Proceedings of the LSA

Perception and social evaluation of gendered adjective-noun combinations in Mandarin

Jiyuan Zhou & Aini Li*

Abstract. Mandarin lacks grammatical gender, but gender meanings can still be conveyed through lexical stereotypes. This study examines how native Mandarin listeners socially evaluate gendered adjective-noun combinations in speech across various personality dimensions, and whether such evaluations are influenced by voice gender and by the disclosed AI nature of the voice. Seventy-seven native Mandarin speakers listened to sentences containing different gendered adjectivenoun combinations, produced in AI-generated female and male voices, and rated the speakers on friendliness, trustworthiness, fluency, openness, and education level. Results showed that gender-congruent combinations were rated more positively than incongruent ones, with the strongest penalty found when female-typical adjectives modified masculine nouns. Female-typical adjectives were also evaluated more favorably in a female voice. Explicit disclosure that the voice were AI-generated did not significantly affect listener ratings, but voices perceived as human were evaluated more positively than those perceived as artificial. These findings provide empirical evidence that, in a language without grammatical gender, social evaluation of gendered language can be shaped by lexical gender stereotypes, paralinguistic voice cues, and perceived humanness.

Keywords. gendered language, gender stereotypes, social evaluation, Mandarin Chinese, adjective-noun combinations, voice gender, AI-generated speech

1. Introduction. Language carries rich social information such as gender but how gender is expressed varies across languages. As Corbett (1991) notes, gender may be expressed through both grammatical and semantic systems. In many Indo-European languages, gender is morphologically encoded and overtly marked on the surface form of words. For instance, in Spanish, grammatical gender distinguishes feminine or masculine nouns, as in *machacho* 'boy' vs. *machacha* 'girl', and *nieto* 'grandson' vs. *nieta* 'granddaughter'. The suffix '-o' typically occurs when the word refers to 'male', while the meaning 'female' appears with the suffix '-a'. In this sense, different morphemes represent different gender meanings (Falk 1978). Because gender is repeatedly and straightforwardly signaled through morphology, it becomes noticeable on the part of listeners when gender agreement is violated. For instance, when a speaker produces a gender mismatch, such as *El faro es luminosa y alto* 'The (m) lighthouse (m) is bright (f) and high'– native listeners can easily tell that the sentence is ungrammatical and that the speaker is linguistically incompetent in this language (Barber et al. 2004; Roulet-Amiot & Jakubowicz 2006).

However, unlike inflectional languages, Mandarin Chinese does not encode gender through grammatical morphology or agreement, but primarily through lexical semantics (Farris 1988). This makes gender expression less transparent and more difficult to localize in any single formal marker. In Mandarin, gender is conveyed through explicit lexical meanings like kinship terms, as well as through implicit social associations carried by occupational nouns and descriptive adjectives (Chao 1956; Zhu & Liu 2020; Jiao & Luo 2021; Su et al. 2021). Since Mandarin adjec-

* Many thanks to the audience at the 2026 LSA Annual Meeting for helpful comments and feedback. Authors: Jiyuan Zhou, University of South Carolina (jiyuan.zhou@sc.edu) & Aini Li, City University of Hong Kong (ainili@city.edu.hk).

tives are grammatically free to modify any noun, the distributional patterns that emerge in usage reflect social expectations instead of syntactic constraints. For example, the combination such as *piàoliàng de bàba* 'beautiful father' is well-formed in grammar, even though it departs from conventional gender expectations, as the adjective *piàoliàng* typically modifies a feminine none instead of a masculine one. Mandarin listeners perceive such expressions as socially atypical or stereotypically incongruent rather than as morphological violations. This raises the question of whether listeners treat these gender-incongruent combinations as a signal of reduced speaker competence or as socially marked departures from conventional gender expectations. While previous studies has documented gendered associations in Mandarin adjectives and nouns, it remains unclear how such combinations affect listeners social evaluations of the speaker. Therefore, the primary goal of this study is to examine how native Mandarin listeners socially evaluate adjective-noun combinations that vary in gender congruence.

However, in spoken language, this evaluation is unlikely depend on lexical content alone, because listeners also infer social information from the speaker's voice. Voice gender is among the most salient of these paralinguistic cues. Listeners perceive speaker gender from acoustic properties of the voice rapidly and with high accuracy (Zimman 2018), and this perception shapes social judgments that go beyond the propositional content of what is said (Ko et al. 2009). How listeners evaluate a gender-incongruent adjective-noun combination may therefore depend not only on the lexical mismatch itself but on whether the voice producing it is perceived as male or female. More recently, the increasing use of Artificial Intelligence (AI)-generated text-to-speech (TTS) voices in everyday contexts raises a further issue. Whether listeners know that a voice is artificially generated or perceive it to be artificial may itself shape how they evaluate the speaker. Prior work suggested that social evaluations of TTS voices across different trait dimensions are driven primarily by how human-like the voice sounds rather than by explicit knowledge of its origin (Lilley et al. 2025; Cohn & Zellou 2020), but whether this dissociation between perception and knowledge holds in the context of gendered language evaluation has not been tested. Thus, we extended our investigation to examine whether listeners' evaluations of gendered adjective-noun combinations are modulated by voice gender and the disclosed AI nature of the voice.

In this context, the present study examines how native Mandarin listeners perceive and socially evaluate gendered adjective-noun combinations in use. In an evaluation experiment, participants heard auditory sentences containing different gendered adjective-noun combinations spoken in AI-generated female and male voices, either with or without disclosure of the nature of the voice (human vs. AI), and rated the voice along different personality traits. The study is guided by three research questions:

RQ1. How do native Mandarin listeners socially evaluated gender-congruent and gender-incongruent adjective-noun combinations across personality dimensions?

RQ2. Does the voice gender modulate listeners' evaluations of these adjective-noun combinations?

RQ3. Does explicit disclosure that a voice is AI-generated affect listeners' evaluations?

