Pilot 3: Test Intervention and Determine Sample Size

The goal of this pilot study was to test our intervention and determine a sample size for our full study.

Participants. We have 80 participants in total for this study recruited through Prolific. All participants are over 18 years old and fluent in English.

Stimuli. We generated four letters (2 with more female-associated words and 2 with more male-associated words) using ChatGPT. The prompts we used for each letter are:

- **Female Prompt 1:** Pretend you are a high school teacher writing a moderately strong recommendation letter for a student who is applying to college. The letter is for a female student described as polite, organized, and intuitive.
- **Female Prompt 2:** Pretend you are a high school teacher writing a moderately strong recommendation letter for a student who is applying to college. The letter is for a female student described as warm, dependable, and insightful.
- **Male Prompt 1:** Pretend you are a high school teacher writing a moderately strong recommendation letter for a student who is applying to college. The letter is for a male student described as considerate, analytical, and logical.
- **Male Prompt 2:** Pretend you are a high school teacher writing a moderately strong recommendation letter for a student who is applying to college. The letter is for a male student described as honest, bright, and tenacious.

We used 1 communal word, 1 grindstone word and 1 agentic word in the prompts for generating female language letters, and used 1 communal word, 1 ability word and 1 agentic word for generating male language letters. For each prompt, we generated 3 letters and all authors read the letters to select stimuli that were sufficiently different while maintaining comparable quality. In this study, we used 1 female language letter and 1 male language letter to test the interventions.

Procedure. Each participant was randomly assigned to one of the eight conditions of a 2 x 4 design based on letter language and intervention. Participants were asked to read the recommendation letter and rate the competitiveness of the applicant on a scale of 1 (Extremely uncompetitive) to 7 (Extremely competitive).

Results. We used the data to determine the sample size for the full study. We conducted a 2-way ANOVA test using langue and intervention as factors and calculated partial η^2 for each factor which were used in G*Power[1] to calculate the sample size with a 0.95 power at the standard .05 alpha error probability. The results were 560 for intervention and 382 for language and we used 560 because our main goal is to test the effect of the interventions.

---

[1] https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower