Reviewer Preferences and Gender Disparities in Aesthetic Judgments

Ida Marie S. Lassen[1], Yuri Bizzoni[1], Telma Peura[1], Mads Rosendahl Thomsen[2] and Kristoffer Nielbo[1]

[1]Center for Humanities Computing Aarhus, Aarhus University, Jens Chr. Skous Vej 4, Building 1483,DK-8000 Aarhus C

[2]School of Communication and Culture - Comparative Literature, Aarhus University, Langelandsgade 139, Building 1580, DK-8000 Aarhus C

Abstract

Aesthetic preferences are considered highly subjective resulting in inherently noisy judgments of aesthetic objects, yet certain aspects of aesthetic judgment display convergent trends over time. This paper presents a study that uses literary reviews as a proxy for aesthetic judgment in order to identify systematic components that can be attributed to bias. Specifically, we find that judgments of literary quality differ across media types and display a gender bias. In newspapers, male reviewers have a same-gender preference while female reviewers show an opposite–gender preference. On the other hand, in the blogosphere female reviewers prefer female authors. While alternative accounts exist of this apparent gender disparity, we argue that it reflects a cultural gender antagonism that is necessary to take into account when doing computational assessment of aesthetics.

Keywords

aesthetic judgement, gender, bias analysis, literary review

1. Introduction

Aesthetic judgments are notoriously complex and subject to considerable variation because aesthetic objects are complex (ex. literature is a complex linguistic phenomenon that conveys information indirectly), aesthetic preferences are subjective (ex. readers have different aesthetic preferences), and there is a general lack of a shared measurement (ex. there is no definitive metric to measure aesthetics or aesthetic judgments). Literary quality, for instance, can be considered one of the most subjective fields of evaluation, and variation is mostly attributable to noise introduced by individual preferences. Yet the perception of literary quality from large amounts of readers over time does show convergent trends: communities tend to establish and update canons [12]; specific texts and narratives manage to remain popular [22] despite the changing of fashions and political phases and certain author names become eponymous of literary quality in different countries and throughout the social spectrum [2]. Some

CHR 2022: Computational Humanities Research Conference, December 12 – 14, 2022, Antwerp, Belgium

idamarie@cas.au.dk (I. M. S. Lassen); yuri.bizzoni@cc.au.dk (Y. Bizzoni); tpeura@cc.au.dk (T. Peura); madsrt@cc.au.dk (M. R. Thomsen); kln@cas.au.dk (K. Nielbo)

0000-0001-6905-5665 (I. M. S. Lassen); 0000-0002-6981-7903 (Y. Bizzoni); 0000-0001-8896-8603 (T. Peura); 0000-0002-4975-6752 (M. R. Thomsen); 0000-0002-5116-5070 (K. Nielbo)

facets of literary quality can be explained in terms of the literary content (ex. predictability of content, coherence of the narrative), while others depend on socio-cultural priors that introduce systematic variation in aesthetic judgments. It is the latter that are the object of this study, specifically the possible effects of gender on the assessment of literary quality as an example of how aesthetic judgments can be biased by contextual factors.

There are two important caveats to consider. First, we are not claiming that variability in aesthetic judgment is undesirable, on the contrary, it facilitates expressive variation and counters aesthetic standardization as has been the norm under some authoritarian regimes [4, 10, 13, 17]. Defining bias as a deviation from statistical parity, we are only interested in the systematic components of aesthetic judgment that can be attributed to such bias, specifically gender bias, and approach this problem from the perspective of fairness challenges in the classification of real-world data [20]. Second, it is not our intention to 'point fingers' or address specific individuals (ex. specific reviewers) or institutional levels of biases (ex. specific outlets). Fairness challenges first and foremost concern a systemic level of biases, that is, macro-relations that are systematic and disadvantaged groups of people based on their identity (gender, race, class, sexual orientation), while at the same time advantaging members of a dominant group. While at the individual level a bias effect may seem small or trivial, it is important to emphasize that systems of bias can result in rampant injustice [15].

