ("How can the boy have two fathers?") amplifies subjective difficulty without necessarily prolonging the objective solution process. By contrast, participants in the 'Parent-role reversal' condition may have spent time double-checking for any trick in what appeared to be a straightforward scenario—thus inflating their response times—but without experiencing the same level of cognitive conflict.

Our findings suggest that GFL strategies—both conservative (periphrases) and innovative (neologisms like "chirurg*")—can help disrupt default male interpretations in Italian. Notably, the innovative "chirurg*" form improved solution rates more substantially (52.5%) than "persona che opera" (37.6%), yet both forms significantly outperformed the generic masculine baseline. Meanwhile, there were no marked differences in perceived difficulty or response times across the two GFL conditions, paralleling the results of Pabst and Kollmayer (2023), who also reported no adverse effects of GFL on text comprehensibility. Together, these points indicate that fears about reduced readability or acceptance of innovative GFL forms may be overstated, and that such strategies can be a potent means of challenging occupational stereotypes.[13]

6.1 Limitations and Future Directions

Despite its strengths—such as a relatively large sample of adolescents from diverse school tracks—this study has some limitations. We lacked detailed parental education data, which might help clarify how socioeconomic factors interact with riddle solving and GFL. Furthermore, our focus on a single occupational stereotype (surgeon) cannot be easily generalized to other professions, especially those seen as more gender-neutral or even female-dominated. Future research could compare multiple professions across languages with and without grammatical gender, systematically examining how each factor (e.g., GFL vs. no GFL, stereotypically male vs. female vs. neutral professions) affects interpretation. Cross-cultural studies comparing Italian adolescents to peers in other countries with different grammatical and cultural contexts may also shed light on whether the strength of the male-as-default bias varies with linguistic structure.

7 Conclusion

This study, by explicitly comparing the generic masculine "chirurgo" with gender-fair alternatives such as "persona che opera" and "chirurg*," makes the effect of linguistic form on mental representations especially clear. Even minimal changes prompted significantly more participants to hypothesize that the surgeon is a woman, thus confirming the potential of GFL to mitigate male-dominant interpretations of professional roles. These findings contribute to ongoing debates on GFL in Italy and other countries with gendered languages by highlighting how deliberate, small-scale linguistic reforms can challenge entrenched stereotypes and foster more inclusive perceptions of who can occupy high-status professions.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

[13] For a discussion of whether innovative strategies may pose challenges for visually impaired and dyslexic individuals, see Giovine (2026).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13164-025-00794-9.

Acknowledgements Cella's work on this publication has been supported by the Juan de la Cierva grant JDC2022-049764-I, by grant CEX2021-001169-M funded by MICIU/AEI/10.13039/501100011033, by the research group LOGOS (2021-SGR-00276), by the project Paradoxes as Philosophical Tools, funded by the Spanish Ministry of Science, Innovation and Universities - Code: PID2023-150066NB-I00, and by the Fundação para a Ciência e a Tecnologia under project UIDP/00183/2020.
Between 2023 and 2025, this research was presented in its early stages at the Faculty of Philosophy of Vita-Salute San Raffaele University (CRESA Work-In-Progress Seminar Series). Earlier versions of this manuscript were presented at the 3rd European Experimental Philosophy Conference, the 15th Conference of the Italian Society for Analytic Philosophy, the CEPDISC'23 Conference on Discrimination, the Logos Seminars series (University of Barcelona), the New Directions in Social Philosophy of language workshop (Nova Institute of Philosophy), the 5th Annual SWIP Italia Conference, the 5th European Experimental Philosophy Conference, the 2nd Linguistic Intersections of Language and Gender conference, the XI Conference of the Spanish Society for Analytic Philosophy, and as a seminar at the Interfaculty Centre for Gender Study (San Raffaele University). We would like to thank Claudia Bianchi, Diego Marconi, and Martina Rosola for their insightful suggestions, all the people who provided feedback during our presentations, and two anonymous reviewers for their comments on previous drafts of the manuscript.

Author Contributions All three authors contributed equally to the conception, design, analysis, and writing of the study.

Funding Open access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability The raw data for the study are available online on the Open Science Framework: https://osf.io/kj8fv/?view_only=1f169360ba474a8d99ffef885f050e5b.

Declarations

Ethics Approval The study was approved by the Ethical Committee of the University of Torino (Protocol No. 0183640 dated 15/03/2023 – [UOR: S1000045 – Classif. III/14]).

Conflict of Interest The authors declare no conflict of interest.

References

Alwin, D.F., and J.A. Krosnick. 1991. Aging, cohorts, and the stability of sociopolitical orientations over the life span. *American Journal of Sociology* 97 (1): 169–195.
Bailey, A.H., A. Williams, and A. Cimpian. 2022. Based on billions of words on the internet, people= men. *Science Advances* 8 (13): eabm2463.

Belle, D., A.B. Tartarilla, M. Wapman, M. Schlieber, and A.E. Mercurio. 2021. I can't operate, that boy is my son!": Gender schemas and a classic riddle. *Sex Roles*. https://doi.org/10.1007/s11199-020-01211-4.

