Proceedings of the LSA

Rethinking how we measure gender associations: An updated norming method for nouns and adjectives

Cassie Davenport & Eszter Ronai*

Abstract. This study introduces a novel norming method for measuring the gender associations of words that avoids binary Likert scales that present *male* and *female* as opposite endpoints. Here, participants provide ratings on two separate scales for a word's likelihood of referring to a woman or a man, respectively. This approach allows for a more nuanced measurement, independently capturing bias toward each gender rather than assuming they are inversely related. Results from the two-scale method are strongly correlated with traditionally used binary scale ratings, which suggests that this new method reliably captures gender bias of words without reinforcing a binary continuum. In addition to nouns, the present study extends gender norming to adjectives, which have received comparatively little attention in prior experimental work. We find that adjectives exhibit gender associations similar to those observed for nouns, which provides evidence that gendered expectations extend beyond nouns in English.

Keywords. gender associations; gender norming; norming methodology; psycholinguistics; gender bias; experimental design; language and gender; adjectives

1. Introduction. Gender norming is a common practice in psycholinguistics to experimentally establish the degree of gendered expectations speakers associate with certain words. Even with the absence of grammatical gender in English, words can still differ in how strongly they evoke a certain gender (e.g., man, woman, non-binary person). For example, in (1), speakers tend to process the reflexive pronoun *herself* more slowly than *himself*. This slowdown in reading times, known as the Gender Mismatch Effect, occurs when a noun phrase such as *the engineer*, though grammatically compatible, is perceived as mismatching the gender of the pronoun (e.g., Carreiras et al. 1996; Sturt 2003).

(1) The engineer burned {herself / himself} while on the job.

In reading experiments this effect can be utilized to test how comprehenders retrieve antecedents and resolve pronoun reference. As such, gender information is often a core feature used in psycholinguistic experiments, manipulated in order to investigate language processing mechanisms such as predictive processing (e.g., Kush & Dillon 2021), agreement checking (e.g., Hagoort & Brown 1999; Barber & Carreiras 2005), and feature-based retrieval (e.g., Dillon et al. 2013; Kush et al. 2015).

In order to determine what gendered words to use in experimental items, researchers either norm the gender associations of words or draw on previous norming results that are publicly available. However, the current accepted practice in the field when it comes to norming gender associations is a Likert scale that represents *female* and *male* as opposite endpoints (e.g., Carreiras et al. 1996; Osterhout et al. 1997; Kennison & Trofe 2003; Irmen 2007; Pyykkönen et al.

* We would like to thank the audience at the 2026 Annual Meeting of the LSA for helpful feedback. This material is based upon work supported by the National Science Foundation under Grant #BCS-2438973. All materials, data, and supplementary analyses for this study are available on the Open Science Framework (OSF): https://osf.io/7mb9j. Authors: Cassie Davenport, Northwestern University (cassandradavenport2026@u.northwestern.edu) & Eszter Ronai, Northwestern University (ronai@northwestern.edu).

2010; Scott et al. 2019), which does not align with contemporary understandings of gender (e.g., Hyde et al. 2019; Schudson & Morgenroth 2022).

Researchers have already drawn attention to potential limitations and harms of using only binary gender measurements for participant demographic questionnaires (Cameron & Stinson 2019; Bedin et al. 2024). It is important to consider these drawbacks not just with demographic questionnaires, but also with the type of experimental tasks participants are asked to complete. Given these concerns, it is worth reassessing the psycholinguistic methodologies used to detect word gender associations in language.

In this paper, we present a novel norming method that captures gender associations of nouns without using a scale that represents *male* and *female* as opposite ends of a binary dimension. We additionally implemented and discuss changes to task instructions and debrief. Altogether, we hope to provide researchers with a method to study the gender associations of words that is informed, ethical, direct, transparent, and sensitive (e.g., Bedin et al. 2024). Lastly, previous gender norming has primarily focused on the gender information carried by nouns. We present novel results showing that adjectives can also have gender associations similar to what has been long documented for nouns.

1.1. PREVIOUS GENDER NORMING STUDIES. The Likert scale used by previous norming studies is a type of bipolar scale, which is a very common rating tool in human behavioral experiments. This type of scale usually presents two endpoints that are considered antonymous (e.g., *extremely dissatisfied* to *extremely satisfied*), often with the midpoint considered neutral. Table 1 shows the scales used in previous psycholinguistic gender norming studies.[1] All studies included in this table used a bipolar scale with *male/female* or *masculine/feminine* on either endpoint. Some of these studies define the midpoint for participants but most do not include a midpoint label.

Study	Scale
Carreiras et al. (1996)	1 (strongly male) → 11 (strongly female)
Osterhout et al. (1997)	1 (extremely male) → 4 (gender-neutral) → 7 (extremely female)
Kennison & Trofe (2003)	1 (mostly female) → 7 (mostly male)
Irmen (2007)	1 (typically female) → 7 (typically male)
Pyykkönen et al. (2010)	1 (extremely masculine) → 7 (extremely feminine)
Scott et al. (2019)	1 (very feminine) → 4 (neuter) → 7 (very masculine)

Table 1. Previous Likert scales used for gender norming

1.2. PROBLEMATIZING BINARY GENDER SCALES. In contemporary gender scholarship, gender is largely considered a multidimensional and non-binary construct (e.g., Hyde et al. 2019; Schudson & Morgenroth 2022). Thus, measuring gender using a binary scale that represents *female* and *male* as opposite poles fails to represent social scientists' current understanding of gender. In this section, we will outline some concerns regarding the use of this binary scale, namely, that it inaccurately presents male and female as conceptual opposites, that it assumes that these

[1] Note that while this is not an exhaustive list, the studies included are some of the most cited norming studies psycholinguists use to select gendered nouns for experimental stimuli.

two genders are inversely related, and that it assumes that these two genders are the only possible gender categories that span the full space of gender. These highlighted issues raise concerns from a gender theory perspective as well as a methodological perspective, potentially limiting the extent to which such measures can accurately capture gendered associations as they are socially perceived.

