Beyond Sentiment: Evaluating Gendered Language Using Zero-Shot Text Classification

Anonymous ACL Submission

Abstract

Situational Judgment Tests (SJTs) present hypothetical job-related situations to assess judgment and decision-making skills. Using zero-shot text classification, we replicated previously established findings of sentiment effects and further explored the influence of gendered language on participant responses in SJTs. Our study demonstrates that negative sentiment in action statements lowers effectiveness ratings and increases response variability. Contrary to gender schema theory, we found no evidence that gender-congruent phrasing led to higher effectiveness ratings. These findings underscore the potential of zero-shot text classification for refining SJT item development and mitigating unintended biases.

1 Introduction

A Situational Judgment Test (SJT) is used to assess how individuals handle realistic job-related scenarios, focusing on practical, interpersonal, and problem-solving skills. Participants might be presented with a scenario where a team member consistently misses deadlines. One possible action statement could be, "Confront the team member directly and demand an explanation." Participants would then rate the effectiveness of this action on a scale from 1 to 5. The use of SJTs in hiring and employee development is of great interest to researchers and practitioners (Christian et al., 2010; McDaniel et al., 2007; Patterson et al., 2017; Webster et al., 2020). Despite the use of SJTs, little is known about how gendered language impacts responses, which this study aims to investigate using zero-shot text classification.

Zero-shot text classification, a method where a model is inferenced to classify text without specific examples of each class in the training data, offers a novel approach to analyze the linguistic features of SJT statements. It can identify the sentiment and gendered language within statements, providing insights into how these elements influence responding behavior. This approach enables a deeper understanding of the nuanced impact of item design on participant responses, facilitating the development of more equitable and valid SJTs.

While our study focuses on SJTs, the implications of zero-shot text classification extend beyond this domain. This method can be applied to sentiment analysis, bias detection in recruitment tools, and content moderation on social media platforms. By leveraging pre-trained language models, researchers and practitioners can gain insights into the nuanced impact of language on human behavior and decision-making. Furthermore, this technique is especially advantageous because it eliminates the need for extensive labeled datasets, which can be resource-intensive to create.

2 Related Work

Previous research has explored how the wording and sentiment of statements affect how respondents evaluate effectiveness in SJTs (Loftus and Palmer, 1974; Kensinger and Corkin, 2003). More specifically, a recent study by Pawirosetiko and Perrotta (2024) examined the impact of SJT action statement sentiment on responding behavior using a zero-shot text classification approach. It was found that the sentiment of action statements, particularly the degree of negative sentiment, influenced both the average ratings of effectiveness and the variability in responding. These findings highlight the importance of carefully considering the language in SJT development, as the emotional tone of action statements can introduce biases.

Of primary interest to this study, gender differences in SJT performance have been well-documented. For instance, a meta-analysis by Whetzel et al. (2008) revealed significant sex differences in SJT performance, while other studies have also illustrated these differences in various

contexts (Arthur et al., 2014; Herde et al., 2020). These gender differences might be due to the wording of the test statements. Research suggests that men and women may use language differently and respond differently to the emotional tone and phrasing (Leaper and Ayres, 2007; Newman et al., 2008). Moreover, studies have found that the use of masculine or feminine phrasing can impact how individuals perceive and respond to various stimuli, including job descriptions and performance evaluations (Gaucher et al., 2011; Madera et al., 2009).

Gender schema theory suggests that people have mental frameworks about gender that influence how they perceive and interpret information Bem, 1981. These schemas categorize attributes as either masculine or feminine and guide individuals' understanding of what is considered appropriate behavior for each gender. That is, these masculine and feminine categorizations often align with societal stereotypes.

Collectively, these studies highlight the critical role of language in shaping SJT responses and underscore the need for careful item design to mitigate biases. As employee assessment methods evolve, understanding SJT design and its impact on responses is crucial. By employing zero-shot text classification, our research aims to show that such tools can be used to further elucidate how sentiment and gendered language in SJTs impact respondent behavior, ultimately contributing to the development of more equitable assessment tools.