Bem, S. L. 1993. *The lenses of gender: Transforming the debate on sexual inequality.* Yale University Press.

Block, N. 1996. Mental paint and mental latex. *Philosophical Issues, 7* (Perception), 19–49.

Chalmers, D. J. 1996. *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Dretske, F. 1995. *Naturalizing the mind*. MIT Press.

Fodor, J. A. 1975. *The language of thought*. Harvard University Press.

Fodor, J. A. 1987. *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press.

Formato, F. 2019. An overview of grammatical gender in Italian. In *Gender, Discourse and Ideology in Italian* (pp. 39–80). Palgrave Macmillan.

Gabriel, U., P. Gygax, O. Sarrasin, A. Garnham, and J. Oakhill. 2008. Au-pairs are rarely male: Norms on the gender perception of role names across English, French and German. *Behavior Research Methods* 40: 206–212.

Garnham, A., U. Gabriel, O. Sarrasin, P. Gygax, and J. Oakhill. 2012. Gender representation in different languages and grammatical marking on pronouns: When beauticians, musicians, and mechanics remain men. *Discourse Processes* 49 (6): 481–500.

Giovine, M. (2026). Reconciling inclusion and accessibility: Solutions for non-binary linguistic strategies in grammatical gender languages. *Language Sciences*, *113*, 101778.

Gygax, P., and U. Gabriel. 2008. Can a group of musicians be composed of women? Generic interpretation of French masculine role names in the absence and presence of feminine forms. *Swiss Journal of Psychology* 67 (3): 143–151.

Gygax, P., U. Gabriel, O. Sarrasin, J. Oakhill, and A. Garnham. 2008. Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. *Language and Cognitive Processes* 23 (3): 464–485.

Gygax, P., U. Gabriel, A. Lévy, E. Pool, M. Grivel, and E. Pedrazzini. 2012. The masculine form and its competing interpretations in French: When linking grammatically masculine role names to female referents is difficult. *Journal of Cognitive Psychology* 24 (4): 395–408.

Gygax, P.M., L. Schoenhals, A. Lévy, P. Luethold, and U. Gabriel. 2019. Exploring the onset of a male-biased interpretation of masculine generics among French speaking kindergarten children. *Frontiers in Psychology* 10: 1225.

Gygax, P., S. Sato, A. Öttl, and U. Gabriel. 2021. The masculine form in grammatically gendered languages and its multiple interpretations: A challenge for our cognitive system. *Language Sciences* 83: 101328.

Hamilton, M.C. 1991. Masculine bias in the attribution of personhood: People= male, male= people. *Psychology of Women Quarterly* 15 (3): 393–402.

Hendriks, I., M. Lubbers, and P. Scheepers. 2024. Individual change in rejection of equal opportunities for foreigners among adolescents and young adults in Switzerland: Testing realistic conflict theory from a dynamic perspective. *PLoS One* 19 (2): e0296883.

Hentschel, T., L.K. Horvath, C. Peus, and S. Sczesny. 2018. Kick-starting female careers. *Journal of Personnel Psychology* 17 (4): 193–203.

Irmen, L., and J. Kurovskaja. 2010. On the semantic content of grammatical gender and its impact on the representation of human referents. *Experimental Psychology* 57: 367–375.

Jubas, K. 2013. Grey('s) identity: Complications of learning and becoming in a popular television show. *Review of Education, Pedagogy, and Cultural Studies* 35 (2): 127–143.

Kollmayer, M., A. Pfaffel, B. Schober, and L. Brandt. 2018. Breaking away from the male stereotype of a specialist: Gendered language affects performance in a thinking task. *Frontiers in Psychology* 9: 985.

Kousta, S.T., D.P. Vinson, and G. Vigliocco. 2008. Investigating linguistic relativity through bilingualism: The case of grammatical gender. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34 (4): 843–858.

Kriegel, U. (Ed.). 2013. *Phenomenal intentionality*. Oxford University Press.

Lindsay, J. 2023. Male default: When images of neutral characters from picture books are not so neutral. *Gender Issues* 40 (1): 119–143.

Linscheid, L.J., E.B. Holliday, A. Ahmed, J.S. Somerson, S. Hanson, R. Jagsi, and C. Deville Jr. 2020. Women in academic surgery over the last four decades. *PLoS One* 15 (12): e0243308.

Menegatti, M., & Rubini, M. (2017). Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication*. Oxford University Press.

