Exploring the Effect of Author and Reader Identity in Online Story Writing: the STORIESINTHEWILD Corpus

Tal August[†] **Maarten Sap**[†] **Elizabeth Clark**[†]
Katharina Reinecke[†] **Noah A. Smith**[†◇]
[†]Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA, USA
[◇]Allen Institute for Artificial Intelligence, Seattle, WA, USA

Abstract

Current story writing or story editing systems rely on human judgments of story quality for evaluating performance, often ignoring the subjectivity in ratings. We analyze the effect of author and reader characteristics and story writing setup on the quality of stories in a short storytelling task. To study this effect, we create and release STORIESINTHEWILD, containing 1,630 stories collected on a volunteer-based crowdsourcing platform. Each story is rated by three different readers, and comes paired with the author's and reader's age, gender, and personality.

Our findings show significant effects of authors' and readers' identities, as well as writing setup, on story writing and ratings. Notably, compared to younger readers, readers age 45 and older consider stories significantly less creative and less entertaining. Readers also prefer stories written all at once, rather than in chunks, finding them more coherent and creative. We also observe linguistic differences associated with authors' demographics (e.g., older authors wrote more vivid and emotional stories). Our findings suggest that reader and writer demographics, as well as writing setup, should be accounted for in story writing evaluations.

1 Introduction

Reading or writing a story is an inherently subjective task that depends on the experiences and identity of the author, those of the reader, and the structure of the writing process itself (Morgan and Murray, 1935; Conway and Pleydell-Pearce, 2000; Clark et al., 2018). Despite this subjectivity, many natural language processing tasks treat human judgments of story quality as the gold standard for evaluating systems that generate or revise text. In creative applications, such as machine-in-the-loop story writing systems (Clark et al., 2018), it is important to understand sources of variation in judgments if we hope to have reliable, reproducible estimates of quality.

In this work, we investigate how an author's and reader's identity, as well as overall writing setup, influence how stories are written and rated. We introduce and release STORIESINTHEWILD,[1] containing 1,630 short stories written on a volunteer-based crowdsourcing platform, paired with author demographics and personality information. For each story, we obtain three sets of ratings from third-party evaluators, along with their demographics and personality.

Our findings confirm that author identity, reader identity, and writing setup affect story writing and rating in STORIESINTHEWILD. Notably, people in general preferred stories written in one chunk rather than broken up into multiple stages. Raters age 45 and over generally rated stories as less creative, more confusing, and liked them less compared to raters under age 45. Additionally, we find that, in our corpus, men were more likely than women to write about female characters and their social interactions, and compared to younger authors, older authors wrote more vivid and emotional stories. We also find evidence of reader and author personality, and their interaction, influencing ratings of story creativity.

Our new dataset and results are first steps in analyzing how writing setup and author and reader traits can influence ratings of story quality, and suggest that these characteristics should be accounted for in human evaluations of story quality.

2 Background and Research Questions

To guide our study, we craft several research questions informed by existing literature on story writing and the relationship between author identity

[1] http://tinyurl.com/StoriesInTheWild

Proceedings of the 1st Joint Workshop on Narrative Understanding, Storylines, and Events, pages 46–54

and language, outlined below.

RQ1 *How are author gender, age, and personality traits associated with language variation in stories?* A wealth of work has shown an association between an author's mental states and their language patterns. Variation in pronoun usage, topic choices, and narrative complexity correlates strongly with the author's age and gender (Nguyen et al., 2016) and moderately with their personality (Yarkoni, 2010). We aim to confirm these differences in a prompted storytelling setting, since most work has focused on self-narratives (e.g., diaries and social media posts; Pennebaker and Seagal, 1999; Hirsh and Peterson, 2009; Schwartz et al., 2013), with the exception of the essays studied by Pennebaker et al. (2014).

RQ2 *How are rater gender, age, and personality traits associated with variation in story quality ratings?* Ratings of stories are often only used to evaluate a story writing system's output (e.g., Fan et al., 2018; Yao et al., 2019) or to develop automatic evaluation metrics (e.g., Hashimoto et al., 2019; Purdy et al., 2018), ignoring the rater's identity. However, prior work has shown differences in crowdsourcing worker's behavior or annotations based on task framing (Levin et al., 2002; August et al., 2018; Sap et al., 2019) or the annotator's own identity or experiences (Breitfeller et al., 2019; Geva et al., 2019). We seek to confirm and characterize these differences in our story rating task. As a follow-up to **RQ2**, we also investigate the interaction between author and rater demographics on story ratings.