3 Methodology

3.1 Hypotheses

Building upon previous work, we aimed to replicate findings on how the sentiment of action statements affects responses (Pawirosetiko and Perrotta, 2024):

- **H1**: Participants will exhibit significantly lower mean responding behavior and higher variability in responding behavior when the text of the action statement is more negative.

Gender schemas act as filters, making individuals more likely to accept information that matches their gender expectations (Bem, 1981). Thus, we sought to explore the role that gendered phrasing may have in SJT mean responding behavior by examining potential sex differences:

- **H2**: Male participants will exhibit significantly higher mean responding behavior when the text of the action statement is phrased in a masculine way.

- **H3**: Female participants will exhibit significantly higher mean responding behavior when the text of the action statement is phrased in a feminine way.

It follows that encountering information that is incongruent with one's gender schema may lead to cognitive dissonance or uncertainty (Festinger, 1957). This psychological discomfort arises from the conflict between their internalized gender expectations and the information presented in the SJT scenario. To resolve this discomfort, individuals may react in various ways, such as adjusting their evaluations, rejecting the information, or seeking additional cues to make sense of the situation. This variability in coping mechanisms was expected to result in a wider range of responses compared to situations where the information aligns with their gender schemas:

- **H4**: Male participants will exhibit significantly higher variability in responding behavior when the text of the action statement is phrased in a feminine way.

- **H5**: Female participants will exhibit significantly higher variability in responding behavior when the text of the action statement is phrased in a masculine way.

3.2 Data Collection and Analysis

We conducted a follow-up study replicating Pawirosetiko and Perrotta (2024) using a larger, diverse sample of SJT action statements. We generated normative data based on effectiveness ratings, using sample sizes ranging from 243 to 530 (M = 370.7, SD = 117.2) participants per statement. For hypotheses 3 through 6, we calculated the mean and standard deviation separately by participant sex to identify any potential differences in how males and females perceive the effectiveness of gendered language. Again, this normative data was generated using sample sizes ranging from 112 to 293 (Male: M = 190.8, SD = 67; Female: M = 175, SD = 49.1).

We then computed probability scores of positive phrasing, negative phrasing, masculine phrasing, and feminine phrasing for each action statement. More specifically, in the present study, the zero-shot classification task was performed using the

deberta-v3-large-zeroshot-v2.0 model from MoritzLauer on HuggingFace (Laurer et al., 2023). This model was fine-tuned on a mix of natural language inference (NLI) task datasets, equipping the model with a broader understanding of language and enabling it to generalize and perform classification analysis on our dataset of action statements. The model was configured to allow for multi-label classification, meaning that the text for each action statement could be associated with multiple labels to varying degrees. This is reflective of the nuanced and multifaceted nature of phrasing within human communication, where a single piece of text can evoke a range of sentiments and gendered meanings.

4 Experiments and Results

To test hypothesis 1, we conducted multiple linear regressions examining the effects of negative and positive phrasing on the mean and standard deviation of SJT effectiveness ratings. Our findings supported replicated Pawirosetiko and Perrotta (2024), showing that negative phrasing significantly lowered mean effectiveness ratings and increased response variability (B = -0.12, p = .001; B = 0.15, p < .001; respectively). Positive phrasing did not significantly predict the mean or standard deviation of effectiveness ratings in either analysis.

Hypotheses 2 and 3 posited that male participants would exhibit higher mean responding behavior to actions phrased in a more masculine way, while female participants would do the same for actions phrased in a feminine way. A multiple linear regression examining male participants' responses to masculine and feminine phrasing revealed a significant negative effect of masculine phrasing on mean ratings (B = -0.20, p = .002). This finding contradicts Hypothesis 2, suggesting that male participants rated actions phrased in a masculine way as less effective. Feminine phrasing did not significantly predict mean ratings for males. Similarly, and counter to Hypothesis 3, a corresponding analysis for female participants found a significant negative effect of masculine phrasing on mean effectiveness ratings (B = -0.17, p = .003). Feminine phrasing did not significantly predict mean ratings for females.