Merritt, R.D., and C.J. Kok. 1995. Attribution of gender to a gender-unspecified individual: An evaluation of the people= male hypothesis. *Sex Roles* 33: 145–157.
Misersky, J., P.M. Gygax, P. Canal, U. Gabriel, A. Garnham, F. Braun, Tania Chiarini, Kjellrun Englund, Adriana Hanulikova, Anton Öttl, Jana Valdrova, Lisa Von Stockhausen, and S. Sczesny. 2014. Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behavior Research Methods* 46: 841–871.
Morehouse, K.N., B. Kurdi, E. Hakim, and M.R. Banaji. 2022. When a stereotype dumbfounds: Probing the nature of the surgeon= male belief. *Current Research in Ecological and Social Psychology* 3: 100044.
Moulton, J., G.M. Robinson, and C. Elias. 1978. Sex bias in language use: "Neutral" pronouns that aren't. *American Psychologist* 33 (11): 1032–1036.
Moulton, J. 1981. The myth of the neutral "man." In *Sexist Language: A Modern Philosophical Analysis*, ed. M. Vetterling-Braggin, 100–116. Adams: Littlefield.
Nitti, P. 2021. Feminine-specific job titles: A research on sexism in the Italian language. *Modern Italy* 26 (1): 1–12.
Pabst, L.M., and M. Kollmayer. 2023. How to make a difference: The impact of gender-fair language on text comprehensibility amongst adults with and without an academic background. *Frontiers in Psychology* 14: 1234860. https://doi.org/10.3389/fpsyg.2023.1234860.
Pitt, D. 2020. Mental representation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition). https://plato.stanford.edu/archives/win2020/entries/mental-representation/
Pulijana, J., & Stevens, G. (2023). What's so bad about being a manageress? *Inquiry*, 1–25. https://doi.org/10.1080/0020174X.2023.2177727
Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. MIT Press.
Redl, T., S.L. Frank, P. De Swart, and H. De Hoop. 2021. The male bias of a generically-intended masculine pronoun: Evidence from eye-tracking and sentence evaluation. *PLoS One* 16 (4): e0249309.
Reynolds, D., A. Garnham, and J. Oakhill. 2006. Evidence of immediate activation of gender information from a social role name. *Quarterly Journal of Experimental Psychology* 59 (5): 886–903.
Rosola, M., Frenda, S., Cignarella, A. T., Pellegrini, M., Marra, A., & Floris, M. (2023). Beyond Obscuration and Visibility: Thoughts on the Different Strategies of Gender-Fair Language in Italian. In *CLiC-it 2023. Proceedings of the 9th Italian Conference on Computational Linguistics. Venice, Italy, November 30-December 2, 2023* (Vol. 3596, pp. 1–10). CEUR-WS.
Schärer, F. 2008. *Père, mère, roi et sorcière: la représentation des deux sexes et de la catégorie du genre dans les manuels scolaires de l'école primaire de la Suisse alémanique et de la Suisse romande* (Vol. 327). Peter Lang.
Schneider, J. W., & Hacker, S. L. 1973. Sex role imagery and use of the generic "man" in introductory texts: A case in the sociology of sociology. *The American Sociologist*, 12–18.
Schunack, S., and A. Binanzer. 2022. Revisiting gender-fair language and stereotypes–A comparison of word pairs, capital i forms and the asterisk. *Zeitschrift Für Sprachwissenschaft* 41 (2): 309–337.
Sczesny, S., M. Formanowicz, and F. Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in Psychology* 7: 154379.
Shu, X., and K.D. Meagher. 2018. Beyond the stalled gender revolution: Historical and cohort dynamics in gender attitudes from 1977 to 2016. *Social Forces* 96 (3): 1243–1274.
Skorinko, J. L. M. 2018. Riddle me this: Using riddles that violate gender stereotypes to demonstrate the pervasiveness of stereotypes. *Psychology Learning & Teaching* 17 (2): 194–208.
Stoeger, H., A. Ziegler, and H. David. 2004. What is a specialist? Effects of the male concept of a successful academic person on performance in a thinking task. *Psychology Science* 46 (4): 514–530.
Stout, J.G., and N. Dasgupta. 2011. When he doesn't mean you: Gender-exclusive language as ostracism. *Personality and Social Psychology Bulletin* 37 (6): 757–769.
Sulis, G., and V. Gheno. 2022. The debate on language and gender in Italy, from the visibility of women to inclusive language (1980s–2020s). *The Italianist* 42 (1): 153–183.
Swim, J.K., K.J. Aikin, W.S. Hall, and B.A. Hunter. 1995. Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology* 68: 199–214.
Vergoossen, H.P., E.A. Renström, A. Lindqvist, and M. Gustafsson Sendén. 2020. Four dimensions of criticism against gender-fair language. *Sex Roles* 83: 328–337.
Vervecken, D., and B. Hannover. 2015. Yes I can! Effects of gender fair job descriptions on children's perceptions of job status, job difficulty, and vocational self-efficacy. *Social Psychology* 46: 76–92.

Vigliocco, G., and J. Franck. 1999. When sex and syntax go hand in hand: Gender agreement in language production. *Journal of Memory and Language* 40 (4): 455–478.

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

Wardle, S.G., S. Paranjape, J. Taubert, and C.I. Baker. 2022. Illusory faces are more likely to be perceived as male than female. *Proceedings of the National Academy of Sciences of the United States of America* 119 (5): e2117413119.

Wardle, S.G., L. Ewing, G.L. Malcolm, S. Paranjape, and C.I. Baker. 2023. Children perceive illusory faces in objects as male more often than female. *Cognition* 235: 105398. https://doi.org/10.1016/j.cognition.2023.105398.