2. Background.

2.1. GENDER ENCODING IN MANDARIN. Gender is encoded differently across languages, sometimes through grammatical systems and sometimes through semantic ones (Corbett 1991). In languages with grammatical gender, such as Spanish and Russian, gender is overtly marked on nouns and their agreeing elements, including adjectives, verbs, and personal pronouns. In

Russian, for instance, the adjective *novyj* 'new' appears as *novyj dom* 'new house' (masculine), *novaja gazeta* 'new newspaper' (feminine), or *novoe taksi* 'new taxi' (neuter), depending on the noun it modifies. Because such systems encode gender repeatedly and overtly, violations of agreement are immediately perceptible to native listeners and can be taken as evidence of language incompetence and social inappropriateness (Barber et al. 2004; Roulet-Amiot & Jakubowicz 2006). However, Mandarin Chinese does not encode gender through grammatical agreement or inflectional morphology in the same way. Instead, gender in Mandarin is often expressed through lexical meaning and social association. As Farris (1988) argued, Chinese exhibits a covert gender system in which gender emerges through patterned relations, contextual interpretation, and socially shared meaning. Thus, gender in Mandarin is better understood as a semantic and sociocultural category than as a grammatical one.

Within Mandarin, lexical gender can be expressed both explicitly and implicitly. Some nouns encode gender directly as part of their semantic content. For example, kinship terms such as *māma* 'mother' and *bàba* 'father' denote biological sex directly, and the third-person pronouns *tā* 'she' and *tā* 'he' refer to different genders, but they are phonetically identical and are distinguished only in writing (Chao 1956). Other nouns, however, are gender-neutral on the surface but carry covert masculine or feminine associations through cultural convention. Farris (1988) illustrated this with *yīshēng* 'doctor, which does not overtly mark masculine gender, yet it defaults to a male referent, such that intended feminine reference requires the *nǚ* 'female' prefix, yielding *nǚ* yīshēng 'female doctor'. The feminine is the marked case, and the masculine serves as the generic, especially when used with occupational nouns. Su et al. (2021) provided corpus evidence for this pattern by tracking the frequency of gendered modifiers before 63 occupational terms in *People's Daily* from 1946 to 2018. They found that *nǚ yīshēng* 'female doctor' is common, while *nán yīshēng* 'male doctor' is unattested. By contrast, *nán bǎomǔ* 'male nanny' appears regularly, whereas *nǚ bǎomǔ* 'female nanny' barely occurs. Similar default feminine associations have also been noted for occupations such as *hùshi* 'nurse', *kèfú* 'customer service' and *wénmì* 'secretary' (Jiang et al. 2023).

In addition to nouns, adjectives also implies masculine or feminine. Zhu & Liu (2020) asked participants to score 466 adjectives in terms of whether they were likely to describe males or females and found gender skewness in adjective meanings. Adjectives such as *piàoliàng* 'beautiful' and *róuruò* 'delicate' are associated with women, whereas adjectives such as *shuàiqì* 'handsome' and *zhuāngshì* 'sturdy' are associated with men. Using word embeddings, Li et al. (2022) further showed that Chinese adjectives reflect mainstream social judgments, with men more strongly associated with action and rationality and women with appearance and emotionality. Jiao & Luo (2021) similarly argued that gender stereotypes are conveyed through adjective choices and that repeated associations between an adjective and a particular gender can help sustain social stereotypes over time. When such items co-occur, they create expectations about which adjective-noun combinations are socially typical or atypical. Parallel evidence from English corpora points in the same direction. Adjectives such as *curvy* and *pretty* collocate strongly with female-denoting nouns, while adjectives such as *burly* and *muscular* collocate strongly with male-denoting nouns (Pearce 2008; Caldas-Coulthard & Moon 2010). These findings trigger our curiosity about how people perceive gender adjective-noun combinations today, especially when such combinations align with or depart from conventional gender expectations.

2.2. SOCIAL EVALUATION OF GENDERED LANGUAGE. Gendered adjective-noun combinations do not merely reflect how men and women are perceived; they may also evoke expectations about how they should be. Gender stereotypes originate in the social roles that men and women typically occupy, and the traits associated with each gender come to be seen as typical and appropriate (Eagly & Karau 2002). Prentice & Carranza (2002) formalized this distinction between descriptive stereotypes, which capture what members of a group are like, and prescriptive stereotypes, which specify what they ought to be like. Violations of prescriptive norms carry negative social evaluation, but violations of descriptive expectations do not. Rudman et al. (2012) tested this asymmetry by asking participants to rate gender-stereotyped trait combinations on desirability. Gender-congruent combinations such as *ambitious men* and *likeable women* received consistently higher ratings than incongruent ones, such as *aggressive women* or *emotional men*. This pattern holds even when the trait itself is positively valued in other contexts. Specifically, *ambitious* is generally desirable, but an *ambitious woman* is evaluated less favorably than an *ambitious man* because the combination violates prescriptive expectations about femininity.

The penalties for norm violation are not symmetric across genders. As *precarious manhood theory* proposes, unlike femininity, masculinity is often seen as a social status that must be continually demonstrated rather than taken for granted, which is why threats to it tend to trigger stronger reactions by men (Vandello & Bosson 2013). Bosson et al. (2022) provided further cross-cultural evidence for this asymmetry: proscriptions against male weakness appear near-universal across 62 countries, even as prescriptions about desirable male traits vary. Overall, masculine norm violations are treated more harshly than feminine ones. Given this asymmetry, a further question is whether different types of gender-incongruent adjective-noun combinations may not be evaluated in the same way.

Meanwhile, any such asymmetry in Mandarin cannot be assumed to be fixed, because the gender meanings linked to adjectives, nouns and even the interpretation of particular referents are themselves socially shaped and open to change. In contemporary Chinese society, women's participation in education, the labor force, and public life has expanded substantially over recent decades (Yang 2020; Ji et al. 2017), and the gender associations of occupational nouns have shifted accordingly (Su et al. 2021). However, such changes do not eliminate implicit gender bias. Using a dictation survey, Dong et al. (2023) examined how Mandarin speakers interpreted the referent of the epicene pronoun *tā*, a form that is not distinguished by gender in speech but is distinguished in writing. They found that referent interpretation remained strongly shaped by male-as-norm expectations and social stereotypes, and that these results varied across listener groups by gender and age. This suggests that gendered expectations in Mandarin are neither absent nor fixed, but remain active, socially mediated, and variable across contexts and populations. If such expectations already influence the interpretation of an underspecified referent, they are likely to matter even more when speakers hear adjectives that explicitly describe personality or social traits. Crucially, research on personality-trait adjectives further finds that gendered adjectives distribute differently across evaluative domains in corpus data. Drawing on the Big Five model, Motschenbacher & Roivainen (2020) classify personality adjectives into five categories: extraversion, agreeableness, conscientiousness, neuroticism, and intellect/openness. They further found that gender-consistent and cross-gender usages coexist across trait domains, indicating that the relationship between these adjectives and gender cannot be adequately explained by a single stereotype. This point motivates our study to explore whether gendered adjective-noun combinations whose social meanings are subject to change invite differentiated social evaluations.