The prevailing idea in gender theory is that gender does not fall into two opposite categories (e.g., Butler 1999). As Likert scales usually present two endpoints that are considered antonymous, representing *male* and *female* genders as opposite poles on a Likert scale assumes that these two genders are themselves opposites. This representation of gender thus contradicts prevailing theories about gender identity. Assuming that many researchers support multidimensional and non-binary models of gender, it is important to consider whether the gender norming practices they employ may nonetheless inadvertently perpetuate assumptions about gender that are at odds with those commitments.

Additionally, presenting participants with a scale where *male* and *female* are opposite endpoints forces participants to consider an inverse correlation between the two endpoints (i.e., if a word is rated more female than it must be rated as less male). This inverse relationship implies mutual exclusivity of these genders, which is not necessarily an accurate reflection of reality. Bem (1974) expressed this concern for measuring the gender identity of participants in psychological experiments and developed a method for measuring masculinity and femininity using separate scales in order to avoid assuming an inverse relationship between the two. This method asked participants to describe their gender identity using separate scales for masculine and feminine characteristics; using this method, Bem (1974) found that the dimensions of masculinity and femininity are empirically and logically independent. Some participants scored high on both scales, others scored low on both, and some scored high on one scale and low on the other. These findings demonstrate that masculinity and femininity are not inversely correlated traits. Considering that it is not possible on the binary Likert scale to rate something as high on both ends of the scale or low on both ends of the scale, it is possible that some words have associations with multiple genders, but that this is obscured in the binary scale method.

Lastly, the binary scale method introduces challenges for representing genders outside the binary, as its visual structure suggests that male and female anchor the full range of gendered associations. Under this design, participants are not provided with a way to indicate associations with gender identities that fall outside these categories (e.g., nonbinary, agender, genderfluid, etc.). Relatedly, previous gender norming has largely interpreted the midpoint of the scale as neutral, but it is often unclear whether participants are to assume that the middle represents a true absence of gender associations or a gender that is both partially male and partially female.

Taken together, these concerns point to limitations of the binary scale method that are both theoretical and methodological. To address these limitations, we introduce a two-scale gender norming method that elicits independent ratings of associations with men and women. The two-scale method we propose in the present paper directly addresses the first two issues discussed above. By eliciting independent ratings for the likelihood of being a man and the likelihood of being a woman, it avoids representing these categories as conceptual opposites and does not impose an inverse relationship between them. This design also clarifies the interpretation of mid-scale values, which can be difficult to interpret in binary scales. A midpoint rating may reflect either a lack of gender association or a combination of associations with both men and women. In the two-scale method, these possibilities can be distinguished: low ratings on both scales indicate

weak or absent association, while higher ratings on both scales could indicate an association with both genders. At the same time, it must be acknowledged that our approach does not resolve the issue of representing gender identities outside the binary. Participants are still not provided with a way to indicate associations with nonbinary, agender, or genderfluid identities. One potential extension would be to include an additional scale capturing associations with nonbinary identities, which may reveal patterns of gender association that are obscured under current binary-based methods.

1.3. GENDER ASSOCIATIONS OF ADJECTIVES. The discussion so far has centered on gender associations of nouns in English. Some adjectives in English have also been argued to carry strong gender associations, but have remained underexplored in psycholinguistics. The existence of gendered English adjectives has been shown through corpus studies that identify adjectival collocates of definitional gendered nouns (e.g., Moon 2014; Raphael 2023), as well as corpus work investigating the use and impact of gendered descriptors in recommendation letters and performance evaluations (e.g., Adams et al. 2022; Khan et al. 2023; Evans et al. 2024).

Moon (2014) presents a corpus-based analysis of adjectives used to describe age-differentiated gender categories (e.g., *young woman/man*, *middle-aged woman/man*, *old woman/man*) in English, examining differences in the frequency with which adjectives are used across these categories. Similarly, Raphael (2023) analyzes the British National Corpus and the Kolhapur Corpus of Indian English to identify adjectives that frequently co-occur with gendered nouns such as *bachelor*, *groom*, and *wife*. In both studies, frequent co-occurrence with gendered referents suggests that certain adjectives carry gendered associations. For example, *beautiful* and *gossipy* are more strongly associated with women, while *handsome* and *chivalrous* are more strongly associated with men.

Khan et al. (2023) conducted a systematic review of studies that assessed gendered language in medical reference letters for residency applications and medical faculty hiring. Results show significant differences in how men and women were described in recommendation letters, specifically in the types of adjectives used to describe the applicants. Women applicants were more likely to be described using what the authors categorize as "communal adjectives", such as *helpful* or *compassionate*, while men applicants were more likely to be described with what the authors term "agentic adjectives", such as *assertive* and *confident*. These findings show discrepancies in the language used to describe men vs. women applicants, which reflects larger societal gender expectations and stereotypes. Based on the observed trends, Khan et al. (2023) argue that these differences in adjective use can negatively affect career advancement by reflecting and reaffirming gender inequality in the workforce.

Analyzing gender patterns in teaching evaluations, both Adams et al. (2022) and Evans et al. (2024) find that instructors receive more positive evaluations (i.e., higher ratings on teaching effectiveness and student satisfaction scales) when their behaviors and attributes conform to societal gender expectations. These expectations are reflected in the adjectives used to describe instructors, which vary systematically by gender. For example, women instructors are more often described with adjectives such as *sensitive*, *nice*, or *friendly*, while men instructors are more often described with adjectives such as *confident*, *smart*, and *authoritative*. Overall, these studies show that gendered expectations shape teacher evaluations and are reflected in the descriptive language used in those evaluations.

The studies discussed in this section demonstrate that adjectives in English can carry gen-

dered associations, reflecting broader societal expectations. However, to our knowledge, such adjectives have not been a primary focus of psycholinguistic research and have therefore not been systematically included in norming studies. We are aware of only one study that includes gender association ratings for adjectives, Scott et al. (2019), which provides ratings for 5,553 English nouns, adjectives, verbs, and adverbs across nine different psycholinguistic dimensions. By focusing more narrowly on the gender associations of nouns and adjectives, the present study is able to examine a larger and more targeted set of potentially gendered adjectives than those included in Scott et al. (2019).

