

Gender-Inclusive Language in Labour Markets: Lessons From a Digital Field Experiment

Anica Waldendorf¹ · Klarita Gërkhani² · Arnout van de Rijt³

¹ University of Oxford, Nuffield College

² Vrije Universiteit Amsterdam, School of Business and Economics

³ European University Institute, Department of Political and Social Sciences

This study leverages a digital field experiment in a labour market setting to examine whether attitudinal effects translate into real-world behaviour, while offering broader lessons for the design and interpretation of such experiments. Gender-inclusive language in job descriptions has been shown to impact perceived fit, belonging, and occupational attractiveness, but its impact on actual behaviour remains uncertain. We used a crowdworking platform as an experimental labour market, where sign-up for our advertised study itself serves as an outcome variable. In this setting, we thus tested whether the use of gender-inclusive language in a job advertisement for a maths task increases both the share of women who sign-up and their subsequent performance. We find no effect of gender-inclusive language on either the share of women or their performance. We provide a look behind the scenes of running a digital field experiment, reflecting on challenges along the way. We used the opportunity to learn about design elements that may have prevented the predicted effect from arising, and the moderating theoretical variables they allude to. One key conjecture that emerged from our assessment is that in tight labour markets, such as the crowdfunding platform we studied, participation is less gendered, with an attenuating effect on gendered performance expectations.

Keywords: gender-inclusive language; supply-side; job advertisements; digital field experiment

1 Introduction

We report on a study that used an online crowdworking platform as an experimental labour market, with study sign-up and ensuing performance as outcome variables. Crowdworking platforms are real-life micro labour markets in which researchers act as employers. Systematic variation in

the message used for recruitment allowed us to measure the behavioural effect of using gender-inclusive language (GIL) in job advertisements for a maths task. This approach was inspired by Mason and Watts (2009), who used the crowdworking platform Amazon Mechanical Turk to test whether the type of compensation (piece rate or quota rate) affects performance.

Our contribution is two-fold. First, we present our findings on the behavioural effect of using GIL in a real-world recruitment context. Previous research has consistently documented attitudinal effects of GIL, including perceived fit, sense of belonging, and attractiveness of occupations, yet demonstrations of downstream behavioural change are scarce. In our study we find no effect of GIL on either of two behavioural outcomes: neither the fraction of women signing-up nor women's performance once signed-up is impacted by GIL. These findings falsify our pre-registered predictions that were derived from signalling and stereotype threat theory. The findings challenge the

Supplementary Information The online version of this article (<https://doi.org/10.18148/srm/2026.v20i3.8554>) contains supplementary material.

Corresponding author: Anica Waldendorf, University of Oxford, Nuffield College, Oxford, United Kingdom (Email: anica.waldendorf@nuffield.ox.ac.uk)

AI Use Disclosure AI was used to improve style and grammar, to shorten sections, as well as to assist with coding, especially troubleshooting.

behavioural generalizability of prior attitudinal findings, which have previously been taken as suggestive evidence that GIL can mitigate gendered self-selection and performance gaps. By testing the signalling and stereotype mechanisms in a setting where individuals make real participation and effort decisions, our results provide a stringent behavioural test of these claims about the power of GIL.

Second, we use the null results in our study as an opportunity to carefully reflect on challenges in running a digital field experiment and what can be learned from the absence of the expected behavioural effects. Null findings in experimental research are often implicitly attributed to limitations in design, implementation, or statistical power, particularly when they contradict well-motivated theoretical expectations. This asymmetry contributes to publication bias and may inadvertently shield theories from meaningful empirical scrutiny. A more productive approach, which we follow, is to link aspects of experimental design to theoretical conditions under which a predicted effect may be expected to arise. Our assessment identifies two features of the experimental setting in particular that may explain the absence of a behavioural effect for our two outcome variables: sign-up and performance.

The first feature is a tight labour market and its consequences for gendered participation. In settings where job competition is low and everyone easily finds a job regardless of their demographics, stereotypes, even if present, are less activated (Gërxhani et al. 2023). In this case, members of negatively stereotyped groups may feel less threatened and are thus less likely to selectively avoid domains where they expect disadvantage. This suggests that GIL may exert a stronger behavioural effect in slack labour markets where job competition is higher, which in turn might more strongly activate stereotype threat.

The second feature is self-selection owing to our field experiment set-up (see Elwert and Winship (2014) on endogenous selection bias). Our study was based on voluntary participation, meaning that individuals could have self-selected into the study based on maths affinity. This selection was a necessary feature for the first outcome (sign-up), but created a conservative test for the second outcome (performance). If GIL had increased sign-up, self-selection would have been reduced, however, since we found no difference in sign-up, self-selection can help explain the null finding on performance: Individuals with low maths affinity may have been less likely to participate, yet the expected effect of GIL on performance is greatest for them.

2 Background and Motivation

In recent years, several languages have experienced an increase in the use of GIL in different forms. GIL refers to

noun constructions that make person nouns inclusive towards women (e.g., Studentinnen und Studenten) or towards all genders (e.g., Student*innen or Student:innen) and is used as an alternative to the generic masculine (GM) (e.g., Studenten).

Studies across different languages and countries have consistently shown that using GIL makes women more salient in people's minds, both for binary forms (see Stahlberg et al. (2007) and Kotthoff and Nübling (2019) for an overview) and more recently also for the gender star, a non-binary inclusive form (e.g., Student*innen) (Körner et al. 2022; Kurz and De Mulder 2023; Zacharski and Ferstl 2023). This effect is present in different languages such as Dutch, French, German, Italian, Polish, and Spanish (Bojarska 2011; Gygas et al. 2008; Horvath et al. 2016; Kaufmann and Böhner 2014; Redl 2020).

One context in which GIL has received considerable attention and where the potential real-world impact is high is the labour market. Nearly a quarter of the gender pay gap can be attributed to women and men working in different occupations, whereby the occupations that men work in tend to have higher salaries (European Commission 2022). Rather than looking at the increasingly studied 'demand side' of the problem, e.g., discrimination against women, this study looks at the 'supply side', i.e., women's job applications. Women's behaviour in applying for jobs may be driven by their own preferences, social norms, or institutional settings. We focus on GIL as a potential tool that may increase female labour supply in the context of occupational gender segregation. The potential of GIL lies in its ability to reduce the stereotype triggered by reading a role noun that is masculine-typed (Duffy and Keir 2004; Gabriel et al. 2008a; Gygas et al. 2021) and in making masculine-typed jobs more attractive or increasing the interest in them for girls and women (Bem and Bem 1973; Hentschel et al. 2018; Stout and Dasgupta 2011; Vervecken et al. 2013).