2.3. VOICE GENDER AND AI-GENERATED SPEECH. In spoken language, an utterance carries not only what is said but information about who is saying it. Prior research has treated the voice as a salient cue to gender perception and has highlighted the role of both acoustic properties and sociocultural practice in shaping how voices are gendered (Zimman 2018). Importantly, such perception can shape social judgments independently of propositional content. Ko et al. (2009) found that female voice influences ratings of competence and warmth even when individuating information about the speaker is available, suggesting that paralinguistic cues operate alongside lexical content rather than being overridden by it. These effects are not limited to the voice in isolation. If voice gender shapes social evaluation independently of what is said, then the alignment or misalignment between voice gender and the gendered content of an utterance may itself modulate listener judgments. Thus, how listeners evaluate a gender-incongruent adjective-noun combination may depend not only on the lexical mismatch itself but on whether the voice producing it is perceived as male or female.

With the rapid development of speech synthesis technology, AI-generated text-to-speech voices have been increasingly common in everyday life. Beyond the voice gender, researchers have also begun to examine how such voices are perceived by listeners. Early work showed that neural TTS voices are rated as more human-like, natural, likeable, and familiar than earlier concatenative systems, suggesting that listeners' responses to synthesized speech are closely tied to perceived naturalness (Cohn & Zellou 2020). More recent research has extended this line of inquiry from general language attitudes to social evaluation. For example, Lilley et al. (2025) found that listeners attribute a range of social characteristics to TTS voices, including solidarity-related traits such as friendliness and honesty and status-related traits such as intelligence and wealth, indicating that social evaluation of synthetic voices is multidimensional rather than unitary. Relatedly, more robotic-sounding voices tended to receive less favorable evaluations across these dimensions (Lilley et al. 2025). Recent work on synthetic voice attractiveness further suggests that although listeners are often fooled into identifying TTS voices as human, synthetic voices are still rated as less attractive and less socially appealing overall, even as they increasingly approximate human voices in perception (Bruder et al. 2025).

As artificial voices become more naturalistic, listeners are more likely to encounter speech whose origin, whether human or AI-generated, is not always transparent. One issue that has drawn particular attention is the distinction between knowing that a voice is artificial and perceiving it as artificial. Reeves & Nass (1996) argued that people respond to mediated voices as they would to real people, regardless of what they consciously know about the medium. Nass & Moon (2000) provided further experimental evidence that social responses to computers and synthetic agents are driven primarily by perceptual cues rather than by reflective knowledge about the source. Consistent with this framework, Lilley et al. (2025) showed that labeling a voice as human or as a device influences some behavioral responses but not others, pointing to a dissociation between declarative knowledge and perceptual experience. If this dissociation extends to gendered language evaluation, then the effect of AI voice disclosure on listener judgments may depend not on whether listeners have been informed of a voice's origin but on whether they independently perceive it as artificial. The present study examines this possibility.

3. Methods. The goal of this experiment is to investigate how native Mandarin listeners evaluate gendered adjective-noun combinations along various social dimensions and whether such evaluations are modulated by the gender of the speaker voice. Specifically, we ask whether the gender

typicality of adjective-noun combinations, that is, whether the adjective's gender associations align with or violate the gender implied by the noun, influences listeners' judgment of speaker personality across dimensions such as friendliness, trustworthiness, fluency, open-mindedness, and education level. We further examine how voice gender interacts with adjective-noun gender typicality in shaping listeners' social evaluations, and whether knowing that the voice was AI-generated altered listeners' evaluations.

3.1. DESIGN. The experiment employed a mixed factorial design. In the within-subjects portion, the stimuli consisted of declarative sentences with adjective-noun combinations. The critical items were manipulated across three factors: adjective gender typicality (female-typical and male-typical), noun gender (feminine and masculine), and voice gender (female and male), yielding a $2 \times 2 \times 2$ repeated-measures design. In the between-subjects portion, AI-generated voice disclosure was manipulated. Participants were randomly assigned either to an AI-informed group, in which they were explicitly told that the voice was AI-generated, or to an uninformed group, in which no such disclosure was provided.

3.2. MATERIALS. The stimuli comprised 16 critical declarative sentences, each including an adjective-noun combination. To create these adjective-noun combinations, two types of adjectives were selected: 1) female-typical adjectives, which co-occur primarily with feminine nouns (e.g., *wǔmèi* 'sultry', *wēnwǎn* 'gentle'), and 2) male-typical adjectives, which co-occur primarily with masculine nouns (e.g., *yīngjùn* 'handsome', *gāngjiàn* 'sturdy'). Apart from the adjectives, the nouns used in the experiment were drawn from two gender categories: feminine nouns and masculine nouns. Feminine nouns included occupational and kinship terms conventionally associated with female referents (e.g., *hùshi* 'nurse', *āyí* 'aunt', *bǎomǔ* 'nanny'). Masculine nouns included occupational and kinship terms conventionally associated with male referents (e.g., *yīshēng* 'doctor', *shūshu* 'uncle', *bǎoān* 'security guard').

Each critical sentence followed a uniform syntactic structure: a gender-neutral proper name subject, followed by a copula, a numeral classifier, and an adjective-noun combination marked by the modification particle *DE* (attributive marker). Common human classifiers, such as *yīwèi*, *yīmíng*, and *yīgè*, all translatable as 'one person', were used consistently across stimuli to avoid classifier reduplication. In Mandarin, personal names are often perceived as gendered by virtue of the characters used (e.g., *juān* is typically associated with female names, while *qiáng* with male names). Thus, names with no identifiable gender associations were chosen to prevent participants from inferring the gender of the subject in a given sentence from the auditory stimulus alone. Example stimuli are provided in Table 1.