RQ3 *Is writing setup associated with different ratings of story quality?* Past work has investigated story writing as a turn-taking game (Clark et al., 2018) or as a distributed activity (Teevan et al., 2016) rather than a single event. We investigate whether writing setup (writing a story all at once or sentence-by-sentence) impacts overall story quality.

3 STORIESINTHEWILD Collection

We introduce and release STORIESINTHEWILD, containing 1,630 short stories (§3.1) paired with author demographics and personality information.[2] We pair these stories with third-party ratings (§3.2) to evaluate the effect of writing setup and author identity on story writing.

		total	written in... *full*	written in... *seq.*
stats	# stories	1630	792	838
	avg. # tokens	592	583	600
	avg. writing time (min.)	9.30	9.98	8.72
	avg. key press time (sec.)	0.96	1.09	0.83
ratings	coherent	4.52	4.78	4.27 **
	confusing	3.44	3.19	3.67 **
	creative	4.00	4.09	3.90 *
	entertaining	3.95	4.10	3.81 **
	grammatical	4.22	4.39	4.06 **
	liked	3.89	4.05	3.73 **

Table 1: Statistics in STORIESINTHEWILD for all stories, as well as broken down by writing setup (*full*: written in full, *seq.*: written sequentially). Discussed in §4.3, rating differences are significant after Holm-correcting for multiple comparisons (*: $p < 0.01$, **: $p < 0.001$), but story length (# tokens), writing time, and writing speed (key press time) are not.

3.1 Crowdsourcing Stories

To construct STORIESINTHEWILD, we first collected 1,630 written stories using a volunteer-based online study platform, LabintheWild (Reinecke and Gajos, 2015).[3] Following best practices in recruiting on LabintheWild (August et al., 2018), we advertised our study as a way for participants to learn more about themselves by seeing how a simple pronoun-based classifier can predict their personality based on their story writing (described in Appendix A.1).

We first collected participants' identity and demographics (age, gender, race, and education level). Then, participants chose the *topic* of their story by selecting one of five preview thumbnails, each representing one of five image strips that participants subsequently used as prompts for their story. We selected the images from the Visual Storytelling dataset of Flickr images (Huang et al., 2016) and a cartoon dataset (Iyyer et al., 2017). All images are shown in Figure 1 in Appendix A.

Writing setup After choosing a topic, all authors are presented with a five image sequence corresponding to the topic they chose to write about. We then randomly assign authors to one of two writing setups: (1) *all at once* or (2) *se-*

[2]Each stage of data collection was approved by the authors' institutional review board (IRB).

[3]LabintheWild recruits study participants using intrinsic motivations (as opposed to monetary compensation, cf. Amazon Mechanical Turk), such as the the desire to compare oneself to others or to support science (Jun et al., 2017).

quential, both shown in Figure 2 in Appendix A. In (1), participants simply write a full 5–10 sentence story. In (2), participants are instructed to write five sets of 1–2 sentences in an accordion of text boxes, each box corresponding to an image in the strip. The second writing setup is inspired by machine-in-the-loop turn-taking for story writing (Clark et al., 2018). Once each text box is submitted, participants can no longer edit that text.

In both setups, participants are instructed to tell a story rather than just describe the images, to make sure their story has a clear beginning, middle, and end, and to use correct punctuation. The task took around 9 minutes in both conditions.

Following the story writing, participants can optionally fill out the Ten Item Personality Measure (TIPI; Gosling et al., 2003), a short personality questionnaire based on the Five Factor Model (FFM; Costa Jr and McCrae, 2008).[4]

Author demographics Of the authors in STORIESINTHEWILD, 57% were women and 40% men (3% declined to state their gender), with an average age of 25±12 years and an average of 14.30±4.20 years of education including primary school. Of the authors, 56% were white, 28% Asian, and 3% African-American (13% selected another ethnicity/race); we did not restrict participation to any specific country. 1,133 (70%) authors took the personality questionnaire.