Ample research indicates a positive effect of GIL on attitudes and preferences of women toward masculine-typed jobs, increasing interest, sense of belonging, application intentions, and perceived suitability (Bem and Bem 1973; Damelang and Rückel 2021; Hentschel et al. 2018; Hetjens and Hartmann 2024; Horvath and Sczesny 2016; Stout and Dasgupta 2011; Vervecken et al. 2013). Under the assumption that attitudes carry over to behaviour, such effects of GIL would have real-world implications for gender inequality and segregation in labour markets. However, existing research on GIL and job applications does not test actual application behaviour, which is also noted by Böhm et al. (2020). While attitudes and behaviour may correlate, they do not necessarily perfectly align (Ajzen and Fishbein 1977; Hainmueller et al. 2015; Petzold and Wolbring 2019; Wille and Derous 2017; Wille and Derous 2018).

Three studies provide first insights into behavioural outcomes of linguistic features closely related to GIL and the labour market. Castilla and Rho (2023) find no impact of gendered wording (Gaucher et al. 2011) on job application behaviour, yet Krichelli-Katz and Regev (2021) find that women perform better in maths when addressed in the feminine, as opposed to the masculine. Finally, Gay et al. (2018) show a negative association between speaking a gender marked language and married immigrant women's labour force participation. The majority of this effect is attributed to culture, but one-third potentially stems from the language itself.

Our contribution is testing whether GIL increases both the quantity of women's supply (i.e., how many apply for a job) and the quality of supply (i.e., how well applicants perform after applying). We test an application effect (RQ1) and a performance effect (RQ2a and RQ2b) in Germany and in Italy (see below on this country selection). We ask: do more women apply for a male-typed task when it is advertised in GIL? Second, do women perform better when a male-typed task is advertised or described using GIL? We look at both application behaviour and subsequent performance because application behaviour is the first channel through which GIL could have a gendered impact before it may carry over to gender differences in performance. We operationalize GIL using the gender star, a fully inclusive variant of GIL that aims to include all genders.

3 Potential Mechanisms

We propose that women are more likely to apply for, and perform better in, masculine-typed tasks or jobs when these are advertised in gender-inclusive rather than generic masculine language. This proposition is grounded in several theoretical perspectives, at the core of which is the assumption that a job which is seen as typically masculine creates gendered expectations about who is well suited to this job and who will perform well at this job. While masculine language can reinforce these stereotypes, GIL may counteract them.

Cultural stereotypes that women are weaker in mathematics are well documented (e.g., Kaiser et al. (2012) for Germany; Cappelletti et al. (2022) for Italy). These cultural beliefs about women's and men's expected competence can impact their choices of occupation (Charles 2011; Correll 2001; Correll 2004).

The theoretical perspectives that can explain this are: a perceived lack-of-fit and lack of belonging (Heilman 1983; Heilman and Caleo 2018), an implicit stereotype that math = male (Nosek et al. 2002), the triggering of a stereotype threat and consequent underperformance (Steele et al. 2002), and gendered performance expectations when gen-

der acts as a status characteristic (Correll and Ridgeway 2006) that can lead to different evaluations despite objectively equal performance (Foschi et al. 1994). As a result, women may decide not to apply to the position in the first place or even if they do apply, they underperform.

We argue that GIL may mitigate these processes through two main mechanisms. First, the use of GIL can interrupt the association between maths and men by sending a *signal* of belonging (Connelly et al. 2011), thereby increasing the perceived fit and sense of belonging. Previous research has shown that signals can be used to increase a sense of belonging in computer science (Cheryan et al. 2009) and in tech fields (Wynn and Correll 2017). These mechanisms have also been specifically studied for GIL, which yields the same pattern: a positive link between using GIL and women's perceived fit (Damelang and Rückel 2021; Hentschel et al. 2018; Hentschel et al. 2021; Horvath and Sczesny 2016) and a higher sense of belonging (Castilla and Rho 2023; Rosenberger and Claypool 2023; Stout and Dasgupta 2011).

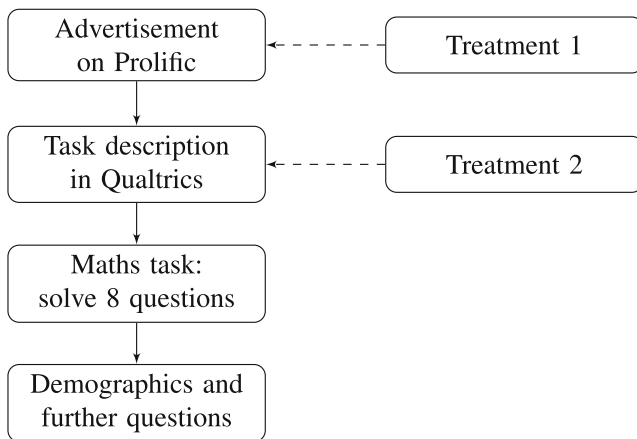
The second mechanism refers to GIL's impact in mitigating stereotype presence, thereby lowering the *stereotype threat* and increasing anticipated performance. Research shows that GIL reduces male bias in mental representations (Stahlberg et al. 2007) and can diminish the stereotypicality of roles, especially when female forms are presented first (Gabriel et al. 2008a).

In short, describing typically masculine positions in GIL can enhance women's perceived fit and sense of belonging, and at the same time reduce stereotype threat. Hence, we expect that using GIL to advertise masculine-typed jobs will increase women's likelihood of applying and enhance their performance. In contrast, men's performance should remain unaffected, as they are not confronted with a negative stereotype.

4 Experimental Design

We conducted a field experiment using the online platform Prolific. Prolific is a survey platform designed for researchers to easily access participants (Palan and Schitter 2018). We leveraged the marketplace setting: researchers (or organizations) recruit participants for their studies, essentially hiring them for a job (i.e., taking the survey). Potential participants have a dashboard on which they can see the different surveys they can opt to take. This allowed us to offer the same job (participating in our study) under different conditions. Thus, we could measure real behaviour in an online (micro-)job market.

It should be noted that Prolific provides specific conditions for research. Prolific operates under a spot-market logic, in which services are instantly exchanged for pay-

**Fig. 1***Experimental Design*

ment. This means that Prolific provides an anonymous environment with no personal contact with others. It is a digital workplace with no colleagues or superiors and no application process. Thus, Prolific presents a set of specific conditions in which the effect of GIL can be tested.

Using a 2×2 factorial design, we advertised maths tasks and varied the use of GIL in the task advertisement as well as the task description (see Fig. 1). The task was advertised using either a masculine role noun or a gender-inclusive one (in Italian and German). With this design, we could untangle differences between selection effects and treatment effects, separating differences in performance from preferences for or against GIL. This meant that some participants were exposed to both GIL and GM. However, inconsistent use of GIL is fairly common in job advertisements (Krome 2021) and other instances, such as newspaper articles (Waldendorf 2024). Therefore, we assumed that individuals would not be irritated or surprised by the variation in GIL use.