2. Experiment 1: Norming with Prior Methodology. While our goal in this paper is to highlight shortcomings of binary scales and to introduce a new norming method, we nonetheless first conduct an experiment using the binary scale method known from previous studies (e.g., Carreiras et al. 1996; Osterhout et al. 1997; Kennison & Trofe 2003; Irmen 2007; Pyykkönen et al. 2010; Scott et al. 2019). This is to allow for a direct assessment of whether findings can be replicated using our new method (Experiment 2).

2.1. PARTICIPANTS. Previous norming studies typically had undergraduate participants. For this reason, we also recruited undergraduate students for our experiments so as to keep participants similar, allowing for later comparisons with prior work. Participants for 1a were 35 American English speakers (age range: 18-21; mean age: 19), recruited through Northwestern University's linguistics subject pool. All subject pool participants received course credit for completing the experiment. Participants for 1b were 40 American English speakers (age range: 18-24; mean age: 20). These participants were recruited on Prolific —where we filtered participants for undergraduate student status —and compensated $3 at a rate of $12 per hour.

2.2. DESIGN AND MATERIALS. For 1a, a list of nouns was created by compiling and modifying noun lists from previous work (Carreiras et al. 1996; Irmen 2007; Kennison & Trofe 2003; Duffy & Keir 2004). The list contained 102 nouns previously found to be biased towards women, 260 nouns previously found to be biased towards men, and 44 nouns previously found to not have gender associations, for a total of 406 nouns for each participant to rate. For Experiment 1b, a list of adjectives was created by compiling and modifying datasets from corpus studies that identify adjectival collocates of gendered nouns (Moon 2014; Raphael 2023) and from work examining gendered language in evaluative contexts such as recommendation letters and performance evaluations (Khan et al. 2023; Evans et al. 2024; Adams et al. 2022), as well as by drawing on the first author's intuitions to supplement the resulting adjective list. The final list contained a total of 291 adjectives: 132 hypothesized to be associated with women, 106 hypothesized to be associated with men, and 53 hypothesized to not have gender associations.

One important departure from prior methodology concerns the labeling of the scale endpoints. In the present study, the scale is anchored with 1 = *more likely to be a* ***woman*** and 7 = *more likely to be a* ***man***. This choice, which replaces using labels such as *male/female* or *masculine/feminine*, is motivated by the type of gender information we aim to capture. Specifically, our goal is to measure gender associations grounded in social roles and expectations. While many prior norming studies share a similar goal, their use of labels such as *male/female* or *masculine/feminine* may invite interpretations tied to biological sex or inherent traits.[2] By contrast,

[2] We note that biological sex is also widely understood to be shaped by social and ideological factors; see Zimman (2014) for discussion.

using *woman/man* emphasizes gender associations as socially constructed rather than inherent. For example, while an *engineer* may be more likely to be perceived as someone who identifies as a man, there is nothing inherently male or masculine about the word *engineer* itself.

The experiments were conducted on PCIbex (Zehr & Schwarz 2018). For both 1a and 1b, participants saw one word at a time on their computer screen and were asked to rate how likely they feel that each word represents a person who is a man or a person who is a woman, using a scale from 1 to 7. Instructions specified that a rating of 1 indicates that a particular noun is very likely to represent a woman and a rating of 7 indicates that a particular noun is very likely to represent a man. Additionally, a rating of 4 indicates that a particular noun is equally likely to represent a woman or a man, while a rating of 2 or 3 and 5 or 6 would indicate different degrees of likelihood that a particular noun represents a woman or a man, respectively. Participants were also instructed to let their ratings be informed by societal gender expectations and representations and to base their ratings on how the world is and not how it ought to be, to not spend a long time thinking about each word, and to try to use the whole range of the scale. The instructions stated that gender associations are probabilistic, based on gendered representations in the world, rather than binary and oppositional. Additionally, participants were told that there is no correct gender assignment for these nouns, nor do we think there should be. Since we want to measure first impressions and initial reactions, participants were told that it may sometimes be difficult to rate a word on the given scale, but to respond as best as they can without thinking too deeply.

2.2.1. TASK DEBRIEF. The goals of the participant debrief were to clearly separate study methodology from the researchers' beliefs about gender in order to support inclusivity and respect and to maintain transparency about study goals. To be transparent about the study goals, the debrief explicitly told participants that we aim to develop other methods of testing gender biases that are more in line with contemporary gender theory, not simply representing women and men on two opposite ends of a gender binary. Participants were also informed that the current accepted method in sentence processing for gender norming is the binary scale they just used, and that their participation in this task will help to improve norming methods in psycholinguistics. In order to support inclusivity and respect for all gender identities, the debrief emphasized that we do not think there are only two genders, and we do not think that representing gender as two opposing ends of a binary scale is an accurate representation of gender. Additionally, the debrief stated that gender exists on a spectrum, meaning that there are a lot of different ways that people can express their gender. This section also provided opportunity for participant feedback to be entered in a text box, in order to continue to improve norming methods and to empower participants (Kost et al. 2025).