3.2 Rating Stories

We create an Amazon Mechanical Turk task to obtain quality ratings for each of the stories collected in our previous task. For each story, we ask U.S.-based workers to rate stories on 6 dimensions (listed in Table 1), using a 7-point Likert scale.[5] Those dimensions include 5 fine-grained quality dimensions (e.g., grammaticality, coherence), as well as an overall impression of the story ("I liked this story"). Each worker also optionally filled out their demographics information (age, race, gender, education level). Additionally, as a measure of intellect and creativity, workers filled out the four openness items from the Mini-IPIP Big 5 personality scale (Donnellan et al., 2006).

Rater demographics 56% of our raters were women and 42% were men. 79% identified as white, 6% as African-American, and 6% as Asian. On average, their age was 40±12 years, and they had 15±3 years of education, including primary school.

4 Analyses

We investigate the effects of author and rater characteristics on the story's language and ratings. Unless otherwise specified, we only consider the male and female gender labels[6] and use a continuous representation of age and personality. We also explore the impact the writing setup—whether authors wrote stories all at once or in sequential chunks—has on story ratings.[7]

Note that our findings are simply measuring associations between aggregate categories (e.g., number of pronouns used, authors over age 45) and should not be interpreted as applying to individual data points with specific contexts.

4.1 Author Identity (RQ1)

To analyze which types of words are associated with different demographic identities, we extract psychologically relevant linguistic categories from stories, using the Linguistic Inquiry Word Count (LIWC; Pennebaker et al., 2015). For each LIWC category, we compute a linear regression model on the z-scored features, controlling for writing setup and topic choice. We only report regression coefficients (βs) that are significant after Holm correction for multiple comparisons (Holm, 1979).

Gender, age We find that the author's age, gender, and personality correlate with differential usage of linguistic categories, controlling for image choice and writing setup.[8] Specifically, we find that men used more personal pronouns ($|\beta| = 0.30$, $p < 0.001$) and social words ($|\beta| = 0.28$,

[4]The FFM delineates five dimensions of personality (openness to experience, conscientiousness, extraversion, agreeableness, neuroticism), each represented as a continuous score. For more details, we refer the reader to http://en.wikipedia.org/wiki/Big_Five_personality_traits.

[5]To ensure the quality of responses, we restrict the task to workers with 99% or above approval rate and at least 1,000 HITs approved. Additionally, we ask that workers write out a short piece of feedback to improve the story, to encourage them to think critically while rating stories.

[6]Gender is a social construct that goes beyond the man-woman binary (Lorber et al., 1991); however, a more complex analysis is not possible given the limited number of individuals not identifying as male or female in our data.

[7]We exclude author and reader education from our findings, as we did not find any signification effects for those variables.

[8]See Appendix B.1 for associations between author demographics and image choice.

$p < 0.001$) to describe characters (specifically, female characters, $|\beta| = 0.33$, $p < 0.001$), compared to women. Controlling for gender effects, our findings show that older authors wrote more emotional and positive stories ($|\beta| = 0.05$ and $|\beta| = 0.04$, respectively, $p < 0.05$) that contained more visual descriptions ($|\beta| = 0.05$, $p < 0.001$), whereas younger authors used past tense more ($|\beta| = 0.04$, $p < 0.05$).

Personality We find significant correlations between LIWC categories and an author's personality traits, controlling for age and gender (see Appendix B.2 for the full set of results). Notably, highly conscientious authors focused on character motivations ($|\beta| = 0.12$, $p < 0.05$) and used a more positive tone ($|\beta| = 0.14$, $p < 0.01$), compared to low-conscientiousness authors who wrote stories that tended to be more negative ($|\beta| = 0.11$, $p < 0.1$). Finally, less agreeable authors used more swearing ($|\beta| = 0.15$, $p < 0.1$), and more differentiating words ($|\beta| = 0.10$, $p < 0.1$) compared to more agreeable authors.

4.2 Rater Identity (RQ2)

We examine the association between rater traits and their story ratings using linear regressions controlling for image type and writing setup (similar to §4.1). We also investigate interaction effects with author demographics, and show the full results of our regressions in Appendix B.3.

Gender, age For age, we first noticed that older workers rated stories noticeably more negatively than younger workers (e.g., $r = -.08$, $p < .001$ for both the like and entertaining ratings). When inspecting the data we noticed this trend was most defined for raters age 45 or older, and so we perform our analyses below using a binarized age variable, splitting raters as either 45 or older ($N = 921$) and younger than 45 ($N = 1916$).

Our findings indicate that, compared to younger raters, raters of age 45 and older liked the stories significantly less ($|\beta| = 0.42$, $p < 0.001$), and rated them as substantially less entertaining ($|\beta| = 0.39$, $p < 0.001$), less creative ($|\beta| = 0.25$, $p < 0.05$), more confusing ($|\beta| = 0.27$, $p < 0.05$), and less grammatical ($|\beta| = 0.30$, $p < 0.05$). Interestingly, there was no significant association between annotator gender and story ratings.

Personality Openness to experience is often linked to creativity (McCrae, 1987), so we explore how ratings of creativity are associated with rater and author openness to experience personality scores. We find significant correlations between story ratings and rater openness to experience. Specifically, raters with higher openness to experience thought stories were generally more creative ($|\beta| = 0.38$, $p < 0.05$) and less confusing ($|\beta| = 0.64$, $p < 0.001$). Additionally, authors with higher openness scores wrote stories that were rated more creative ($|\beta| = 0.35$, $p < 0.1$)

Author-Rater Identity Interactions

We also investigate story ratings through the lens of author and rater demographics to see if any shared traits across raters and authors were associated with rater preferences.

While both reader and writer openness to experience were associated with significantly higher ratings of creativity, the interaction between the two was negative ($|\beta| = 0.50$, $p < 0.1$), meaning that as writer and reader openness to experience increased, the reader's rating of the story's creativity actually decreased. No other interactions (e.g., age, gender) were significant in our sample.

4.3 Differences in writing setup (RQ3)

We quantify the differences in ratings for our two writing setups. We average the ratings for each story, and report differences in Table 1 using Cohen's d. We find that stories written in full are rated to have higher quality across all dimensions, compared to stories written sequentially.

We also find that certain story topics were preferred over others ($F = 26.17$, $p < 0.001$). Specifically, stories written about the dog prompt were liked significantly more than others ($p < 0.001$), and those about the jail prompt significantly less ($p < 0.001$).

5 Conclusion

In this study we find that differences in author characteristics are associated with linguistic differences in stories and that rater characteristics are associated with differences in ratings. For authors, men were more likely than women to write about female characters and their social interactions, and compared to younger authors, older authors wrote more vivid and emotional stories. Raters preferred stories written all at once rather than broken up into multiple stages, and raters age

45 and older rate stories significantly lower than raters under age 45. We release our dataset, STORIESINTHEWILD, containing 1,630 stories with quality ratings and anonymized author and rater demographics.

Our results suggest that author and reader characteristics (e.g., demographics, personality) could explain variations in story writing evaluations. While work has shown that some study designs are more robust against this variation, (e.g., by ranking instead of rating Yannakakis and Martínez, 2015), rater differences could still lead to variation in annotations. We recommend that evaluations include some ability to collect characteristics, such as a short demographics and personality questionnaire, in order to assess any influence of these variables.

Furthermore, future work could explore alternative ways of collecting author and reader characteristics during evaluations. While demographic questionnaires are common and short (e.g., to collect gender and age would require two questions), full personality questionnaires are time consuming, asking multiple questions for each characteristic. Study designers could instead use reduced questionnaires, such as the ten item personality inventory (TIPI; Gosling et al., 2003). Alternatively, focusing on fewer, more highly trained raters—that represent a diverse set of demographics and personality—could reduce the cost of collecting many rater demographics. Finally, future work should investigate whether annotator variance might be better captured with psychological factors related to reading (e.g., propensity for liking long sentences or fiction) rather than stable traits such as personality or demographics.

Our results that author personality and gender were associated with topic selection and story writing also suggest that studies could leverage the behavior of participants to predict personality characteristics. While these results are not yet strong enough to provide robust measures of personality or demographics, future studies could explore how to leverage these associations to predict author characteristics in story writing or other writing evaluations rather than relying on questionnaires.

Acknowledgments

This work was partially funded by NSF award 1651487, an NSF graduate research fellowship, and the DARPA CwC program through ARO (W911NF-15-1-0543). We thank the anonymous reviewers and members of the UWNLP community for their helpful feedback. We would also like to thank the LabintheWild volunteers for writing stories and the Mechanical Turk workers for reading them.

References

Tal August, Nigini Oliveira, Chenhao Tan, Noah Smith, and Katharina Reinecke. 2018. Framing effects: Choice of slogans used to advertise online experiments can boost recruitment and lead to sample biases. *Proceedings of the ACM on Human-Computer Interaction*, (CSCW):1–19.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*, pages 329–340.

Martin Conway and Christopher Pleydell-Pearce. 2000. The construction of autobiographical memories in the self-memory system. *Psychology Review*, 107(2):261–288.

Paul T Costa Jr and Robert R McCrae. 2008. *The Revised NEO Personality Inventory (NEO-PI-R)*. Sage Publications, Inc.

Brent Donnellan, Frederick Oswald, Brendan Baird, and Richard Lucas. 2006. The mini-IPIP scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, 18(2):192.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Association of Computational Linguistics*, pages 889–898.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Empirical Methods in Natural Language Processing*, pages 1161–1166.

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jacob B Hirsh and Jordan B Peterson. 2009. Personality and language use in self-narratives. *Journal of Research in Personality*, 43(3):524–527.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, and Larry S Davis. 2017. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7186–7195.

Eunice Jun, Gary Hsieh, and Katharina Reinecke. 2017. Types of motivation affect study selection, attention, and dropouts in online experiments. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–15.

Irwin P Levin, Gary J Gaeth, Judy Schreiber, and Marco Lauriola. 2002. A new look at framing effects: Distribution of effect sizes, individual differences, and independence of types of effects. *Organizational behavior and human decision processes*, 88(1):411–429.

Judith Lorber, Susan A Farrell, et al. 1991. The social construction of gender. *Newbury Park*, 5.

Robert R McCrae. 1987. Creativity, divergent thinking, and openness to experience. *Journal of personality and social psychology*, 52(6):1258.

Christiana D Morgan and Henry A Murray. 1935. A method for investigating fantasies: The thematic apperception test. *Archives of Neurology & Psychiatry*, 34(2):289–306.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593.

James W Pennebaker, Roger J Booth, Ryan L Boyd, and Martha E Francis. 2015. Linguistic inquiry and word count: LIWC 2015.

James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *Public Library of Science (PloS) one*, 9(12).

James W Pennebaker and Janel D Seagal. 1999. Forming a story: the health benefits of narrative. *Journal of Clinical Psychology*, 55(10):1243–1254.

Christopher Purdy, Xinyu Wang, Larry He, and Mark O. Riedl. 2018. Predicting generated story quality with quantitative measures. In *The Artificial Intelligence for Interactive Digital Entertainment Conference*.

Katharina Reinecke and Krzysztof Z Gajos. 2015. Labinthewild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the ACM on Human-Computer Interaction*, pages 1364–1378. ACM.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Association of Computational Linguistics*.

Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Lukasz Dziurzynski, Stephanie Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *Public Library of Science (PloS) one*, 8(9).

Jaime Teevan, Shamsi T Iqbal, and Curtis Von Veh. 2016. Supporting collaborative writing with microtasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2657–2668.

Georgios N Yannakakis and Héctor P Martínez. 2015. Ratings are overrated! *Frontiers in Information and communications technology*, 2:13.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Association for the Advancement of Artificial Intelligence*, volume 33, pages 7378–7385.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373.

Figure 1: Prompts used in the story writing stage of our data collection.

A StoriesInTheWild Collection

We provide additional details about our data collection process, including the image prompts shown to authors (Figure 1) and the writing setups (Figure 2).

A.1 Motivating LabintheWild authors

Since LabintheWild is a volunteer-based crowdsourcing platform, we design our task such that participants can learn about their personality through story writing as a motivation. The study was advertised on the front page of LabintheWild and posted on social media to recruit participants.

Once a participant finishes their story, we compute their personality estimate (using the Five Factor Model) based on their story language. Specifically, we extract their pronoun usage using the pronoun categories in LIWC (Pennebaker et al., 2015), and predict personality scores using the coefficients from Schwartz et al. (2013). At the end of the task, we display their personality predictions along with short descriptions of which trait is the most present in their writing (i.e., the trait whose score has the highest magnitude).

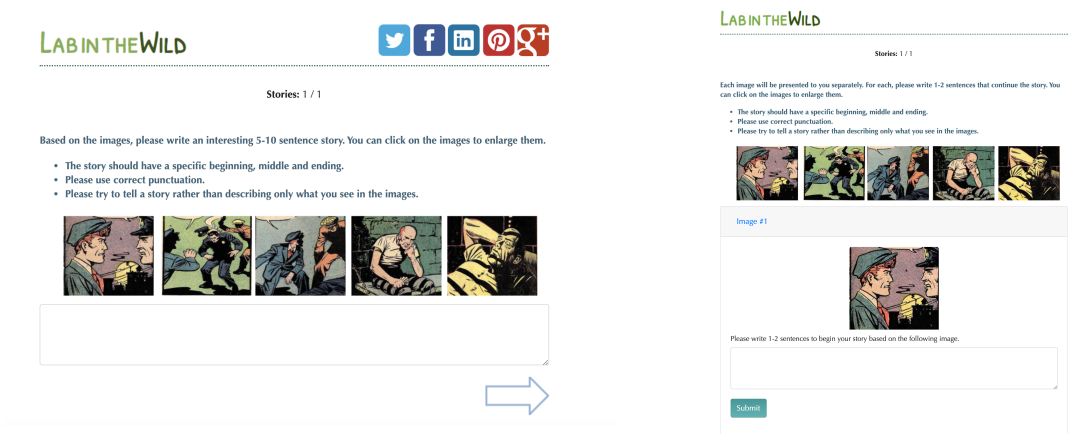
Optionally, participants could take a short personality questionnaire (TIPI; Gosling et al., 2003) before seeing their writing-based personality results. Those who answered these questions could then see their questionnaire-based and their writing-based personality estimates at the end of the task. The end of the task also debriefs participants, explaining the goal of the study and researcher contact information. The debriefing information also includes disclaimers about the personality scores computed from story writing and reiterates that the results should not be used for clinical or diagnostic purposes.

B Analyses

We present further details of our demographic analyses, both between the author demographics and their language use (§B.2) and between the author and reader demographics (§B.3).

B.1 Author demographics and topic choice

To ensure the validity of our other analyses, we examine whether an author's identity was associated with their choosing one of the five topics (Figure 1). We find that only an author's agreeableness affected their choice of image prompt, with highly agreeable authors preferring the dog story (Cohen's $d = 0.30$, $p < 0.001$; Figure 1a) and low agreeableness authors preferring the jail story ($d = 0.41$, $p < 0.001$; Figure 1b). Other demographic variables were comparable for every im-



(a) (b)

Figure 2: Writing interfaces for the crowdsourcing study using the jail cartoon. (a) is all-at-once interface and (b) is the accordion interface. For (b), participants could see all images at the top, but had to write 1-2 sentences about each image separately though an accordian of text boxes.

age prompt (as measured by one-way ANOVAs.)

B.2 Linguistic signal of author demographics

As described in §4.1, we first extract language categories from stories using the LIWC (Pennebaker et al., 2015) lexicon. Then, we use a linear regression model to compute the association between the category and the author's demographics, using z-scored LIWC features for easier interpretation of the regression coefficients (βs).

Our findings, outlined in Table 2, show that an author's identity and personality are somewhat associated with the types of stories they tell (controlling for the type of image prompt they used). Men focused on describing characters (pronoun, social), specifically female characters, whereas women displayed more hierarchical logical storytelling (Analytic; Pennebaker et al., 2014). Controlling for gender, we find that older authors wrote more vivid stories with more emotional tone (Tone, Exclam), more friendship words, and more visual descriptions (percept). In contrast, younger authors wrote in a more past-focused way.