In each condition, we orthogonally varied the use of the gender-exclusive GM and the gender-inclusive gender star. We chose to represent GIL as a gender star, as opposed to a binary GIL approach, because German law mandates that job advertisements be gender inclusive towards all genders, not just women and men. A formally non-binary-inclusive approach could also be achieved through the addition of “m/f/d” to a masculine noun, yet Horvath and Sczesny (2016) and Hetjens and Hartmann (2024) suggest that adding “m/f” or “m/f/d” does little to counter the masculine noun.

We used an alternating block design to advertise the different tasks to mitigate any possible day or daytime effects while ensuring that only one condition was visible at a time. This meant we split each day into four time slots. On day one, the GIL advertisement was online during the first and

third time slots, while the GM advertisement was online during the second and fourth time slots. On day two, we flipped the sequence. We continued rotating the sequence over all the days. Participants who had participated once were blocked from participating again. Once a participant had accepted the task on Prolific, they were led to Qualtrics, where the randomization took place. In each GIL or GM treatment block, half the participants received a GIL task description and half received a GM task description (see Fig. 1).

Two different dependent outcomes were measured: application and performance. Application was the share of women (vs men) participating under each condition (whereby the availability of women was assumed to be balanced in the design, see Appendix 5). The performance was the mean overall performance under each condition. Women’s performance was the mean performance of women under each condition, and men’s performance was the mean performance of men under each condition.

The experiment ran for 10 days (Mondays–Fridays) from 14th June 2023–27th June 2023 with four blocks per language per day. We did not run a pilot in order to maximize the number of participants in the main experiment and to reach maximum power, as we expected to exhaust the available participant pool on Prolific, as we explain in the next section.

4.1 Case Selection

We simultaneously ran the experiment in Italy and Germany and pooled the results (see Appendix 1 for details, especially considerations concerning power). Both countries have grammatical gender languages (meaning that nouns are gender-marked, thus the potential for using GIL is high), language is similarly regulated, the masculinity score following Hofstede et al. (2010) is similar, and they score very closely on the EU Gender Equality Index in 2022 (European Institute for Gender Equality 2024). This means that Germany and Italy are culturally comparable when it comes to the potential key mechanism of masculinity stereotypes. We specifically chose Italy over other European countries (such as France or Spain) because of the closest similarity to Germany, given that we wanted to pool the results.

4.2 The Task

In both conditions, a maths task was advertised. Maths was chosen as a stereotypically masculine domain. Gabriel et al. (2008a) measured the gender stereotypicality of different role names by asking participants to estimate the number of men and women in each role. Their published stereotypi-

cality scores show that mathematician is a male-typed occupation (Gabriel et al. 2008b). To avoid participants thinking they need to be actual mathematicians, we specified ‘hobby mathematicians’ as opposed to ‘mathematicians’.

On Prolific, participants saw a headline reading: “Hobby mathematicians wanted! Solve maths problems.” They saw a short description and could then choose to proceed to the study. The term ‘hobby mathematician’ was treated, being either in gender-inclusive or exclusive language. Specifically, the GIL condition used the gender star in the headline whereas the gender-exclusive version used only the generic masculine (German: GIL=“Hobby-Mathematiker*innen” and GM=“Hobby-Mathematiker”; Italian: GIL=“appassionat* di matematica” and GM=“appassionati di matematica”). In the body of the advertisement, the implementation of GIL varies slightly between the German and Italian. In German, the gender star is always used. In Italian, the gender star was used in all cases apart from the term “The participants”. Here, the gender star could not be used, therefore the pair form was used. This means that in German, “The participants” was “Die Teilnehmer*innen” in GIL and “Die Teilnehmer” in the GM. In Italian, the GIL variant was “I/Le partecipanti” and the GM was “I partecipanti”.

Once they had chosen to proceed, participants were redirected to a survey hosted on Qualtrics. Here, they were presented with an informed consent form, followed by task instructions that used either GIL or the GM. Both the informed consent, which includes a task description, and the task instructions, were treated. As with the task advertisement, the implementation of GIL again varies slightly between the German and Italian for the phrase “the participants”.

The task instructions read: “Thank you for agreeing to participate in this study. On the following pages, *participants* will see eight math problems. The questions will be displayed one after the other. *Participants* have one minute per maths problem. A bonus of £0.50 will be paid if at least four of the eight problems are solved correctly.” The full materials in the original language are included in Appendix 9.

In total, participants were given eight maths tasks, one after the other, with one minute to solve each one. The tasks were designed to be hard to solve. Participants received £1.50 (roughly 1.75 €) for their participation. In addition, a quota-based bonus was offered: if a participant solved at least four out of eight maths problems correctly, they received a bonus of 50p (roughly 60 cents). The bonus aimed to emphasize competence and strengthen the gendered performance expectations and selectivity of workers (Mason and Watts 2009). Moreover, this design feature ensured that participants only applied if they believed they would succeed. We incorporated an attention check (that

mimicked the task, see Appendix 6) and flagged people who did not pass it but allowed them to finish.

4.3 Hypotheses

Overall, based on the theoretical considerations elaborated on earlier, we expect that reading the masculine form increases the salience of gendered expectations that women are worse at maths, reducing both women’s application likelihood and performance. Reading the gender-inclusive form should weaken this stereotype, increasing women’s applications and improving their performance. Given our experimental design features, this leads to the following testable hypotheses (pre-registered at: <https://doi.org/10.17605/OSF.IO/X8FT4>):

H1: When a masculine-typed task is advertised in GIL, a higher share of applicants are women compared to when it is advertised in the generic masculine (GM).

H2a: Women who apply for a masculine-typed task perform better when GIL is used in the task advertisement than when GM is used instead.

H2b: Men who apply for a masculine-typed task have the same performance whether GIL or GM is used in the task advertisement.

H3a: Women who apply for a masculine-typed task perform better when GIL is used in the task description than when GM is used instead.

H3b: Men who apply for a masculine-typed task have the same performance whether GIL or GM is used in the task description.

5 Results

5.1 Descriptives

1321 people participated in our study (569 women, 731 men, and 21 non-binary people). In line with our pre-analysis plan, we analysed only women and men and excluded the 21 non-binary cases (see Appendix 7). We did not achieve our target sample size of 2000. A full data and replication package is available (Waldendorf et al. 2026). Overall, the sample is relatively young and highly educated (mean age of 30, 57% hold a university degree). The political orientation leans slightly to the left (mean of 2.9 on a 1–7 scale). 43% use some kind of GIL (written, spoken,

or both). The mean attitude towards GIL is slightly more negative than positive (4 on a scale of 1 = very positive, 7 = very negative), while the interest in gender issues leans towards a higher interest (4.3 on a scale of 1 = not at all, 7 = very much). We include a breakdown of the descriptives by gender in Appendix 8.

The use of GIL was measured using a post-hoc questionnaire and could therefore include a post-treatment effect. Differentiating GIL use by task advertisement treatment showed slightly different levels between conditions: when GIL was used in the task advertisement, reported GIL use was 44% compared to 41%. This difference is rather small. Looking across all four conditions, the reported GIL

use was 46% when GIL was used consistently and 44% when the GM was used consistently. When the GM was used for the task advertisement and GIL was used for the task description reported GIL use was 38%, compared to 43% vice versa. Thus, reported GIL use was highest when GIL was used for both the task advertisement and task description and lowest when the GM was used for the task advertisement but GIL was used for the task description. Hence, there is no consistent pattern of a post-treatment effect on reported GIL use.

5.2 Pre-Registered Hypotheses

To test our hypotheses, we used a two-tailed proportion test (H1) and two-tailed independent t-tests (H2–3). The two-tailed proportion test compared the proportion of women in the GM condition and the GIL condition. In the two-tailed independent t-tests, the performance of women (resp. men) under each condition was compared. We conducted two-tailed tests as a more robust albeit conservative test of our hypotheses that accommodate unexpected results.

Figure 2 shows the share of women and men in each condition. Hypothesis 1 predicted that when a masculine-typed task is advertised using GIL, a higher share of applicants are women compared to when it is advertised using GM. Therefore, the share of women under a GIL task advertisement was compared with the share of women under a GM task advertisement. There was no difference in the share of women who participated in the task depending on the language used in the task advertisement. When the advertisement used GIL, the share of women (vs men) was 44%, compared to 44% when the advertisement used GM ($\chi^2(1, N=1300)=0.016, p=0.9$).

To recall, hypothesis 2a predicted that women who apply for a masculine-typed task perform better when GIL is used in the task advertisement than when GM is used instead. In contrast, hypothesis 2b predicted that men who apply for a masculine-typed task have the same performance, regardless of whether GIL or GM is used in the task advertisement. To test hypotheses 2a and 2b, women's (men's) mean performance under the GIL task advertisement was compared with women's (men's) mean performance under the GM task advertisement (Fig 3). Women's mean performance when they saw the task advertised in GIL was slightly lower ($M=3.3, SD=1.56$) compared to when the GM was used ($M=3.51, SD=1.55$). This difference was not statistically significant according to a two-tailed independent t-test ($t(567)=1.61, p=0.11$). Men who saw the GIL advertisement had a marginally lower performance ($M=4.3, SD=1.84$), compared to men who saw the GM advertisement ($M=4.36, SD=1.74$). The difference was not statistically significant ($t(729)=0.42, p=0.67$).

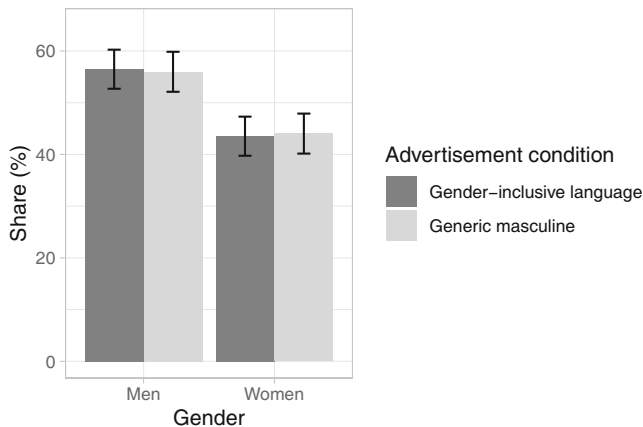


Fig. 2

Share of women and men participating in task after seeing a task advertised in either gender-inclusive language or the generic masculine. $n=1300$. Italy and Germany pooled.

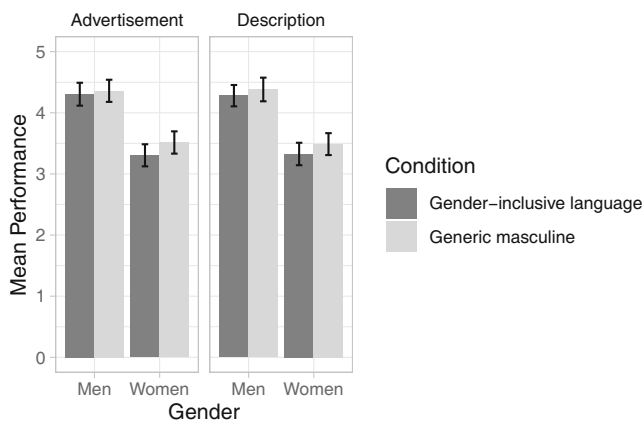


Fig. 3

Performance after seeing a GIL/GM advertisement and a GIL/GM task description. $n=1300$. Italy and Germany pooled.

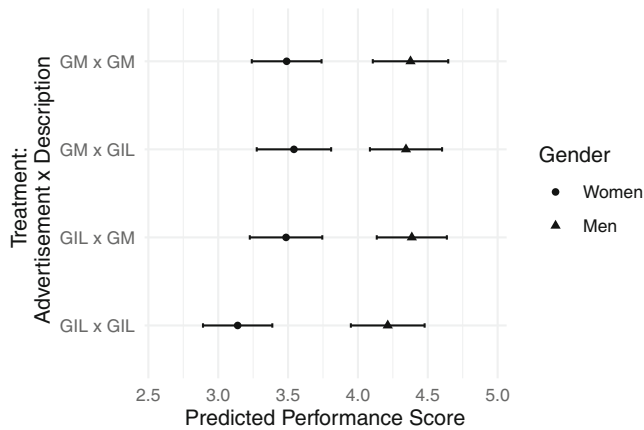


Fig. 4

Interaction between the advertisement and the task description treatments. Women $n = 569$; Men $n = 731$.

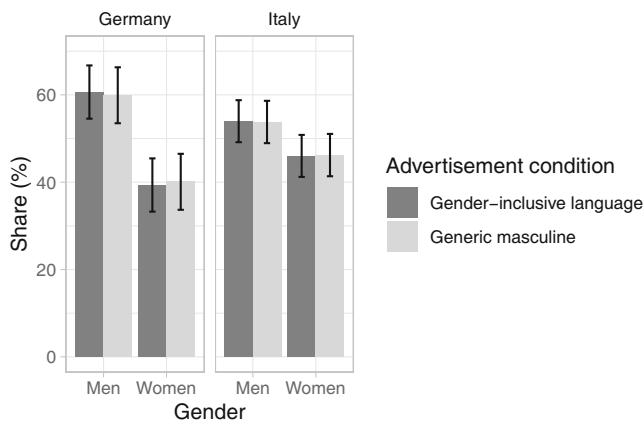


Fig. 5

Share of women participating in task after seeing GIL/GM task advertisement by country. Italy $n = 824$; Germany $n = 476$.