The problem of literary quality's subjective status becomes even more intriguing when we turn to the challenge of its computational assessment. Most studies assume the possibility of one one-dimensional ground truth by modeling literary quality as a single rating or class associated with a text [9, 26, 25]. These ground truths are retrieved from various sources: literary critics, book sale numbers, bestseller lists, or crowd-sourced reader opinions. Such approaches have several limitations: relying only on experts' judgment (ex. awards, prestigious reviews) biases the model to reflect their preferences, but striving for representativity by crowd-sourcing opinions ends up ignoring important differences in the readers' population. To properly understand the scientific value of these ground truths and develop standardized measures of quality, it is necessary to model possible sources of bias.

Recent studies have analyzed the impact of the gender of authors as well as of reviewers in literary reviews. [24] investigates differences in sentiments in Norwegian book reviews and how literary reviewers are describing authors of the same and opposite gender. Their findings show differences in how female and male book authors are positively or negatively described and that the gender of the critics influences this difference. In line with the findings in [23] on Goodreads reviews, the authors point out, that male critics deem crime novels written by female authors and sentimental romance novel by male authors as negative and suggest that this indicate that book reviews contain the social hierarchies tending to ascribe emotional traits to women. In the Goodreads reviews, differences are both found in preferences of types of books as well as within genres, meaning that when reviewers of both genders read and review books of the same genre, differences in grading are found between male and female reviewers. In addition, the results show that within the majority of genres, readers prefer books written by an author of their own gender. Similarly, when Dutch readers were asked to rate both read and unread novels on a scale of 1–7 for their literary and overall quality [16] show that female authors receive significantly lower ratings than their male counterparts.

In the greater context of circulation and reception of books, [21] and [6] address the role

of both review and reviewer in the broader Anglophone literary field. The former point to an imbalance found in the British and Australian review scenes: Most book reviewers are men, and books reviewed are often written by men, resulting in books written by female authors being treated like a niche. The latter offers a historical account of the gendered structure of the literary field and maps out how authors build their reputations and accumulate prestige in contemporary book publishing. By taking the historical perspectives of the literary field into account, it might not be surprising that gender disparities still exist. As the reception and judgment of books exist within a greater societal context where structural oppression occurs, signs of systemic inequality call for further investigation.

In other areas, studies have examined the role of gender in assessment situations. [18] shows how students' ratings of instructors are biased towards a more positive assessment of male instructors compared to female instructors. By conducting their study in an online learning situation, the authors were able to disguise the gender identity of the instructors. The found bias was not dependent on the actual gender of the instructor, but on the perceived gender of the instructor. That allows for a conclusion that points out that the gender bias is not a result of the gendered behavior of the instructors, but actual bias in the students, suggesting that a female instructor would have to work harder than a male to receive comparable ratings. In an academic context, a study from 2020 [14] shows how female applicants were less likely than male applicants to receive access to resources (in terms of telescope time) when the review process was single–, rather than dual–anonymized. In particular, the findings indicate that male reviewers rated female applicants significantly worse than they rated male applicants before dual–anonymization was adopted, and after applying dual–anonymization, the gender bias was reduced. Similar results are shown in the hiring process of orchestra musicians [11] and several studies have shed light upon the effect of gender in hiring processes [3, 5]. Evidence of gender bias across domains may indicate that similar structural dynamics are at play, and hence, not a unique gender bias evolving in the field of literature.

In addition to gender disparities, other social markers might also play a role in aesthetic judgments and requires some awareness in the following analysis. [19] investigates differences between theater reviews written in blogs and newspapers and concludes that even though such reviews are highly similar, differences are found at a subtle level: whereas bloggers tend to focus on categories related to affect and audience relations, reviews written by journalists rely on descriptive approaches to the play at hand. With a focus on book blogs, the analysis in [8] shows how the blog media enables mass participation in reader culture. However, even within the blogosphere, a hierarchy of 'reader capital' exist, and some bloggers obtain status as 'tastemakers.' These findings indicate a need for clustering of reader types when modeling reader preferences [1].

2. Methods

2.1. Data

The data set covers book reviews published in Danish media in the years 2010-2021. The data are retrieved from the online platform bog.nu's API which collects book reviews published in Danish media. This includes reviews written in national newspapers, literary magazines,

online media as well as in personal blogs. See table 1 for a brief overview of the data set.

