Methodological Note

Austin Coffelt

February 21, 2026

1 Data Limitations

The data for my project currently has 2 significant limitations.

1.1 Namsor API

The first concerns the Namsor gender-checker API, which is probabilistic and may return incorrect results. In testing the API before use, I found that it returned the correct answer for 95 percent of names, matching the rate advertised on the Namsor website. Because misclassification is unlikely to be related to writing style, the effect on my project is small.

This is the same procedure for classifying gender that Hengel (2022) uses in her paper, which provides precedent for this method.

1.2 OpenAI LLM Evaluations

The second limitation concerns the LLM evaluations, which are biased by the alignment and the inherent randomness of the GPT-5 model. To account for these factors, I specifically engineered and tuned the prompt for structural consistency and clarity, and systematically validated the outputs to ensure compliance with the schema.

Despite the controls and processes put in place to validate the output, bias cannot be fully eliminated in the model, as slight changes to the model, the prompt, or the rubric influence scoring tendencies, and constraints on context length make it impossible to fully expand on our rubric. Preliminary results suggest that the scoring behavior is stable, but full validation (e.g., multi-run reliability analysis or inter-model comparison) has not yet been conducted.

2 Data Decisions

A significant portion of the dataset structure is designed to align with Hengel (2022) to facilitate replication and extension. The decision to use the 'Female Authorship' variable as the main variable of analysis comes from Hengel (2022), where she sees the readability scores starting to deviate when the ratio of women to men reaches .50. The variables used for robust analysis also come from Hengel (2022). This design ensures comparability with prior findings while allowing extension through new LLM-based measures.

The decisions I had to make about the data concerned the evaluation rubric and what to scrape from the DOI links. I chose the specific rubric to find the optimal balance between LLM context size and the capture of as many dimensions of writing style and tone as possible. Adding more criteria to the rubric increases the LLM's context size, making the output less reliable, while removing criteria provides less information about the paper's writing style and tone. The final 16-dimension rubric was selected to capture tone, assertiveness, readability, and evidentiary style, while remaining within a context length that preserves output consistency and allows structured numeric output parsing. The rubric may be revised in the future depending on LLM output validation.

Scraping targeted metadata necessary for replication of core variables in Hengel (2022). This led to the scraping of the abstract, the department that accepted the paper, and the editor responsible for its acceptance. Combining this data with the gender guesses yields most of the variables in Hengel (2022), though the writers' institutions and children are still missing.

3 Remaining Gaps

The remaining data gaps for this project include the institution, children, and native language, all of which are needed for the full replication and extension of Hengel (2022). To gather this information, I will likely need to either scrape the DOI link further to look for institutional information on the authors or manually look through the authors' information online (e.g., CV, personal website, Facebook)

References

Hengel, E. (2022). Publishing while female: Are women held to higher standards? evidence from peer review. *The Economic Journal*, 132(648):2951–2991.