The final set of pre-registered hypotheses (H3a and H3b) examined women's and men's performance dependent on which type of language was used in the task description, independent of which type of language was used in the task advertisement (see Fig. 3). Women's performance was slightly lower under the GIL condition ($M = 3.33$, $SD = 1.57$) than the GM condition ($M = 3.49$, $SD = 1.54$) but the difference was not statistically significant ($t(566) = 1.24$, $p = 0.22$). For men, performance was also slightly lower under GIL ($M = 4.28$, $SD = 1.69$) compared to under GM ($M = 4.38$, $SD = 1.89$) yet this was not a statistically significant difference ($t(722) = 0.76$, $p = 0.44$).

In short, there was no support for hypothesis 1 that the share of women applying for a masculine-typed job in-

creased when GIL was used. While the performance of men remained stable across conditions, in support of hypotheses 2b and 3b, so did the performance of women, contrary to the expectations of hypotheses 2a and 3a.

6 Exploratory Analyses

In addition to our pre-registered analyses, we investigated the interaction effect between the two treatments. Moreover, we ran separate analyses for Germany and Italy. We found no interaction effect, and the results do not seem to be driven by one country. Overall, the main finding that non-binary inclusive GIL did not have an effect seems robust.

6.1 Interaction Effect

To look at a possible interaction between the two treatments, we ran OLS regressions with an interaction between the two treatment conditions for women and men, respectively.¹

We tested the interaction between the advertisement condition and task description condition on women's performance score using a linear regression. The results of the regression did not find a statistically significant effect of either treatment or the interaction between the two treatment conditions ($R^2 = 0.011$, $F(3, 565) = 2.09$, $p = 0.10$). Neither the main effect of the advertisement condition ($\beta = 0.00$, $p = 0.98$), the main effect of the task description condition ($\beta = 0.05$, $p = 0.78$), nor the interaction term ($\beta = -0.40$, $p = 0.13$) were statistically significant.

We performed the same analysis for men and also do not find a statistically significant effect ($R^2 = 0.001$, $F(3, 727) = 0.356$, $p = 0.79$). Again, neither the main effect of the advertisement condition ($\beta = 0.01$, $p = 0.96$), the main effect of the task description condition ($\beta = -0.03$, $p = 0.86$), nor the interaction term ($\beta = -0.14$, $p = 0.60$) were statistically significant.

We predicted marginal means from the linear regression models for women and men. In both models, none of the treatment effects reached statistical significance. Figure 4 shows the predicted marginal means of the performance score.

6.2 Individual Country Analyses

Overall, GIL was used more (52% vs 37%) and seen more favourably (3.8 vs 4.2, with 1 = very positive and 7 = very negative) in the German sample (see Appendix 8). While there was, in general, a higher proportion of women par-

¹ We thank both anonymous reviewers for this suggestion.

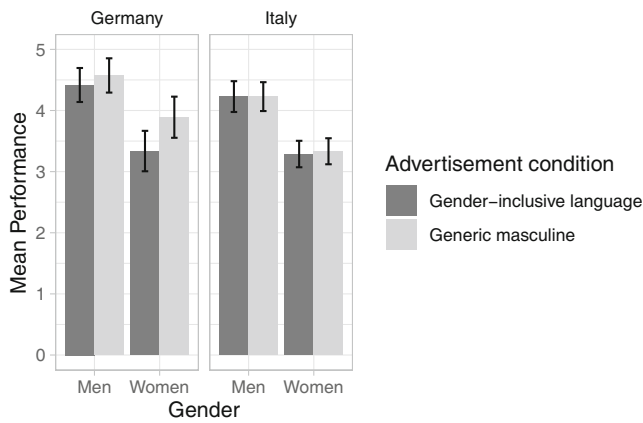


Fig. 6

Performance after seeing a GIL/GM task advertisement by country. Italy $n = 824$; Germany $n = 476$.

ticipants in the Italian sample (around 46%), there was no difference in the share of women between the GIL and the GM advertisement for either country (Fig. 5).

Concerning the performance after seeing either advertisement, women in Germany, on average, performed worse after seeing a GIL advertisement, scoring 0.55 points less (GIL advertisement: $M = 3.34$, $SD = 1.65$; GM advertisement: $M = 3.89$, $SD = 1.62$). This difference was statistically significant ($t(186.46) = 2.33$, $p = 0.02$) (Fig. 6). This was surprising as it went against our expectation. However, this statistical significance should be seen with caution. Due to the small sample size, this estimate is likely not reliable. The unreliability can be seen in the confidence intervals, which are relatively wide and overlap (GIL advertisement [3.0, 3.67]; GM advertisement [3.55, 4.22]). The SD of

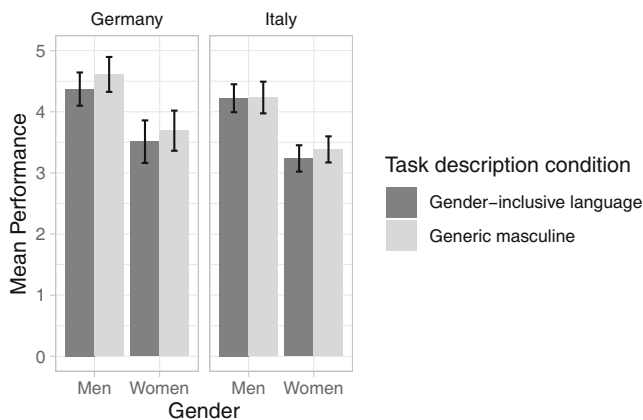


Fig. 7

Performance after seeing a GIL/GM task description by country. Italy $n = 824$; Germany $n = 476$.

around 1.6 (for a scale of 0–8) for both conditions shows that there is quite some variance amongst the performance of women within each condition. There was no difference for men in Germany, or for either gender in Italy.

The difference in performance dependent on the task description was not statistically significant between the GIL and the GM condition for either country or gender (Fig. 7). All test statistics are reported in Appendix 8.2.

7 Discussion

Overall, advertising a male-typed task on Prolific using non-binary inclusive GIL did not increase the share of women who participated. There was also no difference in task performance based on the type of language used in either the advertisement or the task description.