In addition to the 16 critical items, 16 filler sentences were constructed using various sentence structures with non-adjectival modifiers or non-human noun combinations to keep the task diverse and unpredictable. The full stimulus set thus comprised 32 sentences.

All sentences were converted to audio using a Microsoft server-based TTS system [1]. Audio files were produced in 16 kHz, 16-bit, mono PCM WAV format. Two AI voices were used: a female voice generated using the model Xiao Chen, labeled by the provider as a 'young female' voice, and a male voice generated using the model Yun Yi, labeled as a 'young male' voice. Two native speakers of Mandarin further confirmed that both voices were representative of prototypical female and male vocal characteristics in Mandarin. Finally, 32 sentences were rendered in

[1] *Text To Speech*, a web-based text-to-speech platform used to generate the audio stimuli for this study, accessed March 30, 2026, https://www.text-to-speech.cn/.

Sentence	Adjective gender	Noun gender
Chén Xīng shì yí wèi wūmèi de āyí.		
'Chen Xing is a sultry aunt.'	Female-typical	Feminine
Féng Nán shì yíwèi xiánhuì de jiùjiu.		
'Feng Nan is a virtuous uncle.'	Female-typical	Masculine
Wáng Nuè shì yìmíng yīngjùn de yīshēng.		
'Wang Nuo is a handsome doctor.'	Male-typical	Masculine
Wú Qí shì yíwèi shuàiqì de yòushī.		
'Wu Qi is a handsome preschool teacher.'	Male-typical	Feminine

Table 1. Example stimulus sentences classified by adjective gender and noun gender

both voices.

3.3. PARTICIPANTS. A total of 77 native Mandarin Chinese speakers participated in this experiment (45 female, 31 male, 1 participant did not report gender). All participants were recruited from mainland China and were compensated with RMB 5 for their time. Their ages ranged from 21 to 35, and their education backgrounds ranged from bachelor's to doctoral level. No one reported any hearing deficit at the time of participation.

Participants were randomly assigned to one of two between-subjects conditions differing in AI-generated voice disclosure: an AI-informed group (N = 48; 27 female, 21 male) in which participants were explicitly told that the voices they would hear were AI-generated, and an uninformed group (N = 29; 18 female, 10 male, 1 did not report gender) in which no such disclosure was provided.

3.4. PROCEDURE. The evaluation task was implemented online via Qualtrics and conducted entirely in Mandarin Chinese. Before the task started, participants provided informed consent. The between-subjects manipulation of AI disclosure was separated at the instruction stage. In the AI-informed condition, participants were explicitly told that the sentences they would hear were AI-generated. In the uninformed condition, participants were told nothing. After the rating task, they were additionally asked to indicate whether they believed the voice they heard was female or male and whether they thought the recordings were produced by a human speaker or generated by AI. Except for this difference, all instructions and task steps were identical across the two conditions.

During the evaluation task. Participants heard the 32 audios (in a randomized order) and rated the voice on five social-personality dimensions (friendliness, trustworthiness, fluency, open-mindedness, and education level) using a 5-point Likert scale (1 = very unfriendly, 5 = very friendly). These dimensions were selected based on a norming study in which five individuals provided judgments about people who either used or were described by the gender-incongruent adjective-noun combinations.

After the evaluation task, participants reported their age, gender, and education level in a demographic questionnaire. Participants in the uninformed condition were additionally asked to indicate (a) whether they perceived the voice they heard as female or male, and (b) whether they believed the recordings were produced by a human or generated by AI. It took participants 10 minutes on average to complete the entire experiment.

4. Analysis and Results. All ratings were normalized within each participant. Because exploratory analysis showed that rating patterns were broadly similar across the five personality dimensions, we computed a composite evaluation score for each trial by averaging the five personality ratings. The higher the ratings, the more positive the evaluations were. This mean score was used as the dependent variable in a linear mixed-effects regression model, fitted with the *lme4* package (version 1.1-36; Bates et al. 2015). Plots were created using *ggplot2* (version 3.5.1; Wickham 2016) in *R* Statistical environment version 4.4.3 R Core Team 2013). Fixed effects of this model included adjective gender, noun gender, voice gender, their three-way interaction and the between-subject AI disclosure condition (AI-informed vs. uninformed). Participant age, gender, and education level were added as covariates. Random intercepts were specified for participants and stimuli to account for repeated measurements across listeners and items. All categorical predictors were sum-coded so that model coefficients reflect deviations from the grand mean.

To examine whether this overall pattern varied across traits, we also inspected the ratings for each five personality dimensions separately.

4.1. ADJECTIVE-NOUN GENDER AND VOICE GENDER. Figure 1 shows the mean ratings across various adjective-noun gender combinations, separated by voice gender. As can be seen, under the female-voice condition, sentences containing gender-congruent adjective-noun combinations (i.e., female-typical adjectives with feminine nouns, male-typical adjectives with masculine nouns) received higher ratings than those containing gender-atypical combinations (i.e., female-typical adjectives with masculine nouns, male-typical adjectives with feminine nouns). Specifically, feminine nouns were rated even higher when modified by female-typical adjectives (M = 3.78, SD = 0.89) than by male-typical adjectives (M = 3.53, SD = 0.93), whereas masculine nouns received lower ratings with female-typical adjectives (M = 3.38, SD = 1.00) than with male-typical adjectives (M = 3.53, SD = 0.96). Under the male-voice condition, however, a somewhat different pattern emerged. Sentences containing male-typical adjectives received relatively high ratings regardless of noun gender. Feminine nouns were rated more favorably with male-typical adjectives (M = 3.64, SD = 0.85) than with female-typical adjectives (M = 3.36, SD = 1.03), and a similar advantage for male-typical adjectives was observed for masculine nouns (M = 3.60, SD = 0.94 vs. M = 3.19, SD = 1.13). Across both voice conditions, there was a consistent pattern: the combination of female-typical adjectives modifying masculine nouns consistently received the lowest mean ratings (female voice: M = 3.38; male voice: M = 3.19).

The linear mixed-effects regression model further confirmed these patterns. The results are summarized in Table 2. Three significant main effects were identified. First, there was a main effect of adjective gender typicality: overall, sentences containing female-typical adjectives received lower ratings than those containing male-typical adjectives ($\beta = -0.07$, $p < .05$), indicating that ascribing stereotypically feminine traits to a referent (e.g., *wēnwán* 'gentle', *wūmèi* 'sultry') led to lower evaluations relative to ascribing stereotypically masculine traits (e.g., *yīngjùn* 'handsome', *gāngjiàn* 'sturdy'), holding noun gender and voice gender constant. Second, there was a main effect of noun gender: sentences with feminine nouns (e.g., *hùshi* 'nurse', *bāomû* 'nanny') received higher ratings than those with masculine nouns ($\beta = 0.08$, $p < .01$), suggesting that referents denoted by feminine-gender nouns were evaluated more positively in general. Third, there was a main effect of voice gender: sentences delivered in a female AI voice received higher mean ratings than those delivered in a male AI voice ($\beta = 0.05$, $p < .05$) regardless of adjectives and nouns. This indicates that listeners prefer a female AI voice to a male AI voice.

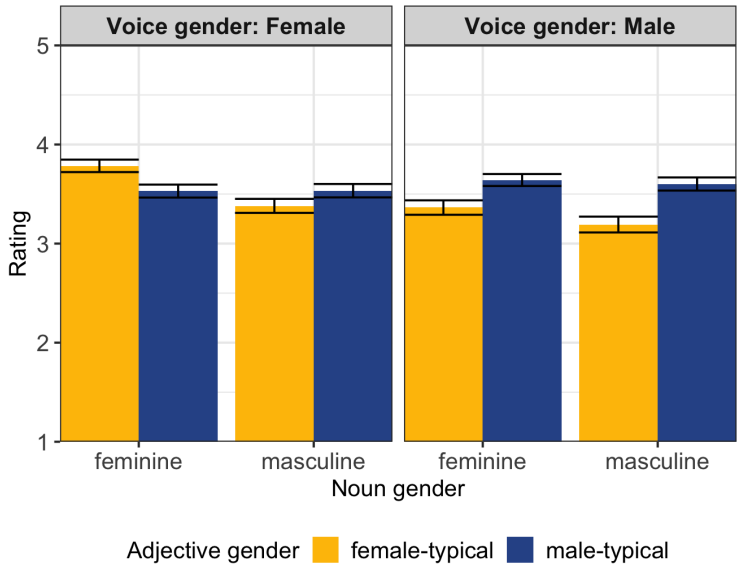

Figure 1. Mean ratings by voice gender, adjective gender, and noun gender

No significant main effects were found for the AI disclosure condition or for participant-level covariates (age, gender, and education level).

Two significant two-way interactions were identified. There was a significant interaction between adjective gender and noun gender ($\beta = 0.07, p < .05$): gender-typical combinations were rated highest, while gender-atypical combinations received the lowest ratings. Among atypical pairings, female-typical adjectives modifying masculine nouns consistently attracted the lowest evaluations (e.g., *Féng Nán shì yí wèi xiánhuì de jiùjiu*, 'Feng Nan is a virtuous uncle'; $\beta = -0.30, p < .0001$). Notably, male-typical adjectives modifying feminine nouns (e.g., *Zhōu Yún shì yíge cūlu de jiějie*, 'Zhou Yun is a rude sister') received ratings comparable to gender-typical male combinations (e.g., *yīngjùn de yīshēng*, 'handsome doctor'). This asymmetry reflects an unequal tolerance for gender-crossing: attributing masculine traits to female referents is more socially acceptable than attributing feminine traits to male referents.

There was also a significant interaction between adjective gender and voice gender ($\beta = 0.10, p < .01$): for sentences containing female-typical adjectives, ratings were significantly higher when produced in a female voice than in a male voice, suggesting that congruence between adjective gender and voice gender enhances social evaluations. No comparable modulation by voice gender was found for male-typical adjectives, where ratings were similar across female and male voices. The three-way interaction among adjective gender, noun gender, and voice gender did not reach statistical significance ($\beta = 0.03, p = .16$).

Figure 2 presents the descriptive patterns for each personality dimension separately (friendliness, trustworthiness, fluency, open-mindedness, education). The overall pattern was consistent across dimensions, although the magnitude of the differences varied somewhat by these dimension. Under female-voice conditions, ratings tended to track the gender typicality of the adjective-noun combination, whereas under male-voice conditions, ratings aligned more closely with adjective gender alone, with male-typical adjectives receiving higher scores regardless of noun gender. The differentiation between gender-typical and gender-atypical combinations ap-

Predictor	β	SE	t	p
(Intercept)	3.09	0.24	12.69	$< .001 **$
female-typical adjective	-0.07	0.02	-3.26	$< .05*$
feminine noun	0.08	0.02	3.40	$< .01 **$
female voice	0.05	0.02	2.39	$< .05*$
AI informed	0.04	0.06	0.68	.50
age [20–25]	0.17	0.17	1.02	.31
age [26–30]	-0.14	0.18	-0.77	.45
age [31–35]	0.39	0.31	1.26	.21
gender [female]	0.28	0.23	1.23	.22
gender [male]	0.21	0.22	0.97	.33
undergraduate	-0.13	0.15	-0.81	.42
graduate	0.07	0.12	0.61	.54
female-typical adjective × feminine noun	0.07	0.02	3.00	$< .05*$
female-typical adjective × female voice	0.10	0.02	4.39	$< .01 **$
feminine noun × female voice	0.02	0.02	1.05	.33
female-typical adjective × feminine noun × female voice	0.03	0.02	1.54	.16

Table 2. Linear mixed-effects regression results for composite social-evaluation ratings. Model: *ratings* ∼ *adjective gender* × *noun gender* × *voice gender* + *AI-disclosure* + *participant age* + *participant gender* + *participant education level* + $(1 \mid \textit{participant}) + (1 \mid \textit{stimulus})$.

peared strongest for education and fluency, weaker for friendliness and open-mindedness, and somewhat smaller for trustworthiness.

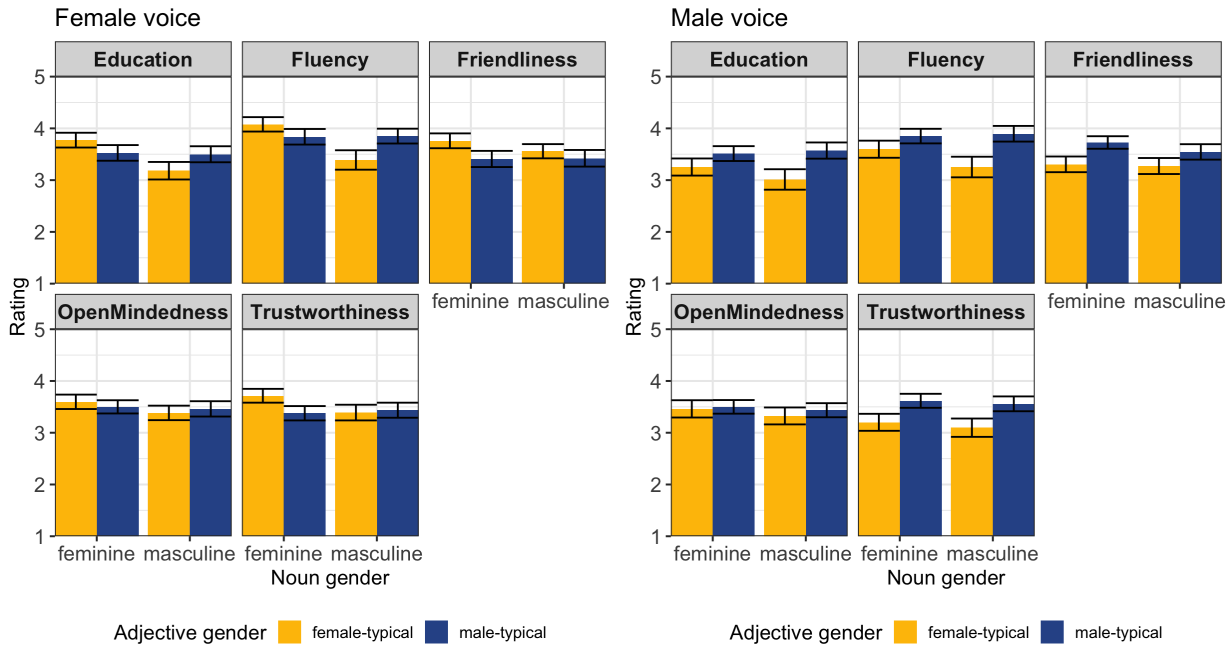

Figure 2. Evaluation ratings by personality dimension, adjective gender, noun gender, and voice gender

4.2. ROLE OF AI DISCLOSURE AND PERCEIVED VOICE ORIGIN. Based on the results shown in Table 2, the between-subject manipulation of AI disclosure (informed vs. uninformed) did not yield a statistically significant effect on composite social-evaluation ratings ($\beta = 0.04$, $p = .50$). We then looked at whether evaluations varied as a function of listeners own perceptual categorization of the voice.

Ratings within the uninformed condition varied depending on how listeners perceived the nature of the voice. We therefore further examined ratings according to whether listeners classified the voice as human, AI-generated, or were unsure of its origin (Figure 3). Voices that listeners subjectively classified as human received the highest composite evaluations ($\beta = 0.46$, $p < .05$), including cases in which the voice was in fact AI-generated but went undetected. By contrast, voices that listeners independently identified as AI-generated, without having been explicitly informed, received the lowest ratings ($\beta = -0.23$), reflecting a substantial penalty associated with the perceptual detection of artificial voice. Voices for which participants remained uncertain occupied an intermediate position ($\beta = -0.03$). Taken together, these results suggest that listeners' evaluations depended more on whether a voice was perceived as human or artificial than on whether its AI origin was explicitly disclosed.

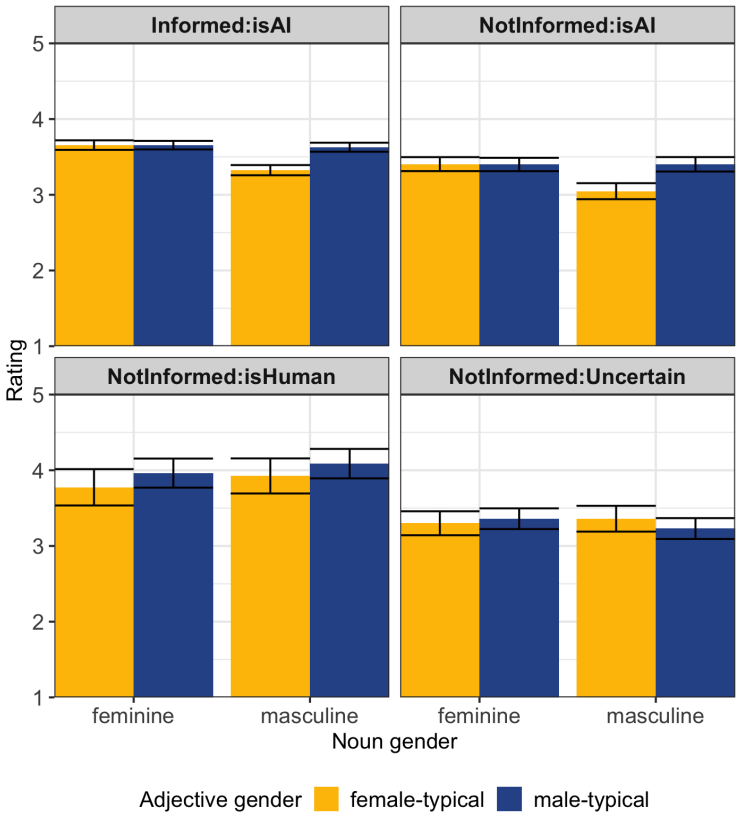

Figure 3. Mean ratings by AI disclosure condition

5. General discussion. This study investigated how native Mandarin listeners socially evaluate the use of declarative sentences including gendered adjective-noun combination and whether such evaluations are modulated by voice gender and AI voice disclosure. First, we found that

sentences with gender-congruent adjective-noun combinations were evaluated more favorably than gender-incongruent ones, with the least favored combination occurring when female-typical adjectives modify masculine nouns (e.g., *Féng Nán shì yíwèi xiánhuì de jiùjiu.* 'Feng Nan is a virtuous uncle.'). In addition, listeners overall tended to give more positive ratings when female-typical adjectives were produced in a female voice. Moreover, explicitly disclosing that a voice was AI-generated did not significantly affect listener evaluations, whereas listeners' expectations about whether the voice was human or artificial did. These findings together suggest that the social evaluation of adjective-noun combinations in Mandarin is shaped not only by lexical gender stereotypes but also by paralinguistic and perceptual cues. We discuss each of these findings specifically in the following subsections.