2.3. RESULTS. To examine the distribution of gender associations across items, we computed mean gender ratings and plotted their distribution as histograms. Mean ratings for each word were grouped into bins representing adjacent intervals on the scale (1-2, 2-3, 3-4, 4-5, 5-6, and 6-7). This procedure aggregates continuous values into discrete ranges, allowing the overall distribution of ratings to be visualized more clearly. In other words, the continuous values were grouped into intervals so that the frequency of observations within each range can be visualized. The height of each bar in the histograms corresponds to the number of words whose mean rating falls within the corresponding interval. To illustrate the types of words represented in each region of the scale, a small number of representative items from each bin were included as labels in the figures. Figure 1 shows the distribution of the mean ratings for nouns and Figure 2 shows

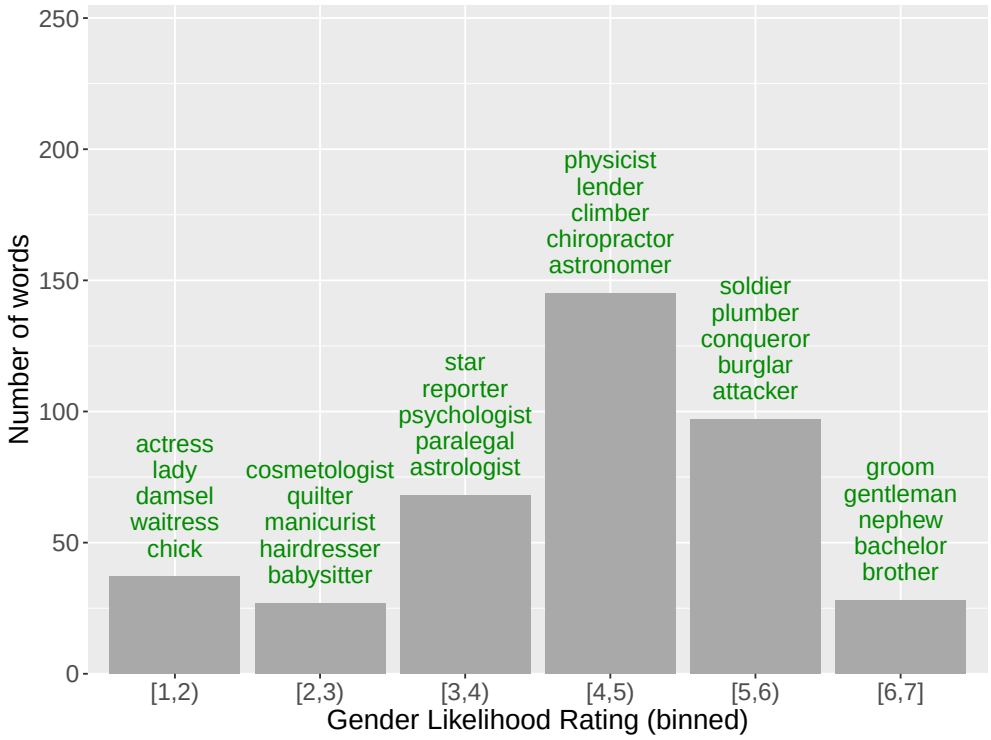


Figure 1. Experiment 1a: Noun Ratings. Continuous mean ratings grouped into bins representing adjacent intervals on the scale. Labels represent example words from each bin.

the distribution of the mean ratings for adjectives. In line with with prior work (e.g., Carreiras et al. 1996; Osterhout et al. 1997; Kennison & Trofe 2003; Irmen 2007; Pyykkönen et al. 2010; Scott et al. 2019), we find that many English nouns carry strong gender associations. Additionally, Experiment 1 produced the novel finding that English adjectives also show similar patterns of gender associations. For both nouns and adjectives, gender associations are not categorical in their strength (e.g., words are not simply either male or female); instead, gender association is a continuous property. In other words, it is not the case that words either simply have a gender association or they do not; rather, there is variability in the relative strength of gender associations. We see this variability for both nouns and adjectives.

3. Experiment 2: Novel 2-Scale Method. Experiment 2 employed a novel rating method where participants rated each word on its likelihood (0-3) of referring to a woman (Scale A), and separately, on its likelihood (0-3) of referring to a man (Scale B). This approach allows for a more nuanced measurement that independently assesses the presence or absence of bias towards either gender. This two scale method avoids some, but not all, of the issues of the binary scale that were outlined in Section 1.2. By not using a singular scale that represents *female* and *male* as opposite ends, this novel method avoids representing male and female as conceptual opposites, which would contradict prevailing theories about gender identity. Further, it avoids the assumption that these two genders are inversely related (i.e., if a word is rated more female than it must be rated as less male).

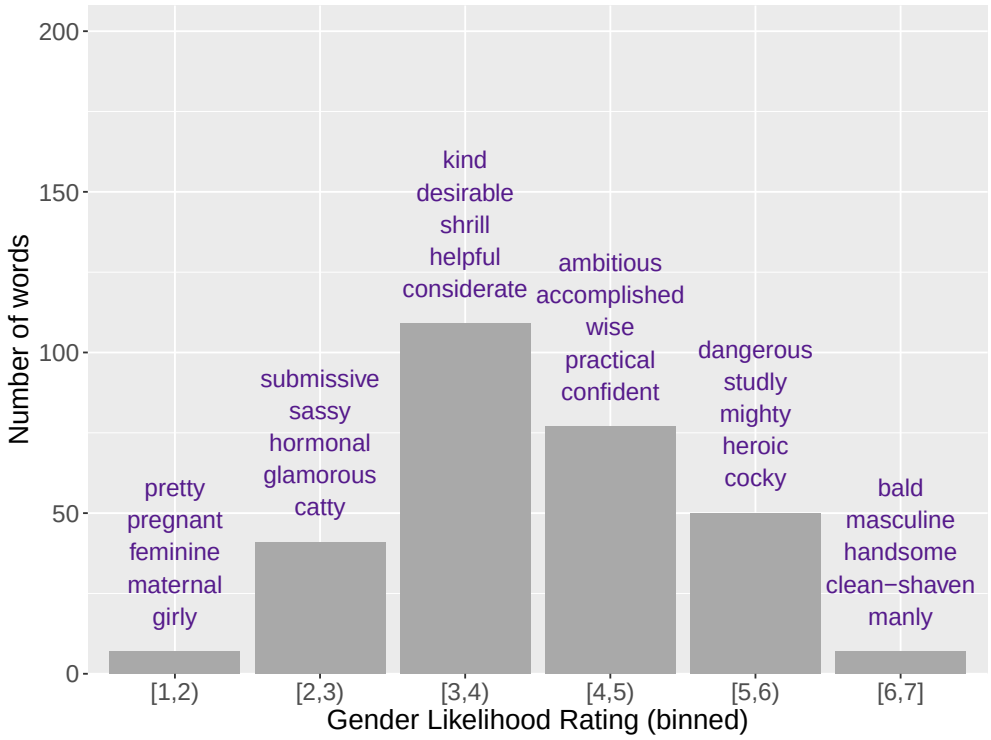


Figure 2. Experiment 1b: Adjective Ratings.

3.1. PARTICIPANTS. Participants for 2a were 38 American English speakers (age range: 18-21; mean: 19.5), recruited through Northwestern University's linguistics subject pool. All subject pool participants received course credit for completing the experiment. Participants for 2b were 41 undergraduate student American English speakers (age range: 18-24; mean: 21), recruited through Prolific and compensated $6 at a rate of $12 per hour.

3.2. DESIGN AND MATERIALS. The noun list for Experiment 2a and the adjective list for 2b were identical to those used in the previous experiment. In both experiments, participants saw one word at a time and rated its likelihood of being associated with a woman (0–3) or a man (0–3). Half of the participants completed the woman scale first, followed by the man scale, while the other half completed the scales in the reverse order. This counterbalancing was used to control for potential order effects. After completing trials with the first scale, participants were given a short break before receiving instructions and practice for the second scale.

Participants were told that a rating of 3 would indicate that a particular word is very likely to represent the given gender. A rating of 0 would indicate that a particular word is not very likely to represent the given gender. A rating of 1 or 2 would indicate different degrees of likelihood that a particular word represents the given gender. Figure 3 shows example trials from each scale.

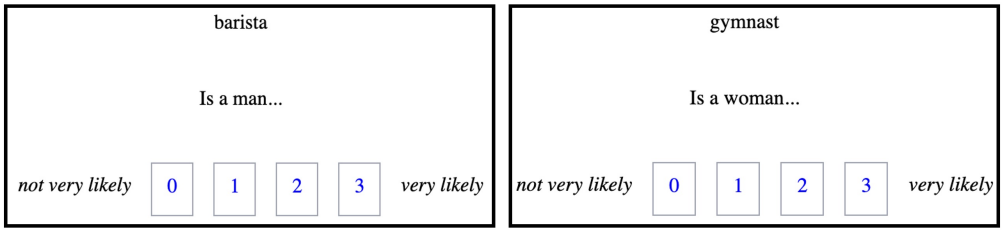
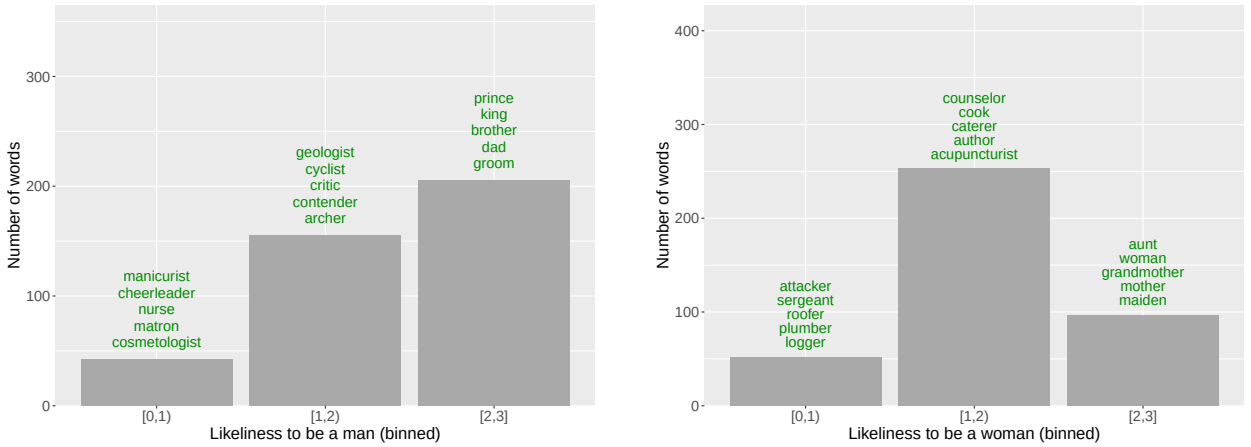


Figure 3. Experiment 2: Two-Scale Examples.

3.3. RESULTS. The average mean rating for each noun and adjective for each scale was computed. Means were grouped into discrete bins in the same way as in Experiment 1 and visualized with histograms. Figure 4 shows the distribution of the mean ratings of the nouns for both scales, and Figure 5 shows the distribution of the mean ratings of the adjectives for both scales. Overall, the patterns observed largely mirror those found in Experiment 1: gender associations are evident across items for both nouns and adjectives, but they are not categorical.

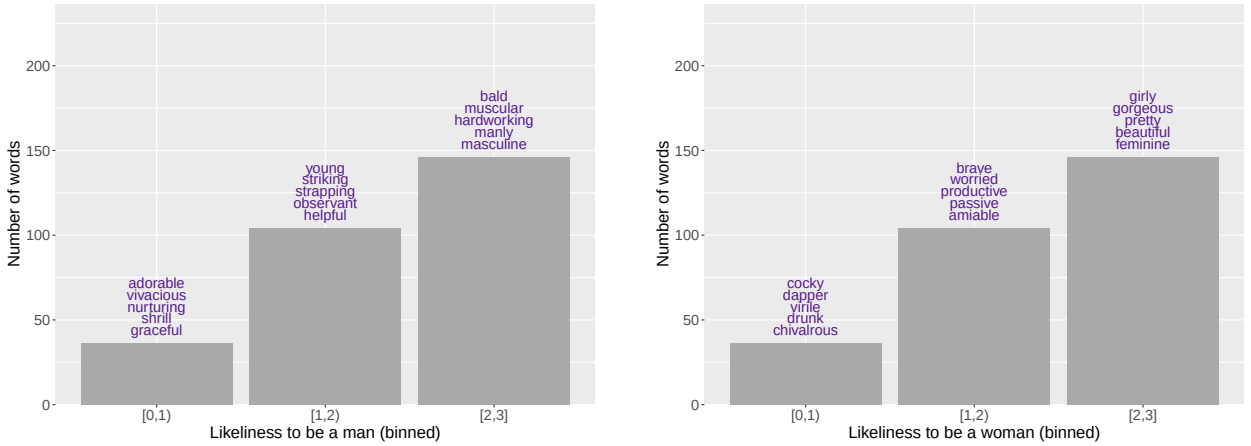


(a) Likelihood to be a man.