It seems that simply changing language does not achieve behavioural change when it comes to choosing a task on a crowd-working platform. From a wide array of experiments, there is strong support that the GM creates an image skewed in favour of men (Kotthoff and Nübling 2019; Stahlberg et al. 2007). We also know that GIL affects preferences and attitudes regarding occupations and job choice (e.g., Bem and Bem (1973); Vervecken et al. (2013)). However, this did not translate into behavioural change when choosing a task to complete on a crowd-working platform, which might mean that the effect of GIL found for intentions and attractiveness simply does not result in behavioural change. Our results are in line with recent research that finds no influence of GIL on competitiveness, leadership, or feelings of inclusion in a laboratory experiment (Balafoutas et al. 2025) and no impact of gendered wording on job application behaviour (Castilla and Rho 2023). Alternatively, the study's design and conditions may have negated possible effects of GIL. There is an indication that GIL may only have an impact under specific scope conditions (Gay et al. 2018; Hollenbach and Rohmann 2026). In combination with insights from our analysis of the open comments in our survey (see Appendix 8.3), this gives us reason to take a closer look at the conditions in our study.

8 Implementation Aspects and Theoretical Inferences

In the following, we discuss the general challenges of running a field experiment within a survey experiment framework. We then identify and relate aspects of our experimental design to theory in order to reflect on the conditions under which a behavioural effect of GIL may or may not be observed. The null findings provide an opportunity to reflect on challenges in implementing a digital field experiment and based on this to delineate boundary conditions for

the operation of the theorized mechanisms, clarifying when attitudinal effects are unlikely to translate into behavioural outcomes. We organize this discussion around six key features of the experimental context that may explain our null finding. These features may have inhibited the operation of theoretical mechanisms, prevented the statistical detection of their behavioural consequences, or contributed to self-selection into the study.

8.1 A Tight Labour Market

We used Prolific to field our study. It is known that competition for jobs on platforms such as Prolific is weak. Still we had anticipated that there would be ample participants (see discussion on power below). However, the labour market on Prolific turned out tighter than expected. Beyond not reaching our sample size, this meant that the conditions were such that all women and men could easily find as many jobs as they would like. Under such conditions the stereotype threat perceived by participants may have been less salient, which in turn may have hindered the emergence of gendered performance expectations.

8.2 A Lower-Stakes Labour Market

We assumed that different advertisements would vie with each other, and participants must decide which tasks to complete. We believed participants would choose tasks that they could perform with the highest reward in terms of relative payment. If payment depends on performance, this creates a potential for gendered performance expectations to impact task selection. Yet, the statistical results, as well as the analysis of open questions (see Appendix 8.3), did not indicate the presence of gendered performance expectations. Our bonus functioned to render performing poorly consequential. The stakes may nonetheless have been too small. In a typical job, performance is tied to securing an income. It is plausible that poor performance in a typical job may lead to future financial consequences (e.g., not getting a promotion or being fired) or social consequences (e.g., letting down the team by performing badly and being excluded from social team activities). The risk of not getting a bonus was perhaps not high enough to discourage participants who do not think they are good at maths from participating (as suggested by the analysis of the open questions).

8.3 No Task-Related Stereotype

It is possible that the task may not have been perceived as stereotypically masculine. We deliberately advertised for ‘hobby’ mathematicians, because we wanted to avoid the impression that formal qualifications or training were required. The addition of ‘hobby’ may have suggested that enjoyment of maths, rather than high maths ability, was required. Excelling at maths is ascribed to men (Bian et al. 2018; Leslie et al. 2015), so the emphasis on ‘hobby’ may then have undermined the intended masculine stereotyping of the task. Alternatively, mathematics in general may no longer be seen as typically masculine. Actual differences in maths performance seem to be small, if present at all (Hyde et al. 2008). At the same time, Nosek et al. (2002) show that while the explicit link between maths and gender has decreased, the implicit link remains, and there is mixed evidence as to whether stereotype threat in maths performance persists (Kim et al. 2023; Stoet and Geary 2012). We did not measure participants’ beliefs regarding the stereotypicality of the task, therefore, it remains unclear whether the task was perceived as typically masculine. Hollenbach and Rohmann (2026) investigate the impact of GIL for a job that is neutral (in contrast to the previously cited studies that looked at stereotypically masculine jobs) and only find higher application intentions for trans, intersex, agender, and non-binary participants. This supports our reasoning that our null finding could be because our participants did not perceive maths as stereotypically masculine.

8.4 Achieving Power

As mentioned above, the unexpected labour market conditions proved troublesome. A further consequence was a smaller than anticipated sample size. Our power calculations indicated that we needed a sample size of 2000 participants. However, despite Prolific’s large number of active German-speaking participants in Germany (around 2000), we could only realistically expect a sample of 1000 based on Prolific’s self-reported expected participation rate of 40–50% (Prolific Team 2023). Based on our power analysis, we chose to expand from just a German sample to a German and Italian sample to ensure a sufficient number of participants. Italy had around 2400 active Italian-speaking participants. Despite this pre-emptive adjustment, we did not reach the target sample size of 2000. Instead, we reached 1321 participants (1300 if we only include women and men).

We observed different participation rates for Italy and Germany (34% compared to 23%, respectively). Especially for Germany, we had a very small sample of 476 individuals (249 participants in the GIL condition and 227 in

the GM condition). If we further split this by gender, there were only 189 German women and 287 German men. If we then look at the sample size for the performance analyses, there were 92 women in the GIL task description condition and 97 women in the GM task description condition. This makes the conclusions we can draw from this analysis more tentative. It also makes subgroup analyses unfeasible. We have reported confidence intervals in our results to indicate the reliability.

8.5 Two-Language and Two-Country Setting

Conducting the experiment in two countries meant we had to operationalize GIL in slightly different ways. Both languages have a form of GIL that uses an asterisk, albeit in slightly different ways. In German, the asterisk is used to combine the masculine and feminine ending (e.g., *Mathe-matiker*innen*), while in Italian the noun is truncated to remove all gendered endings (e.g., *appassionat* di matematica*). Furthermore, not all nouns required the same modifications owing to unique features in each language. For example, “participants” (*Teilnehmer/partecipanti*) became *Teilnehmer*innen* in German. However, in Italian, the noun is identical in the feminine and the masculine form and does not provide the possibility of truncation, so gender inclusivity had to be signalled using articles (*I/Le parteci-panti*). Including a second country ameliorated the problem of power. However, we still did not reach our desired power, which may have allowed a positive effect to have gone unnoticed.

8.6 Self-Selection

Participation in the study was voluntary, meaning that we only measured performance effects for participants who selected into our study based on maths affinity. Individuals who have a low maths affinity may be more susceptible to gendered performance expectations, especially when they do a compulsory test, and may have therefore not applied. Hence, our study can be seen as a conservative test.

Future research can address this self-selection in two ways, either by observing or by circumventing selection. To observe selection, the study design must start one step earlier than the present study. To do so, participants could be recruited to participate in a study and then in a second step choose which task to complete from a range of tasks that vary in their gender stereotypicality and use of GIL. Which task is chosen is then an outcome variable. In contrast to the present study, the researcher can experimentally control which tasks are being chosen from which makes selection observable. Alternatively, the selection problem can be cir-

cumvented, for example by leveraging compulsory testing in schools or universities. A field experiment in which the use of GIL in compulsory testing is experimentally varied would rule out selection and can measure the impact of GIL on performance.