To examine Hypotheses 4 and 5, which predicted increased response variability for incongruent gender phrasing, we conducted further multiple linear regressions. For male participants, masculine phrasing significantly predicted increased variability in responding (B = 0.15, p = .004), contradicting Hypothesis 4. This suggests that actions phrased in a more masculine way led to more diverse effectiveness ratings from male participants. Feminine phrasing did not significantly predict variability for males. For female participants, masculine phrasing also significantly predicted increased response variability (B = 0.11, p = .024), supporting Hypothesis 5. Feminine phrasing did not significantly predict response variability for females.

5 Discussion

The present study sought to expand upon prior research examining the influence of language on responding behavior within SJTs. Replicating the findings of Pawirosetiko and Perrotta (2024), we verified that negative phrasing found in action statements significantly decreased mean effectiveness ratings and increased response variability. This suggests that the emotional tone of language can introduce bias into SJT responses, potentially affecting the validity of these assessments. As Pawirosetiko and Perrotta (2024) point out, this offers a potential explanation for certain evaluative biases that influence responding behavior in SJTs. Yet, through the use of these zero-shot classification models, researchers and practitioners can not only get an idea of the negative sentiment in their statements but also iteratively decrease such sentiment during the statement creation process. Thus, they can lower a potential barrier for adequately measuring the construct of interest.

Contrary to our additional hypotheses (i.e., 2 and 3) based on gender schema theory, we found no evidence that gender-congruent phrasing led to higher mean effectiveness ratings for either male or female participants. Instead, both sexes rated actions phrased in a more masculine way as less effective, irrespective of their own gender. This unexpected result challenges traditional assumptions about the alignment of gender schemas and behavior evaluations, warranting further investigation into the complex interplay of gender, language, and perception within SJTs. It is also worth considering that actions perceived as masculine might be associated with dominance, which could carry a negative tone (Roberts and Utych, 2020). This implies that respondents might be reacting to the dominance conveyed by the actions rather than their masculine nature, highlighting the need for further exploration

into how dominance influences SJT responding.

Interestingly, our findings partially supported the hypotheses (i.e., 4 and 5) regarding response variability. Masculine phrasing increased response variability for both male and female participants, indicating that actions aligning with traditional masculine stereotypes elicited a wider range of evaluations. This suggests that gendered language, even when incongruent with one's own gender, can introduce uncertainty and trigger diverse coping mechanisms during SJT responding. However, the lack of a similar effect for feminine phrasing raises questions about the differential impact of masculine and feminine stereotypes on cognitive processing.

Overall, the results regarding the impact of gendered language on responding behavior were mixed and unexpected. These findings have practical implications for the use of zero-shot text classification in SJT item writing. To illustrate, consider the following examples: a masculine-phrased item might be "Implement a competitive performance-based bonus system to drive productivity among team members," whereas a feminine-phrased item could be "Acknowledge each employee's unique contributions and provide personalized feedback to support their growth." Test developers should consider using zero-shot text classification to help balance the inclusion of actions that are more feminine versus more masculine, or perhaps even strip them of gendered language, to mitigate unintended biases and ensure a more equitable assessment. The present study contributes to the growing body of literature on SJT design by highlighting the multifaceted influence of language on responding behavior. It underscores the importance of using zero-shot text classification to meticulously craft action statements to mitigate unintended biases and ensure equitable assessment for all individuals, regardless of gender. Finally, the unexpected findings regarding gender-congruent phrasing suggest a re-evaluation of gender schema theory's applicability within SJTs.

6 Limitations and Future Research Directions

The study presents several limitations that warrant consideration. Firstly, while the zero-shot text classification model efficiently quantifies linguistic features, it may not fully capture the nuanced and context-dependent nature of gendered language. This is evident in the restricted range of feminine phrasing probability scores (0.0003-0.0557). While the actions themselves might have been mostly masculine and not very feminine, it's also possible that the zero-shot model didn't accurately capture feminine phrasing. Having subject matter experts (SMEs) rate statements on feminine and masculine scales might improve model coverage.