2.2. Grade Transformation and Estimation

As different media use different grading scales, the grades on bog.nu are transformed to a 100-point grading scale. This approach, however, results in a sparse distribution of grades as the use of the original grading scales maps onto different intervals on the 100-point scale. Instead of this naive approach, we have used the original grade and applied a linear transformation to map all grades to a shared 6-point scale.[1] Mapping from an a-b–point scale to a 1-6–point scale:

$$Y = \lfloor (B-A)\frac{(x-a)}{(b-a)} + A \rfloor = \lfloor (6-1)\frac{(x-a)}{(b-a)} + 1 \rfloor = \lfloor \frac{5(x-a)}{b-a} + 1 \rfloor \quad (1)$$

Figure 1 shows the distributions of grades in Danish Newspapers transformed into a shared 6–point scale. Some media do not provide a grade in a given review, but only a qualitative review. Bog.nu does however provide a quantification of the review, which is estimated by a human editor. For reviews written in Danish newspapers, this estimation procedure is used in less than 25% of the cases. Two important clarifications are needed: first, these estimates are made for both genders of both reviewers and authors. Secondly, to test the robustness of these estimates, the analysis below was performed both on the full data set and on the subset with original quantitative grades given in the reviews. We see that the same trends occur when excluding the reviews with estimated grades.

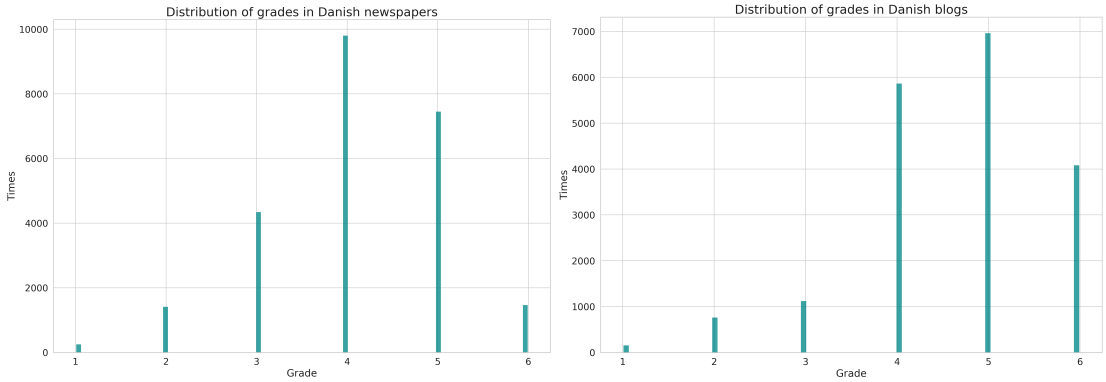

Figure 1: Histogram over grades, in newspapers and blogs, after grades are transformed into shared 6-point Likert scale.

2.3. Feature Distributions

The original data set from bog.nu does not contain gender for all authors and reviewers. We used a gendered name list to retrieve the missing gender variables. We are working with a

[1]It should be noted that most six-point scales have become the standard in many review outlets.

Table 1
An overview of the dataset presented in this paper. The category of Online media includes (literary) sites that fall between online newspapers and personal blogs.

Data set overview					
No. of reviews	**57 369**	**No. of different titles**	**14 647**	**No. of reviews by media type**	
Male reviewers	19 119	Male authors	8 056	Newspapers	22 131
Female reviewers	29 084	Female authors	6 591	Blogs	16 791
Unknown	9 166			Online media	10 635
				Blog-like websites	3 456
				Weekly magazines	1 566
				Professional magazines	168

binary understanding of gender, and we have used the API genderize.io that returns the probability of a name being either male or female, based on a data set of 250,000 names.[2] We are aware of the problems with this method and how it rules out other gender identities [7]. However, a binary understanding of gender is necessary for our analysis to understand the existing structures between men and women in contemporary society – and the literary review scene.

Looking at feature distributions in our data set, we see that both gender variables and grades differ across media types. As shown in Figure 2, we see a highly skewed gender distribution across media types: at the number of *reviews*, male reviewers reviewing male authors are the dominant group in newspapers, whereas female reviewers reviewing female authors are the dominant group in the blogosphere.

Focusing on the number of *reviewers* the gender distribution reflects the one shown in the number of reviews. In the data set, we have 621 unique reviewers writing in Danish newspapers. Out of these, 239 are women, 378 are men, and 4 are unknown according to the gender retrieval method described above. As the bog.nu data set does not contain reviewer id for blog reviewers, a similar calculation cannot be made for blogs.

Besides the distribution of gender, we have furthermore identified different 'grading behaviors' in newspapers and in blogs (see right-hand side of Figure 2). Hence, due to different distributions of gender as well as grades given across media types, we have in the rest of this study divided our analysis in two: newspapers and blogs.

2.4. Model

In order to estimate the relative effect of author and reviewer gender on reviewer assigned grade (six-point scale), we fit the following linear model:

$$y_i^{grade} = x_i\beta_{author} * x_i\beta_{reviewer} + \epsilon_i \tag{2}$$

with the null model that $\beta_{author} = \beta_{reviewer} = 0$

[2]Testing the accuracy of genderize on a gendered name list from Statistic Denmark: ACC = 0.93 for n = 10,000

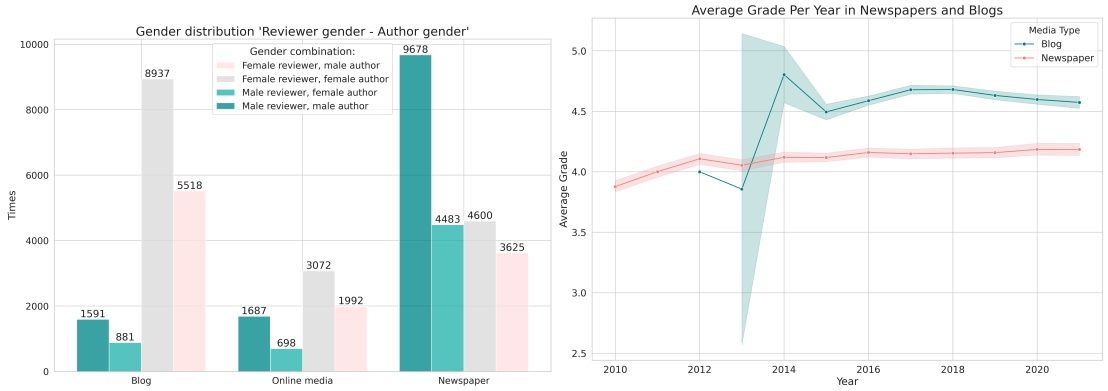

Figure 2: The histogram shows the distribution of reviewer- and author gender across the media types blog, online media, and newspapers. The line plots to the right show the average grade in newspapers and blogs respectively.

Where y_i is the grade of review i, x_i is the predictor value (gender) of review i, β represents unknown parameters and ϵ is the error terms. A linear model is fitted for both blogs and newspapers respectively. We test

3. Results

For newspaper reviews, fitting y_i (grade of review) in the model above with ordinary least squares (OLS), we get the results shown in Table 2. The model and all contrasts are statistically significant ($p < .0001$). Conceptually, same gender (male reviewer–male author, female reviewer–female author) reviews span the extreme values, while the opposite gender (female reviewer–male author, male reviewer–female author) represents the middle of the distribution, see left-hand side of Figure 3. Female reviewers reviewing female authors account for the on average lowest grade. Male reviewers reviewing male authors results in the highest grade, with a 0.2 average grade increase. Opposite gender reviews are statistically speaking indistinguishable, but they differ on average by 0.1-grade point from the same gender scoring.

Table 2
Newspapers: Results for an OLS predicting grade based on the gender combination: reviewer and author. The t-test shows that all results are statistically significant.

Gender combination (reviewer * author)	Average grade	*SD*	t	$p < \|t\|$	*CI* 95%
Intercept [female * female]	3.9830	0.015	268.616	0.0001	[3.954, 4.012]
female * male	0.0881	0.022	3.946	0.0001	[0.044, 0.132]
male * female	0.1093	0.021	5.179	0.0001	[0.068, 0.151]
male * male	0.2079	0.018	11.544	0.0001	[0.173, 0.243]

For the data set on the blogs, fitting y_i (grade of review) in the model above with ordinary least squares (OLS), we get the results shown in Table 3. The model and all contrasts besides

the male-male combination are statistically significant ($p < .0001$ for female-female and female-male, and $p < .05$ for male-female) and we see that in contrast to the results in newspapers, female reviewer–female author account for the on average highest grade while the combination female reviewer–male author result in the, on average, lowest grade with a difference of 0.1-grade point between those two. The large standard deviation for male reviewer–female author is due to the low number of reviews with this combination (n=881). See right-hand side of Figure 3

Table 3
Blogs: Results for an OLS predicting grade based on the gender combination: reviewer and author. The *t*-test shows that all results besides the male-male combination are statistically significant.

Gender combination (reviewer * author)	Average grade	*SD*	*t*	$p < \|t\|$	*CI* 95%
Intercept [female * female]	4.6737	0.012	401.865	0.0001	[4.651, 4.697]
female * male	-0.1212	0.019	-6.402	0.0001	[-0.158, -0.084]
male * female	-0.0931	0.038	-2.440	0.015	[-0.168, -0.018]
male * male	-0.0586	0.030	-1.983	0.047	[-0.117, -0.001]

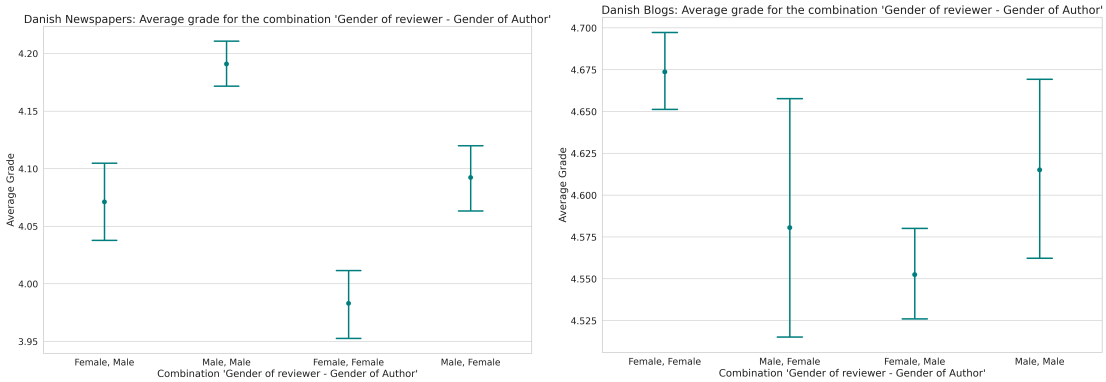

Figure 3: Point plots showing the average grades in each of the four gender combinations. The lines indicate the standard deviations with a confidence interval of 95%. The left-hand figure shows grades given in newspapers, the right-hand figure shows grades given in blogs.

Finally, as mentioned in section 2 and shown in Figure 2, men account for the majority of reviews in the newspapers. Men actually dominate in the number of reviews (63% are written by men and out of these reviews 69% are reviews of male authors). Figure 4 shows the development over the years 2010-2021. Here we see that the fraction of female authors being reviewed is slightly increasing, but the fraction of male reviewers is relatively stable through the years.

For blogs, we see an extreme overweight of female reviewers reviewing female authors (85% are written by women and out of these reviews 60% are reviews of female authors) with little to almost no changes in the years 2016-2021. Be aware that the gender distribution shown in Figure 4 are made on the number of reviews and not on the unique reviewers. As some bloggers are highly productive, the picture might look different if we looked at unique reviewers. How-

ever, as mentioned in section 2 this is not possible for blogs as reviewer id for blog reviewers is lacking from the bog.nu data set. Nevertheless, looking at the number of published reviews show the gender distribution in media coverage.

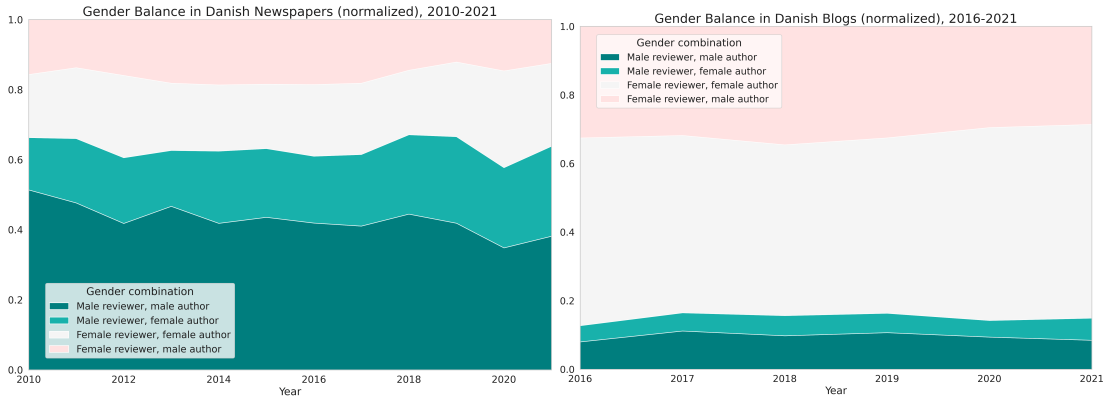

Figure 4: Stacked areas plots showing the percentage distribution of the four different gender combinations over the years (2010-2021 for newspapers and 2016-2021 for blogs) on the number of published reviews. The left-hand figure shows gender balance in newspapers, the right-hand figure shows gender balance in blogs.

4. Discussion

In line with the results in [23, 24, 16], we show that the gender of authors as well as of reviewers play a role in literary reviews. In particular, the results above show that

- in blogs, which women strongly dominate, women review same-gender authors more positively than opposite-gender authors.
- in newspapers, which men dominate, men review same gender authors more positively than opposite gender authors, and women show the reverse pattern, that is, same gender authors are reviewed more negatively than opposite gender authors.

From this, we can conclude that female grading behavior differs in media type. We see a preference for female authors in blogs and an opposite preference in newspapers. Still, we also note that this difference correlates with the gender majority in the media type – female reviewers prefer female authors in blogs where women dominate, and like male authors in newspapers where men dominate several reviewers and authors.

Where the blogosphere is a new medium, the newspaper outlet is a well-established form of traditional media, which historically has excluded women and minority people, potentially influencing gender distribution today. A partial explanation of the grading behavior is that if males display a same-gender preference and male reviewers make up the majority of newspaper reviewers, the gender minority adapts to this preference and develops a same–gender antagonism. At a more general level, the same gender preference of males in aesthetic judgment may reflect a cultural gender antagonism that follows a long historical trajectory. The

female opposite gender preference in newspapers is likely to follow the same cultural gender antagonism. As the number of male blog reviewers reviewing male authors is too low to show statistically significant results, a similar conclusion cannot be drawn for the blogs.

There are caveats to this interpretation. First, the gender bias may be confounded with expertise bias, that is, that specific literary language leads to higher literary appreciation. If women, in general, write more genre literature, then the observed difference may stem from a difference in the complexity of linguistic features. To resolve this, we would need genre classification for all reviewed books and a model of the distribution of genres across media types. Second, although the average differences in grades are highly significant, the effect size is not considerable (ex., 0.2 points on a six-point scale for the same gender in newspaper reviews). This, however, begs the question, how large is a systematic difference supposed to be before it counts as a bias? We would argue that whenever we find the systematic variation that co-insides with demographic variables, we will likely see an indication of a relevant bias irrespective of the effect size. Conversely, if only a difference of a large magnitude (ex. two to three points on a six-point scale) were to count, then biases would only reflect common-sense propositions that most of us would share irrespective of their truth (ex. if women were, on average reviewed two to three points lower, most of us would agree that they were worse writers).

The last caveat points to an important issue; we are not arguing that a specific newspaper, or all newspapers for that matter, follow an explicit exclusionary strategy formulated by male reviewers and editors – nor that bloggers purposely exclude male authors. There are two sources of error whenever we make a judgment: bias and noise. While noise is randomly distributed and lacks a systematic explanatory mechanism, biases are systematic and can be explained in terms of a mechanism. Demographic biases often originate in the systemic oppression of minority groups. For the specific review cases, the results are likely to mirror existing societal oppressive structures such as those found in [6, 18, 14, 11, 3, 5]. We expect that majority groups, in general, will define norms and values that result in biased judgments irrespective of societal domain.

References

[1] Y. Bizzoni, I. Lassen, T. Peura, M. Thomsen, and K. Nielbo. "Predicting Literary Quality How Perspectivist Should We Be?" In: *1st Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Marseille, France: European Language Resources Association (ELRA), 2022, pp. 20–25.

[2] H. Bloom. *The western canon: The books and school of the ages*. Houghton Mifflin Harcourt, 2014.

[3] A. Booth and A. Leigh. "Do employers discriminate by gender? A field experiment in female-dominated occupations". In: *Economics Letters* 107.2 (2010), pp. 236–238.

[4] E. Buch, I. C. Zubillaga, and M. D. Silva. *Composing for the State: Music in Twentieth-Century Dictatorships*. Routledge, 2016.

[5] M. S. Cole, H. S. Feild, and W. F. Giles. “Interaction of recruiter and applicant gender in resume evaluation: a field study”. In: *Sex Roles* 51.9 (2004), pp. 597–608.

[6] A. Dane. *Gender and Prestige in Literature*. Springer, 2020.

[7] S. Dev, M. Monajatipoor, A. Ovalle, A. Subramonian, J. M. Phillips, and K. Chang. “Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies”. In: *CoRR* abs/2108.12084 (2021).

[8] E. Driscoll. “Book blogs as tastemakers”. In: *Participations. Journal of Audience & Reception Studies* 16 (2019), pp. 280–305.

[9] C. Ferrer. “Canonical values vs. the Law of large numbers: The Canadian Literary Canon in the Age of Big Data”. In: *Rupkatha Journal on Interdisciplinary Studies in Humanities* 5.3 (2013), pp. 81–90.

[10] V. Frajese. *Nascita dell’Indice: la censura ecclesiastica dal Rinascimento alla Controriforma*. Vol. 13. Morcelliana, 2006.

[11] C. Goldin and C. Rouse. “Orchestrating impartiality: The impact of” blind” auditions on female musicians”. In: *American economic review* 90.4 (2000), pp. 715–741.

[12] J. Guillory. *Cultural capital: The problem of literary canon formation*. University of Chicago Press, 1993.

[13] A. Herrero-Olaizola. *The censorship files: Latin American writers and Franco’s Spain*. SUNY Press, 2012.

[14] S. K. Johnson and J. F. Kirk. “Dual-anonymization yields promising results for reducing gender bias: A naturalistic field experiment of applications for Hubble Space Telescope time”. In: *Publications of the Astronomical Society of the Pacific* 132.1009 (2020), p. 034503.

[15] D. Kahneman, O. Sibony, and C. R. Sunstein. *Noise: A Flaw in Human Judgment*. New York: Little, Brown Spark, 2021. 464 pp.

[16] C. Koolen, K. van Dalen-Oskam, A. van Cranenburgh, and E. Nagelhout. “Literary quality in the eye of the Dutch reader: The National Reader Survey”. In: *Poetics* 79 (2020), p. 101439.

[17] Y. Kwon and J. Wood. “Literature and art in North Korea: Theory and policy”. In: *Korea Journal* 31.2 (1991), pp. 56–70.

[18] L. MacNell, A. Driscoll, and A. N. Hunt. “What’s in a name: Exposing gender bias in student ratings of teaching”. In: *Innovative Higher Education* 40.4 (2015), pp. 291–303.

[19] M. Maignant, G. Brison, and T. Poibeau. “Text Zoning of Theater Reviews: How Different are Journalistic from Blogger Reviews?” In: *Workshop on Natural Language Processing for Digital Humanities*. 2021, pp. 138–143.

[20] T. Miconi. “The impossibility of” fairness”: a generalized impossibility result for decisions”. In: *arXiv preprint arXiv:1707.01195* (2017).

[21] C. Squires. “The Review and the Reviewer”. In: Routledge, 2020, pp. 117–132.

[22] J. Stephens and R. McCallum. *Retelling stories, framing culture: traditional story and metanarratives in children’s literature*. Routledge, 2013.

[23] M. Thelwall. “Reader and author gender and genre in Goodreads”. In: *Journal of Librarianship and Information Science* 51.2 (2019), pp. 403–430.

[24] S. Touileb, L. Øvrelid, and E. Velldal. “Gender and sentiment, critics and authors: a dataset of Norwegian book reviews”. In: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. 2020, pp. 125–138.

[25] M. Walsh and M. Antoniak. “The Goodreads ‘Classics’: A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism”. In: *Journal of Cultural Analytics* 4 (2021), pp. 243–287.

[26] X. Wang, B. Yucesoy, O. Varol, T. Eliassi-Rad, and A.-L. Barabási. “Success in books: predicting book sales before publication”. In: *EPJ Data Science* 8.1 (2019), pp. 1–20.

5. Online Resources

See https://zenodo.org/record/7050235 for code.