(b) Likelihood to be a woman.

Figure 4. Experiment 2a: Noun Ratings. Continuous mean ratings grouped into bins representing adjacent intervals on the scale. Labels represent example words from each bin.

Visually, the nouns show different distributions for each scale, with a larger number falling in the [1,2) range on the *woman* scale. To examine this further, we identified 152 nouns with mean ratings in [2,3] on the *man* scale and in [1,2) on the *woman* scale. In Experiment 1, these nouns received mean ratings ranging from 3.9 to 5.9 (mean = 4.9). Of these, 85 fall between 3.9 and 5, a range that is near the midpoint of the scale and is often treated as relatively gender-neutral in prior work (e.g., Carreiras et al. 1996; Osterhout et al. 1997; Kennison & Trofe 2003; Irmen 2007; Pyykkönen et al. 2010). This pattern highlights a limitation of the binary 7-point scale: words that appear near the midpoint of a single continuum may nonetheless show asymmetric gender associations when measured independently. Specifically, participants rated some nouns as relatively neutral when considering the likelihood of being a woman, yet rated those same nouns as reasonably more likely to be a man. Under a single-scale approach, these nouns may be interpreted as neutral, thereby obscuring their asymmetric association with one gender.



(a) Likelihood to be a man.

(b) Likelihood to be a woman.

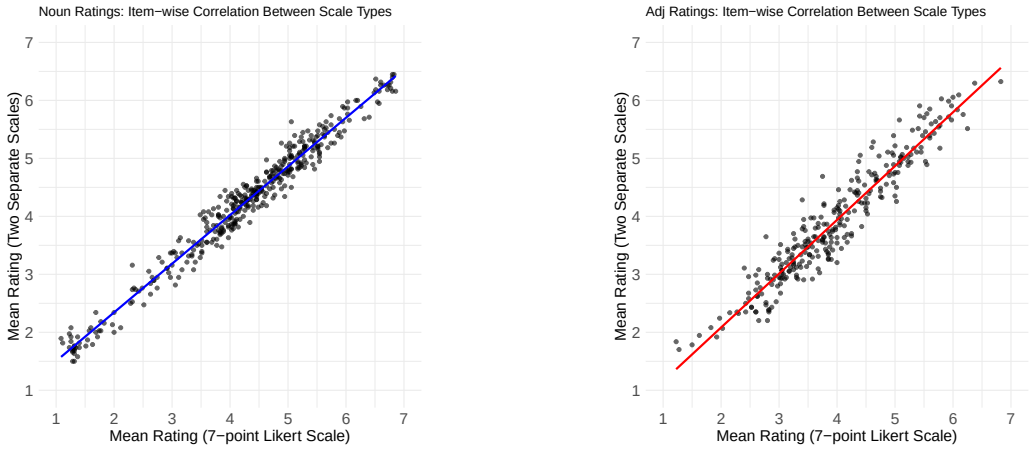
Figure 5. Experiment 2a: Noun Ratings.

These asymmetries also align with broader patterns of gender markedness in English. In particular, the asymmetry between Experiments 1 and 2—whereby nouns that appear near the midpoint of the scale in Experiment 1 are revealed to have stronger associations with men in Experiment 2—mirrors a well-documented tendency for masculine categories to function as the default. Prior research on masculine generics suggests that masculine forms are often interpreted as gender-neutral while still preferentially activating male referents (e.g., Gygax et al. 2008; Sczesny et al. 2016; Stahlberg et al. 2007), reflecting a broader asymmetry in which terms associated with men are treated as the unmarked category. Many of the nouns showing this pattern are gender-neutral occupational terms (e.g., *salesperson*, *chairperson*, *firefighter*, *actor*, *waiter*). While these nouns are widely accepted as referring to individuals of any gender, they may still carry residual male associations (e.g., *salesperson*, *chairperson*, and *firefighter* derive from *salesman*, *chairman*, and *fireman*).

Another factor that may contribute to the observed asymmetries is the composition of the stimulus set itself: there were more nouns previously identified as male-biased than female-biased. Of the 406 nouns in our stimulus set, 260 (64%) were previously identified as male-biased, 102 (25%) as female-biased, and the remaining 44 (11%) as relatively neutral. This imbalance mirrors broader patterns in English, particularly for occupational nouns, where male-biased forms are more frequent (e.g., Ali et al. 2023), and many nouns intended to be neutral still carry male associations (e.g., Gygax et al. 2008). Thus, part of the asymmetry observed in Experiment 2—that there were more nouns in the [1,2) range on the *woman* scale—likely reflects the underlying distribution of gender associations in the stimulus set. At the same time, as discussed above, the two-scale method reveals additional asymmetries that are not captured by a single-scale approach, suggesting that the observed pattern reflects both the underlying distribution of items and the increased sensitivity of the two-scale method to asymmetric gender associations.

4. Comparing Experiment 1 and Experiment 2 Results. In order to compare our new two-scale method with the binary 7-point Likert scale method, we converted the two-scale results to the 7-point scale. This calculation was done by subtracting the likelihood to be a woman ratings from the likelihood to be a man ratings and adding 4. For example, in Experiment 2a, the

word *babysitter* received a mean rating of 1 on the *likelihood to be a man* scale and 2.24 on the *likelihood to be a woman* scale. To convert this to the 7-point rating, we calculate 1 - 2.24 + 4, which results in a rating of 2.76. For comparison, note that in Experiment 1a, *babysitter* received a mean rating of 2.48 on the 7-point scale. It is important to note that this conversion is done for scale comparison purposes only—the two-scale method is not intended to be converted into a binary scale when used by researchers for stimuli creation purposes. To assess whether the different rating experiments produced comparable patterns of gender associations, we calculated the Pearson correlation between item-level mean ratings obtained from each method. The results revealed extremely strong positive correlations between the two methods for both noun ($r = 0.985$, $p < .001$) and adjective ($r = 0.954$, $p < .001$) ratings. In other words, items rated as more strongly associated with men or women under the binary scale tended to receive correspondingly higher ratings on the independent man and woman scales. While the correlation for adjectives is slightly lower than nouns, it nevertheless remains very strong, suggesting that the two-scale method generalizes well across lexical categories. Figure 6 shows the rating correlations for the two different norming experiments. These findings indicate that the two-scale method captures highly correlated overall patterns of gender associations compared to the traditional binary Likert scale. At the same time, as discussed above, the two-scale method reveals asymmetries that are less apparent in the single-scale approach, suggesting that the two methods are broadly comparable while differing in the level of detail they provide.



(a) Correlation for Noun Ratings

(b) Correlation for Adjective Ratings.

Figure 6. Correlation between mean gender ratings obtained using the binary Likert scale (Experiment 1) and the two-scale method (Experiment 2). Each point represents an item, illustrating a strong overall correlation between the two methods.

5. Discussion. A substantial body of psycholinguistic research has examined gender in language, often relying on norming studies to quantify gender associations of words. However, this work has largely adopted a common methodological approach: representing gender along a single binary continuum. In the present study, we identified several limitations of this approach and evaluated an alternative method designed to capture gender associations without using the 7-point Likert scale.

The present study evaluates a two-scale method for norming gender associations that avoids

representing gender as a single binary continuum. Across experiments, ratings obtained using the two-scale method were strongly correlated with those obtained using a traditional binary Likert scale, suggesting that the proposed method produces comparable results while allowing for more flexible representations of gender associations. At the same time, the two-scale method shows potential to reveal asymmetries that were less apparent in the binary scale. In Experiment 2, some items that appeared relatively neutral on the binary scale of Experiment 1 showed asymmetric ratings when evaluated independently for likelihood of referring to a woman versus a man. In particular, several items received relatively neutral ratings for likelihood of referring to a woman while receiving moderately higher ratings for likelihood of referring to a man. These findings suggest that neutrality on the binary scale may obscure asymmetric gender associations that emerge when ratings are collected independently.

In our experiments, participants were also provided with a transparent debrief explaining the goals of the study and the motivation for evaluating alternative norming methods. This approach emphasized that participant responses contribute to improving measurement practices for gender associations in language. Including this information promotes transparency and acknowledges the role of participants in the development of new methodological approaches, which is particularly important when studying socially meaningful constructs.

In addition to nouns, the present study extended gender norming to adjectives, which have received comparatively little attention in prior psycholinguistic, and relatedly, norming work. Adjectives exhibited gender associations similar to those observed for nouns, suggesting that gendered expectations extend beyond nouns in English. This finding expands the range of stimuli available for experimental research on gender expectations in English and enables new lines of inquiry into related psycholinguistic questions, such as the interaction between nouns and adjectives (see Davenport & Ronai 2025).

While the two-scale method improves upon traditional binary Likert scales, there are also opportunities for further methodological development. For example, although the present approach separates ratings for women and men, it still relies on binary gender categories. Future work may therefore test whether the addition of a third scale assessing the likelihood of association with a nonbinary person reveals gender biases that may be overlooked by current rating systems. Expanding independent rating scales in this way may allow researchers to capture more nuanced patterns of gender associations in language. Additionally, future studies may examine how different scale labels, instructions, or participant populations influence the interpretation of independent gender ratings.

Overall, the present findings suggest that independent gender rating scales provide a flexible and informative alternative to traditional binary norming methods. The two-scale approach allows researchers to detect possible asymmetric gender associations, reduces assumptions about gender as a binary concept, and produces results comparable to traditional gender norming methods. More broadly, this work demonstrates that relatively small methodological adjustments, such as separating gender ratings into independent scales and refining scale design, can meaningfully improve the measurement of gender associations as well as participant experience in psycholinguistic research.

References

Adams, Sophie, Sheree Bekker, Yanan Fan, Tess Gordon, Laura J. Shepherd, Eve Slavich & David Waters. 2022. Gender bias in student evaluations of teaching: punish[ing] those who

fail to do their gender right. *Higher Education* 83(4). https://doi.org/10.1007/s10734-021-00704-9.

Ali, Dr. Mansoor, Dr. Abdus Samad & Mr. Tariq Amin. 2023. Gender discrimination in job titles in english language: A corpus-based critical study. *sjesr* 6(3). 1323. https://doi.org/10.36902/sjesr-vol6-iss3-2023(13-23).

Barber, Horacio & Manuel Carreiras. 2005. Grammatical gender and number agreement in spanish: An erp comparison. *Journal of Cognitive Neuroscience* 17(1). https://doi.org/10.1162/0898929052880101.

Bedin, Cooper, Montreal Benesch, Marina Zhukova & Lal Zimman. 2024. Current norms and best practices for collecting and representing sex/gender in linguistics: Towards ethical and inclusive methodologies. *Proceedings of the Linguistic Society of America* 9(1). https://doi.org/10.3765/plsa.v9i1.5668.

Bem, Sandra L. 1974. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology* 42(2). https://doi.org/https://doi.org/10.1037/h0036215.

Butler, Judith. 1999. *Gender trouble: feminism and the subversion of identity*. New York: Routledge 10th edn.

Cameron, Jessica J & Danu Anthony Stinson. 2019. Gender (mis)measurement: Guidelines for respecting gender diversity in psychological research. *Social and Personality Psychology Compass* 13(11). https://doi.org/10.1111/spc3.12506.