5.1. GENDER CONGRUENCY AND ASYMMETRIC EVALUATIONS. Our results found that gender-congruent adjective-noun combinations were evaluated more positively than gender-incongruent ones in spoken Mandarin. This suggests that even in a language without grammatical gender, listeners have implicit knowledge of gender stereotypes that are not overtly marked, and that these stereotypes can be socially evaluated. This congruence effect is consistent with previous work showing that Mandarin encodes gender through lexical semantics and distributional associations, including gendered patterns in role nouns and adjectival descriptions, and that information about gender stereotypes continues to shape gender inference during language perception (Farris 1988; Zhu & Liu 2020; Dong et al. 2023).

Importantly, the penalty for gender incongruence was asymmetric, with the lowest ratings observed when female-typical adjectives modified masculine nouns. However, male-typical adjectives modifying feminine nouns were penalized less strongly and, in some cases, were rated comparably to gender-congruent masculine combinations. This asymmetry suggests that Mandarin listeners were not merely responding to linguistic mismatch itself, but were also projecting social hierarchies and gender norms onto their evaluations of adjective-noun combinations. That is, violations of masculine norms incur more social costs than violations of feminine norms. Listeners therefore seem to evaluate not only whether a combination sounds natural, but also whether crossing gender norms is socially acceptable. This interpretation aligns with research on prescriptive gender stereotypes (Prentice & Carranza 2002), which holds that gender stereotypes function not only descriptively but also prescriptively, as well as with precarious manhood theory, according to which masculinity is a socially tenuous status that requires continual validation and is more harshly sanctioned when threatened (Vandello & Bosson 2013). A complementary explanation for the relatively weaker penalty for assigning male-typical descriptions to feminine referents is the ongoing transformation of gender roles in contemporary Chinese society, where women's more active participation in public life has broadened the range of traits that can be socially associated with them.

5.2. THE ROLE OF VOICE GENDER. Beyond the effects of lexical gender, the voice gender also played a role in shaping listeners' social evaluations of Mandarin adjective-noun combinations. Sentences produced in a female voice received higher ratings overall, and female-typical adjectives were evaluated more favorably when they were produced in a female voice than in a male voice. This indicates that listeners integrate lexical gender information with paralinguistic cues from the speaker's voice, and that congruence across adjective gender and voice gender enhances positive evaluations. In this sense, the social evaluation of these combinations was shaped not only by what was said, but also by how the gendered qualities of the speaker's voice aligned with

the lexical content.

However, the effect of voice gender was not uniform across all conditions, and the three-way interaction among adjective gender, noun gender, and voice gender was not significant. This indicates that listeners' evaluations were primarily shaped by the gender meanings of the adjective-noun combinations, while voice gender only intensified some of these meanings under certain conditions. That is, voice cues do matter in impression formation, but their effect depends on the specific semantic and social expectations triggered by the content of the utterance.

5.3. AI-DISCLOSURE AND PERCEIVED HUMANNESS. Finally, our results showed that explicit disclosure of AI voice origin did not significantly affect listeners social evaluations. However, ratings decreased when listeners subjectively perceived a voice as AI-generated, whereas voices that were in fact AI-generated but perceived as human received the most favorable evaluations. This difference between these two cases lies in whether the voice matched listeners' expectations of how it should sound, rather than in disclosure itself. When listeners were told in advance that a voice was artificial, that information alone did not necessarily change their evaluations if the voice still sounded natural. By contrast, when listeners spontaneously perceived artificiality, the voice appears to have violated their default expectation of human-like speech. The resulting penalty thus seems to follow from perceptual mismatch rather than from explicit knowledge.

This distinction has practical implications for how AI-generated speech is received. If the evaluative cost follows from violated perceptual expectations rather than from disclosed information, then disclosure on its own may do little to shift listener evaluations. As TTS technology improves and synthetic voices become increasingly difficult to distinguish from human voices, the gap between expectation and perception is likely to narrow. Evaluative penalties may therefore diminish, not because listeners necessarily become more accepting of synthetic speech, but because fewer voices sound artificial enough to trigger negative judgments(Lilley et al. 2025; Cohn & Zellou 2020; Reeves & Nass 1996).

More broadly, the present study highlights the theoretical importance of examining gendered language in Mandarin, a language without grammatical gender but with strong lexical gender associations. The findings show that gendered social evaluation can emerge even in the absence of overt grammatical marking, as listeners still rely on social expectations tied to adjective-noun combinations. Meanwhile, these evaluations are not determined by lexical meaning alone, but are jointly shaped by vocal gender cues in spoken interaction. Furthermore, in the context of AI-generated speech, what matters more is not simply whether listeners know that a voice is artificial, but whether it sounds human enough to sustain a socially credible impression. Taken together, the study suggests that gendered meaning in spoken Mandarin is constructed across multiple levels, including lexical association, paralinguistic cues, and perceived humanness.

References

Barber, Horacio, Elena Salillas & Manuel Carreiras. 2004. Gender or genders agreement? In *The on-line study of sentence comprehension*, 309–328. Psychology Press.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01.

Bosson, Jennifer K., Mariah Wilkerson, Natasza Kosakowska-Berezecka, Paweł Jurek & Michał

Olech. 2022. Harder won and easier lost? Testing the double standard in gender rules in 62 countries. *Sex Roles* 87(1–2). 1–19. https://doi.org/10.1007/s11199-022-01297-y.

Bruder, Camila, Pamela Breda & Pauline Larrouy-Maestri. 2025. Attractive synthetic voices. *Computers in Human Behavior: Artificial Humans* 100211.