Controlling for age and gender, we find effects of the author's agreeableness and conscientiousness personality traits on the types of language used in stories. We don't see significant effects on the extraversion, openness, or neuroticism scales, likely due to our small sample size of 1.6k (e.g., compared to the 75k users in Schwartz et al., 2013). Shown in Table 2, less conscientious authors wrote more negative stories, whereas more conscientious authors were more positive and focused on character motivations (drives, reward). Less agreeable authors used more swear words.

B.3 Rater and author interaction

As explained in §4.2, we analyze how rater and author traits relate to story ratings. We run linear regression models using story ratings as dependent variables and rater demographics and personality traits as independent variables. We include author demographics and interaction features in these regression models to see if any shared traits across raters and authors were associated with rater preferences. As in all previous analyses, we include story and image type in each model as controlling variables. We report p-values and β coefficients for each regression feature. Full details on the regression results are in Table 3.

	gender β	age β	agreeableness β	conscientiousness β
Analytic	0.228**	*n.s.*	*n.s.*	*n.s.*
Tone	*n.s.*	0.047*	*n.s.*	0.144**
function	-0.192*	*n.s.*	*n.s.*	*n.s.*
pronoun	-0.224**	*n.s.*	*n.s.*	*n.s.*
ppron	-0.292***	*n.s.*	*n.s.*	*n.s.*
you	*n.s.*	0.037*	*n.s.*	*n.s.*
shehe	-0.191†	*n.s.*	*n.s.*	*n.s.*
conj	-0.263***	*n.s.*	*n.s.*	*n.s.*
verb	*n.s.*	-0.034†	*n.s.*	*n.s.*
number	*n.s.*	-0.033†	*n.s.*	*n.s.*
posemo	*n.s.*	0.04*	*n.s.*	*n.s.*
negemo	*n.s.*	*n.s.*	*n.s.*	-0.115†
sad	*n.s.*	-0.035†	*n.s.*	*n.s.*
social	-0.283***	*n.s.*	*n.s.*	*n.s.*
friend	*n.s.*	0.056***	*n.s.*	*n.s.*
female	-0.334***	*n.s.*	*n.s.*	*n.s.*
differ	-0.191*	*n.s.*	-0.097†	*n.s.*
percept	*n.s.*	0.051***	*n.s.*	*n.s.*
see	*n.s.*	0.04**	*n.s.*	*n.s.*
hear	*n.s.*	0.036*	*n.s.*	*n.s.*
drives	*n.s.*	*n.s.*	*n.s.*	0.116*
reward	*n.s.*	*n.s.*	*n.s.*	0.101†
focuspast	-0.184*	-0.038*	*n.s.*	*n.s.*
leisure	*n.s.*	0.036†	*n.s.*	*n.s.*
swear	*n.s.*	*n.s.*	-0.155†	*n.s.*
Exclam	*n.s.*	0.076***	*n.s.*	*n.s.*

Table 2: Results of our LIWC analyses, showing β coefficients between usage of each category with the author's gender, age (gender-controlled), personality (age- and gender-controlled). We additionally control for topic choice. Gender is coded 0 for men, 1 for women. Only results that are significant after applying Holm-correction are shown (*n.s.*:$p > 0.1$;†: $p < 0.1$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$). Extraversion, Opennness, and Neuroticism are omitted since there were no significant correlations for those traits (likely due dearth of data).

Traits	Like	Creative	Coherent	Confusing	Entertaining	Grammatical
Rater Age (45+)	-0.42***	-0.25*	-0.37***	0.27*	-0.39 ***	-0.30 *
Author Age (45+)	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
Rater Age:Author Age	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
Rater Gender (Woman)	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
Author Gender (Woman)	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
Rater Gender:Author Gender	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*n.s.*
Rater Openness	*n.s.*	0.38*	*n.s.*	-0.64***	*n.s.*	*n.s.*
Author Openness	*n.s.*	0.35†	*n.s.*.	*n.s.*	*n.s.*	*n.s.*
Rater Openness:Author Openness	*n.s.*	-0.50†	*n.s.*.	*n.s.*	*n.s.*	*n.s.*

Table 3: Table with regression results for associations between rater and author traits and story ratings. Corrected for multiple hypothesis testing. (*n.s.*:$p > 0.1$;†: $p < 0.1$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$).