Additionally, relying on self-reported sex as a binary variable fails to capture the complex spectrum of gender identities. Gender schema theory suggests that people process gendered information based on their identification with masculine and feminine traits, beyond just biological sex. A potential mechanism for this could be differential memory relating to masculine and feminine words. Differential recall of masculine and feminine words, as found by Brown et al. (1980), may be moderated by gender identity. This suggests that recall bias related to gendered words could affect how respondents process and evaluate SJT statements. Future studies should incorporate validated measures of gender identity and gender-role attitudes, such as the Bem Sex-Role Inventory, for a more nuanced understanding of gender schemas and linguistic cues in SJT responses.

Another limitation is the operationalization of gendered phrasing. While the zero-shot classification model identified statements with varying degrees of masculine and feminine phrasing, the content of these statements may have confounded the results. Statements rated as more masculine might also be perceived as less desirable or effective due to their content. Future research should disentangle the effects of content and phrasing by using experimental manipulations or balancing the content of gendered statements.

7 Conclusion

This study highlights how language affects SJT responses, confirming that negative phrasing decreases average effectiveness ratings and increases response variability. Contrary to gender schema theory, gender-congruent phrasing did not lead to higher effectiveness ratings. Instead, masculine phrasing increased response variability for both genders. These findings underscore the potential of zero-shot text classification in refining SJT item development, ensuring more equitable assessments by mitigating biases introduced by emotional valence and gendered language.

References

W. Arthur, R. M. Glaze, S. M. Jarrett, C. D. White, I. Schurig, and J. E. Taylor. 2014. Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology*, 99(3):535–545.

S. L. Bem. 1981. Gender schema theory: A cognitive account of sex typing. *Psychological Review*, 88(4):354.

A. S. Brown, M. B. Larsen, S. A. Rankin, and R. A. Ballard. 1980. Sex differences in information processing. *Sex Roles*, 6:663–673.

M. S. Christian, B. D. Edwards, and J. C. Bradley. 2010. Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, 63(1):83–117.

L. Festinger. 1957. *A theory of cognitive dissonance*. Stanford University Press, Stanford, CA.

D. Gaucher, J. Friesen, and A. C. Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, 101(1):109.

C. N. Herde, F. Lievens, D. J. Jackson, A. Shalfrooshan, and P. L. Roth. 2020. Subgroup differences in situational judgment test scores: Evidence from large applicant samples. *International Journal of Selection and Assessment*, 28(1):45–54.

E. A. Kensinger and S. Corkin. 2003. Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words? *Memory & Cognition*, 31(8):1169–1180.

M. Laurer, W. van Atteveldt, A. Casas, and K. Welbers. 2023. Building efficient universal classifiers with natural language inference. arXiv. Http://arxiv.org/abs/2312.17543.

C. Leaper and M. M. Ayres. 2007. A meta-analytic review of gender variations in adults' language use: Talkativeness, affiliative speech, and assertive speech. *Personality and Social Psychology Review*, 11(4):328–363.

E. F. Loftus and J. C. Palmer. 1974. Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13(5):585–589.

J. M. Madera, M. R. Hebl, and R. C. Martin. 2009. Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology*, 94(6):1591.

M. A. McDaniel, N. S. Hartman, D. L. Whetzel, and W. L. Grubb. 2007. Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, 60(1):63–91.

M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236.

F. Patterson, S. Lopes, S. Harding, E. Vaux, L. Berkin, and D. Black. 2017. The predictive validity of a situational judgement test, a clinical problem solving test and the core medical training selection methods for performance in specialty training. *Clinical Medicine*, 17(1):13–17.

J. S. Pawirosetiko and J. Perrotta. 2024. What's in a word?: The impact of sjt action sentiment on responding behavior. Poster presented at the Society for Industrial and Organizational Psychology Annual Conference.

D. C. Roberts and S. M. Utych. 2020. Linking gender, language, and partisanship: Developing a database of masculine and feminine words. *Political Research Quarterly*, 73(1):40–50.

E. S. Webster, L. W. Paton, P. E. S. Crampton, and P. A. Tiffin. 2020. Situational judgement test validity for selection: A systematic review and meta-analysis. *Medical Education*, 54(10):888–902.

D. L. Whetzel, M. A. McDaniel, and N. T. Nguyen. 2008. Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, 21(3):291–309.