Caldas-Coulthard, Carmen Rosa & Rosamund Moon. 2010. curvy, hunky, kinky: Using corpora as tools for critical analysis. *Discourse & Society* 21(2). 99–133. https://doi.org/10.1177/0957926509353843.

Chao, Yuen Ren. 1956. Chinese terms of address. *Language* 32(1). 217–241.

Cohn, Michelle & Georgia Zellou. 2020. Perception of concatenative vs. neural text-to-speech (tts): Differences in intelligibility in noise and language attitudes, https://doi.org/10.21437/Interspeech.2020-1336.

Corbett, Greville G. 1991. *Gender* Cambridge Textbooks in Linguistics. Cambridge University Press.

Dong, Ming, Rong Chen & Lin He. 2023. Gender bias in the chinese epicene pronoun ta. *Language Sciences* 97. 101543. https://doi.org/https://doi.org/10.1016/j.langsci.2023.101543.

Eagly, Alice H. & Steven J. Karau. 2002. Role congruity theory of prejudice toward female leaders. *Psychological Review* 109(3). 573–598. https://doi.org/10.1037//0033-295X.109.3.573.

Falk, Julia S. 1978. *Linguistics and language: A survey of basic concepts and implications*. New York: Wiley 2nd edn.

Farris, Catherine S. 1988. Gender and grammar in chinese: with implications for language universals 14(3). 277–308. https://doi.org/10.1177/009770048801400302.

Ji, Yingchun, Xiaogang Wu, Shengwei Sun & Guangye He. 2017. Unequal care, unequal work: toward a more comprehensive understanding of gender inequality in post-reform urban china. *Sex Roles* 77(11). 765–778. https://doi.org/10.1007/s11199-017-0751-1.

Jiang, Tingting, Yao Li, Shiting Fu & Ye Chen. 2023. Creating a chinese gender lexicon for detecting gendered wording in job advertisements. *Inf. Process. Manage.* 60(5). https://doi.org/10.1016/j.ipm.2023.103424.

Jiao, Meichun & Ziyang Luo. 2021. Gender bias hidden behind Chinese word embeddings: The case of Chinese adjectives. In Marta R. Costa-jussà, Hila Gonen, Christian Hardmeier & Kellie Webster (eds.), *Proceedings of the 3rd workshop on gender bias in natural language processing*, 8–15. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.gebnlp-1.2.

Ko, Sei Jin, Charles M. Judd & Diederik A. Stapel. 2009. Stereotyping based on voice in the presence of individuating information: Vocal femininity affects perceived competence but not warmth. *Personality and Social Psychology Bulletin* 35(2). 198–211. https://doi.org/10.1177/0146167208326477. PMID: 19141624.

Li, Jiali, Shucheng Zhu, Ying Liu & Pengyuan Liu. 2022. Analysis of gender bias in social perception and judgement using Chinese word embeddings. In Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky & Hila Gonen (eds.), *Proceedings of the 4th workshop on gender bias in natural language processing (gebnlp)*, 8–16. Seattle, Washington: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.gebnlp-1.2.

Lilley, Kevin D., Ellen Dossey, Michelle Cohn, Cynthia G. Clopper, Laura Wagner & Georgia Zellou. 2025. Social evaluation of text-to-speech voices by adults and children 166. 103163. https://doi.org/10.1016/j.specom.2024.103163.

Motschenbacher, Heiko & Eka Roivainen. 2020. Personality traits, adjectives and gender. *Journal of Language and Discrimination* 4(1). 16–50. https://doi.org/10.1558/jld.40370.

Nass, Clifford & Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 56(1). 81–103. https://doi.org/https://doi.org/10.1111/0022-4537.00153.

Pearce, Michael. 2008. Investigating the collocational behaviour of man and woman in the bnc using sketch engine. *Corpora* 3(1). 1–29. https://doi.org/10.3366/E174950320800004X.

Prentice, Deborah A. & Erica Carranza. 2002. What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly* 26(4). 269–281. https://doi.org/10.1111/1471-6402.t01-1-00066.

R Core Team. 2013. R: A language and environment for statistical computing. `http://www.R-project.org/`. Last viewed April 26, 2013.

Reeves, Byron & Clifford I. Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge: Cambridge University Press.

Roulet-Amiot, Leslie & Clia Jakubowicz. 2006. Production and perception of gender agreement in french sli. *Advances in Speech Language Pathology* 8(4). 335–346. https://doi.org/10.1080/14417040601009420.

Rudman, Laurie A., Corinne A. Moss-Racusin, Julie E. Phelan & Sanne Nauts. 2012. Status incongruity and backlash effects: defending the gender hierarchy motivates prejudice against female leaders 48(1). 165–179. https://doi.org/10.1016/j.jesp.2011.10.008.

Su, Qi, Pengyuan Liu, Wei Wei, Shucheng Zhu & Chu-Ren Huang. 2021. Occupational gender segregation and gendered language in a language without gender: Trends, variations, implications for social development in China. *Humanities and Social Sciences Communications* 8(1). 133. https://doi.org/10.1057/s41599-021-00799-6.

Vandello, Joseph A & Jennifer K Bosson. 2013. Hard won and easily lost: A review and synthesis of theory and research on precarious manhood. *Psychology of men & masculinity* 14(2). 101.

Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. `https://ggplot2.tidyverse.org`.

Yang, Juhua. 2020. Women in china moving forward: Progress, challenges and reflections. *Social Inclusion* 8(2). 23–35. https://doi.org/10.17645/si.v8i2.2690.

Zhu, Shucheng & Pengyuan Liu. 2020. Weida de nanren he juejiang de nvren: jiyu yuliaoku de xingrongci xingbie piandu yanjiu (great males and stubborn females: A diachronic study of corpus-based gendered skewness in Chinese adjectives). In Maosong Sun, Sujian Li, Yue Zhang & Yang Liu (eds.), *Proceedings of the 19th chinese national conference on computational linguistics*, 31–42. Haikou, China: Chinese Information Processing Society of China.

Zimman, Lal. 2018. Transgender voices: Insights on identity, embodiment, and the gender of the voice. *Language and Linguistics Compass* 12(8). e12284. https://doi.org/https://doi.org/10.1111/lnc3.12284. E12284 LNCO-0742.