Carreiras, Manuel, Alan Garnham, Jane Oakhill & Kate Cain. 1996. The use of stereotypical gender information in constructing a mental model: Evidence from english and spanish. *The Quarterly Journal of Experimental Psychology Section A* 49(3). https://doi.org/10.1080/713755647.

Davenport, Cassie & Eszter Ronai. 2025. Pretty plumbers to the rescue: Adjectives aid in gender mismatch recovery. Poster presented at the 38th Annual Conference on Human Sentence Processing (HSP 38). March 27–29.

Dillon, Brian, Alan Mishler, Shayne Sloggett & Colin Phillips. 2013. Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language* 69(2). https://doi.org/10.1016/j.jml.2013.04.003.

Duffy, Susan A. & Jessica A. Keir. 2004. Violating stereotypes: Eye movements and comprehension processes when text conflicts with world knowledge. *Memory Cognition* 32(4). https://doi.org/10.3758/BF03195846.

Evans, C. A., K. Adler, D. Yucalan & L. M. Schneider-Bentley. 2024. Gender patterns in engineering phd teaching assistant evaluations corroborate role congruity theory. *International Journal of STEM Education* 11(1). https://doi.org/10.1186/s40594-023-00460-5.

Gygax, Pascal, Ute Gabriel, Oriane Sarrasin, Jane Oakhill & Alan Garnham. 2008. Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. *Language and Cognitive Processes* 23(3). https://doi.org/10.1080/01690960701702035.

Hagoort, Peter & Colin Brown. 1999. Gender electrified: Erp evidence on the syntactic nature of gender processing. *Journal of Psycholinguistic Research* 28(6).

Hyde, Janet Shibley, Rebecca S. Bigler, Daphna Joel, Charlotte Chucky Tate & Sari M. Van Anders. 2019. The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist* 74(2). https://doi.org/10.1037/amp0000307.

Irmen, Lisa. 2007. Whats in a (role) name? formal and conceptual aspects of comprehending personal nouns. *Journal of Psycholinguistic Research* 36(6). https://doi.org/10.1007/s10936-007-9053-z.

Kennison, Shelia M. & Jessie L. Trofe. 2003. Comprehending pronouns: A role for word-specific gender stereotype information. *Journal of Psycholinguistic Research* 32(3). https://doi.org/10.1023/A:1023599719948.

Khan, Shawn, Abirami Kirubarajan, Tahmina Shamsheri, Adam Clayton & Geeta Mehta. 2023. Gender bias in reference letters for residency and academic medicine: a systematic review. *Postgraduate Medical Journal* 99(1170). https://doi.org/10.1136/postgradmedj-2021-140045.

Kost, Rhonda G., Joseph Andrews, Ranee Chatterjee, Alex C. Cheng, Lisa Connally, Ann Dozier, Carrie Dykes, Daniel Ford, Nancy S. Green, Caroline Jiang, Sana Khoury-Shakour, Sierra Lindo, Karen Marder, Liz Martinez, Adam Qureshi, Jamie Roberts & Natalie Schlesinger. 2025. What research participants say about their research experiences in empowering the participant voice: Outcomes and actionable data. *Journal of Clinical and Translational Science* 9(1). e43. https://doi.org/10.1017/cts.2025.3.

Kush, Dave & Brian Dillon. 2021. Principle b constrains the processing of cataphora: Evidence for syntactic and discourse predictions. *Journal of Memory and Language* 120. https://doi.org/10.1016/j.jml.2021.104254.

Kush, Dave, Jeffrey Lidz & Colin Phillips. 2015. Relation-sensitive retrieval: Evidence from bound variable pronouns. *Journal of Memory and Language* 82. https://doi.org/10.1016/j.jml.2015.02.003.

Moon, Rosamund. 2014. From gorgeous to grumpy: adjectives, age, and gender. *Gender and Language* 8(1). https://doi.org/10.1558/genl.v8i1.5.

Osterhout, Lee, Michael Bersick & Judith Mclaughlin. 1997. Brain potentials reflect violations of gender stereotypes. *Memory Cognition* 25(3). https://doi.org/10.3758/BF03211283.

Pyykkönen, Pirita, Jukka Hyönä & Roger P. G. van Gompel. 2010. Activating gender stereotypes during online spoken language processing: Evidence from visual world eye tracking. *Experimental Psychology* 57(2). https://doi.org/10.1027/1618-3169/a000016.

Raphael, Esaya Britto. 2023. Gendered representations in language: A corpus-based comparative study of adjective-noun collocations for marital relationships. *Theory and Practice in Language Studies* 13(5). https://doi.org/10.17507/tpls.1305.12.

Schudson, Zach C. & Thekla Morgenroth. 2022. Non-binary gender/sex identities. *Current Opinion in Psychology* 48. https://doi.org/10.1016/j.copsyc.2022.101499.

Scott, Graham G., Anne Keitel, Marc Becirspahic, Bo Yao & Sara C. Sereno. 2019. The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods* 51(3). https://doi.org/10.3758/s13428-018-1099-3.

Sczesny, Sabine, Magda Formanowicz & Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in Psychology* 7. https://doi.org/10.3389/fpsyg.2016.00025. http://journal.frontiersin.org/Article/10.3389/fpsyg.2016.00025/abstract.

Stahlberg, Dagmar, Friederike Braun, Lisa Irmen & Sabine Sczesny. 2007. Representation of the sexes in language. In Klaus Fiedler (ed.), *Social communication*, 163–187. New York: Psychology Press.

Sturt, P. 2003. The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language* 48(3). https://doi.org/10.1016/S0749-596X(02)00536-3.

Zehr, Jeremy & Florian Schwarz. 2018. Penncontroller for internet based experiments (ibex) https://doi.org/https://doi.org/10.17605/OSF.IO/MD832.

Zimman, Lal. 2014. The discursive construction of sex: Remaking and reclaiming the gendered body in talk about genitals among trans men. In Lal Zimman, Jenny Davis & Joshua Raclaw (eds.), *Queer excursions: Retheorizing binaries in language, gender, and sexuality*, Oxford University Press.