9 Conclusion

This study provides rare behavioural evidence on how non-binary inclusive gender-inclusive language (GIL) in job advertisements affects job seekers’ application and performance. By moving beyond attitudinal measures and examining actual behavioural outcomes, the study contributes to a growing literature on inclusive language and labour market inequality. We hope that reflecting on our challenges can help other researchers wishing to develop similar research designs. While the null findings should be interpreted cautiously given the study’s constraints, they offer an important empirical foundation for future research. Specifically, they show that attitudinal effects of linguistic inclusivity need not impact labour market outcomes, and they help identify the conditions under which they do not. More broadly, this study demonstrates the feasibility of a field experimental approach to understanding how features of job advertisements influence (online) labour market participation.

Acknowledgements We would like to thank Katrin Auspurg and Isabell Diekmann as well as the editors of this special issue, Lena Hipp and Sebastian E. Wenz, and two anonymous reviewers for providing valuable comments. We would also like to thank William Foley for providing us with the maths task and Moreno Berva, Arianna Gatta, Mya Giungi, and Ophelia Nicole-Berva for the Italian translation of the experimental material.

Funding This work was supported by mission funding granted by the European University Institute.

Declarations

Ethical Standards The Ethics Committee at the European University Institute granted ethics clearance for this project (project code 20220704_WALDENDORF).

References

- Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical relations. *Psychological Bulletin*, 84(5), 888–918.
- Balafoutas, L., Fornwagner, H., Hauser, E., & Hauser, O. (2025). Gender-inclusive language and economic decision-making. *Management Science*, 72(9), 7381–8307.

- Bem, S.L., & Bem, D.J. (1973). Does sex-biased job advertising 'aid and abet' sex discrimination? *Journal of Applied Social Psychology*, 3(1), 6–18.
- Bian, L., Leslie, S.J., Murphy, M.C., & Cimpian, A. (2018). Messages about brilliance undermine women's interest in educational and professional opportunities. *Journal of Experimental Social Psychology*, 76, 404–420.
- Böhm, S., Linnyk, O., Kohl, J., Weber, T., Teetz, I., Bandurka, K., & Kersting, M. (2020). Analysing gender bias in IT job postings: A pre-study based on samples from the German job market. In Proceedings of the 2020 on Computers and People Research Conference, SIGMIS-CPR'20 (pp. 72–80), ACM, <https://doi.org/10.1145/3378539.3393862>
- Bojarska, K. (2011). Wpływ androcentrycznych i inkluzywnych płciowo konstrukcji językowych na skojarzenia z płcią [Influence of androcentric and gender-inclusive lexical constructions on gender-associational responses]. *Studia Psychologiczne*, 49(2), 53–68.
- Cappelletti, D., Levati, M. V., & Ploner, M. (2022). Math ability, gender stereotypes about math ability, and educational choices. Working Paper Series, Department of Economics, University of Verona, (7):1–49.
- Castilla, E.J., & Rho, H.J. (2023). The gendering of job postings in the online recruitment process. *Management Science*, 69(11), 6912–6939.
- Charles, M. (2011). What gender is science? *Contexts*, 10(2), 22–28.
- Cheryan, S., Plaut, V.C., Davies, P.G., & Steele, C.M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology*, 97(6), 1045–1060.
- Connelly, B.L., Certo, S.T., Ireland, R.D., & Reutzel, C.R. (2011). Signaling theory: A review and assessment. *Journal of Management*, 37(1), 39–67.
- Correll, S.J. (2001). Gender and the career choice process: The role of biased self-assessments. *American Journal of Sociology*, 106(6), 1691–1730.
- Correll, S.J. (2004). Constraints into preferences: Gender, status, and emerging career aspirations. *American Sociological Review*, 69(1), 93–113.
- Correll, S.J., & Ridgeway, C.L. (2006). Expectation states theory. In J. Delamater (Ed.), *Handbook of Social Psychology* (pp. 29–51). Boston: Springer US.
- Damelang, A., & Rückel, A.K. (2021). Was hält Frauen von beruflichen Positionen fern? Ein faktorieller Survey zum Einfluss der Gestaltung einer Stellenausschreibung auf deren Attraktivitätseinschätzung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 73(1), 109–127.
- Duffy, S.A., & Keir, J.A. (2004). Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. *Memory & Cognition*, 32(4), 551–559.
- Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40(1), 31–53.
- European Commission. (2022). The gender pay gap situation in the EU. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/gender-equality/equal-pay/gender-pay-gap-situation-eu_en#why-do-women-earn-less
- European Institute for Gender Equality (2024). *Gender Equality Index 2024: Sustaining Momentum on a Fragile Path*. Luxembourg: Publications Office of the European Union.
- Foschi, M., Lai, L., & Sigerson, K. (1994). Gender and double standards in the assessment of job applicants. *Social Psychology Quarterly*, 57(4), 326–339.
- Gabriel, U., Gygax, P., Sarasin, O., Garnham, A., & Oakhill, J. (2008a). Au pairs are rarely male: Norms on the gender perception of role names across English, French, and German. *Behavior Research Methods*, 40(1), 206–212.
- Gabriel, U., Gygax, P., Sarasin, O., Garnham, A., & Oakhill, J. (2008b). Supplementary material for 'Au pairs are rarely male: Norms on the gender perception of role names across English, French, and German'. *Behavior Research Methods*, 40(1), 206–212.
- Gaucher, D., Friesen, J., & Kay, A.C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, 101(1), 109–128.
- Gay, V., Hicks, D.L., Santacreu-Vasut, E., & Shoham, A. (2018). Decomposing culture: An analysis of gender, language, and labor supply in the household. *Review of Economics of the Household*, 16(4), 879–909.
- Görxhani, K., Brandts, J., & Schram, A. (2023). Competition and gender inequality: A comprehensive analysis of effects and mechanisms. *American Journal of Sociology*, 129(3), 715–752.
- Gygax, P., Gabriel, U., Sarasin, O., Oakhill, J., & Garnham, A. (2008). Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. *Language and Cognitive Processes*, 23(3), 464–485.
- Gygax, P., Sato, S., Öttl, A., & Gabriel, U. (2021). The masculine form in grammatically gendered languages and its multiple interpretations: A challenge for our cognitive system. *Language Sciences*, 83, 1–9.

- Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, 112(8), 2395–2400.
- Heilman, M.E. (1983). Sex bias in work settings: The Lack of Fit model. *Research in Organizational Behavior*, 5, 269–298.
- Heilman, M.E., & Caleo, S. (2018). Combatting gender discrimination: A lack of fit framework. *Group Processes & Intergroup Relations*, 21(5), 725–744.
- Hentschel, T., Horvath, L.K., Peus, C., & Sczesny, S. (2018). Kick-starting female careers: Attracting women to entrepreneurship programs. *Journal of Personnel Psychology*, 17(4), 193–203.
- Hentschel, T., Braun, S., Peus, C., & Frey, D. (2021). Sounds like a fit! Wording in recruitment advertisements and recruiter gender affect women's pursuit of career development programs via anticipated belongingness. *Human Resource Management*, 60(4), 581–602.
- Hetjens, D., & Hartmann, S. (2024). Effects of gender sensitive language in job listings: A study on real-life user interaction. *PLOS ONE*, 19(8), Article e0308072.
- Hofstede, G.H., Hofstede, G.J., & Minkov, M. (2010). *Cultures and organizations: Software of the mind: Intercultural cooperation and its importance for survival* (3rd edn.). New York: McGraw-Hill.
- Hollenbach, C., & Rohmann, A. (2026). Can gender-inclusive language function as an identity safety cue? The impact of gendered language forms on job application decisions among gender/sex minorities and majorities. *Journal of Social Issues*, 82(2), Article e70065.
- Horvath, L.K., & Sczesny, S. (2016). Reducing women's lack of fit with leadership positions? Effects of the wording of job advertisements. *European Journal of Work and Organizational Psychology*, 25(2), 316–328.
- Horvath, L.K., Merkel, E.F., Maass, A., & Sczesny, S. (2016). Does gender-fair language pay off? The social perception of professions from a cross-linguistic perspective. *Frontiers in Psychology*, 6, 1–12.
- Hyde, J.S., Lindberg, S.M., Linn, M.C., Ellis, A.B., & Williams, C.C. (2008). Gender similarities characterize math performance. *Science*, 321(5888), 494–495.
- Kaiser, G., Hoffstall, M., & Orschulik, A.B. (2012). Gender role stereotypes in the perception of mathematics: An empirical study with secondary students in Germany. In H. Forgasz & F. Rivera (Eds.), *Towards Equity in Mathematics Education: Gender, Culture, and Diversity* (pp. 115–140). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Kaufmann, C., & Böhner, G. (2014). Masculine generics and gender-aware alternatives in Spanish. *ZGOnZeit. Onlinezeitschrift des interdisziplinären Zentrums für Geschlechterforschung (IZG)*, 3(1), 8–17.
- Kim, J., Liu, R., & Zhao, X. (2023). A big (male) fish in a small pond? The gendered effect of relative ability on STEM aspirations under stereotype threat. *European Sociological Review*, 39(2), 177–193.
- Körner, A., Abraham, B., Rummer, R., & Strack, F. (2022). Gender representations elicited by the gender star form. *Journal of Language and Social Psychology*, 41(5), 553–571.
- Kotthoff, H., & Nübling, D. (2019). *Genderlinguistik. Eine Einführung in Sprache, Gespräch und Geschlecht*. Tübingen: A. Francke Verlag.
- Krichelli-Katz, T., & Regev, T. (2021). The effect of language on performance: Do gendered languages fail women in maths? *npj Science of Learning*, 6(9), 1–9.
- Krome, S. (2021). Gendern zwischen Sprachpolitik, orthografischer Norm, Sprach- und Schreibgebrauch. Bestandsaufnahme und orthografische Perspektiven zu einem umstrittenen Thema. *Sprachreport*, 37(2), 22–29.
- Kurz, P., & De Mulder, H. (2023). A star is born? The German gender star and its effects on mental representation. *Psychology of Language and Communication*, 27, 384–404.
- Leslie, S.J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219), 262–265.
- Mason, W., & Watts, D.J. (2009). Financial incentives and the “performance of crowds”. In Proceedings of the ACM SIGKDD Workshop on Human Computation (pp. 77–85). ACM. <https://doi.org/10.1145/1600150.1600175>
- Nosek, B.A., Banaji, M.R., & Greenwald, A.G. (2002). Math = male, me = female, therefore math $\neq$ me. *Journal of Personality and Social Psychology*, 83(1), 44–59.
- Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22–27.
- Petzold, K., & Wolbring, T. (2019). What can we learn from factorial surveys about human behavior? A validation study comparing field and survey experiments on discrimination. *Methodology*, 15(1), 19–30.
- Prolific Team. (2023). Audience. Prolific, available from: <https://researcher-help.prolific.co/hc/en-gb/articles/4407449546002-Audience#heading-2> [accessed 22.01.2023].

- Redl, T. (2020). *Masculine generic pronouns. Investigating the processing of an unintended gender cue*. PhD Thesis, Radboud Universiteit Nijmegen.
- Rosenberger, E.J., & Claypool, H.M. (2023). The impact of gender-exclusive language on women's anticipated ostracism: A preregistered replication of Stout and Dasgupta (2011). *PLOS ONE*, 18(9), Article e0290709.
- Stahlberg, D., Braun, F., Irmen, L., & Sczesny, S. (2007). Representation of the sexes in language. In K. Fiedler (Ed.), *Social Communication*. Frontiers of Social Psychology. (pp. 163–187). New York: Psychology Press.
- Steele, C.M., Spencer, S.J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology*, 34, 379–440.
- Stoet, G., & Geary, D.C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, 16(1), 93–102.
- Stout, J.G., & Dasgupta, N. (2011). When he doesn't mean you: Gender-exclusive language as ostracism. *Personality and Social Psychology Bulletin*, 37(6), 757–769.
- Vervecken, D., Hannover, B., & Wolter, I. (2013). Changing (s)expectations: How gender fair job descriptions impact children's perceptions and interest regarding traditionally male occupations. *Journal of Vocational Behavior*, 82(3), 208–220.
- Waldendorf, A. (2024). Words of change: The increase of gender-inclusive language in German media. *European Sociological Review*, 40(2), 357–374.
- Waldendorf, A., Gërkhani, K., & van de Rijt, A. (2026). Gender-inclusive language in labour markets – replication package. [Data set and code] OSF. <https://doi.org/10.17605/OSF.IO/9QG46>
- Wille, L., & Derous, E. (2017). Getting the words right: When wording of job ads affects ethnic minorities' application decisions. *Management Communication Quarterly*, 31(4), 533–558.
- Wille, L., & Derous, E. (2018). When job ads turn you down: How requirements in job ads may stop instead of attract highly qualified women. *Sex Roles*, 79(7–8), 464–475.
- Wynn, A., & Correll, S. (2017). Gendered perceptions of cultural and skill alignment in technology companies. *Social Sciences*, 6(2), 45.
- Zacharski, L., & Ferstl, E.C. (2023). Gendered representations of person referents activated by the nonbinary gender star in German: A word-picture matching task. *Discourse Processes*, 60(4–5), 294–319.