"This is a Problem, Don't You Agree?" Framing and Bias in Human Evaluation for Natural Language Generation

Stephanie Schoch* **Diyi Yang†** **Yangfeng Ji***
* Department of Computer Science, University of Virginia, Charlottesville, VA 22904
† College of Computing, Georgia Institute of Technology, Atlanta, GA 30332
{sns2gr,yangfeng}@virginia.edu diyi.yang@cc.gatech.edu

Abstract

Despite recent efforts reviewing current human evaluation practices for natural language generation (NLG) research, the lack of reported question wording and potential for framing effects or cognitive biases influencing results has been widely overlooked. In this opinion paper, we detail three possible framing effects and cognitive biases that could be imposed on human evaluation in NLG. Based on this, we make a call for increased transparency for human evaluation in NLG and propose the concept of human evaluation statements. We make several recommendations for design details to report that could potentially influence results, such as question wording, and suggest that reporting pertinent design details can help increase comparability across studies as well as reproducibility of results.

1 Introduction

Human evaluation is widely considered the gold standard for evaluating natural language generation (NLG), in part because existing automatic metrics display low correlations with human judgments (Belz and Reiter, 2006; Liu et al., 2016; Reiter and Belz, 2009; Novikova et al., 2017). As a result, human evaluation is frequently used to demonstrate state-of-the-art results for generative tasks. However, this has the potential to be problematic due to the lack of consistency in how human evaluation is carried out (Gkatzia and Mahamood, 2015; van der Lee et al., 2019). Beyond producing variability in results, this has implications for validity of human evaluation results due to the influence of evaluation design choices. To address this, a number of papers have proposed recommended best practices for different aspects of NLG human evaluation (Amidei et al., 2019; van der Lee et al., 2019). However, overlooked have been the issues of transparency and the potential for question framing effects and other cognitive biases influencing results.

Cognitive biases refer to heuristics that arise in judgment or decision-making (Tversky and Kahneman, 1974). Framing effects (Tversky and Kahneman, 1981) are types of cognitive biases that refer to *how* something is asked as opposed to *what* is asked. In the context of natural language generation research, these effects refer to the wording of questions asked and accompanying task descriptions and instructions, as opposed to what the target quality is that is being assessed.

In this opinion paper, we demonstrate the lack of transparency in NLG human evaluation through empirically demonstrating the extent to which question wording is not included in evaluation design details, finding that only 15.68% of human evaluation studies in papers we surveyed explicitly reported the actual questions asked. We discuss three types of framing and cognitive biases that could influence results in NLG human evaluation: positive and negative framing, demand characteristics and response bias, and anchoring and adjusting. Using concrete examples from studies in human-computer-interaction and psychology and hypothetical examples for NLG, we demonstrate the importance of including question wording when using human evaluation in NLG. Finally, we propose the concept of "human evaluation statements" and suggest a set of design parameters that should be included pertaining to human evaluation study design.

2 Transparency in Human Evaluation

There is currently no standardized approach or consensus for how human evaluation for NLG should be carried out (Gkatzia and Mahamood, 2015; van der Lee et al., 2019). As a result, it is currently very difficult to compare results across different studies due to the variability in evaluation design. Past efforts to address this have included

Proceedings of the 1st Workshop on Evaluating NLG Evaluation, pages 10–16,
Online (Dublin, Ireland), December 2020.

overviews of evaluation design practices used during a particular time span (Amidei et al., 2018; Gkatzia and Mahamood, 2015; van der Lee et al., 2019) with corresponding recommendations for best practices (van der Lee et al., 2019) and empirical studies or overviews investigating the effects of different question types and scales (Amidei et al., 2019; Novikova et al., 2018). Consistently, these studies have approached variability as a factor impacting the reliability of results.

However, yet to be addressed is the lack of transparency in how studies are designed and reported, which has implications for comparability across studies, as well as replicability and validity of results. While transparency has yet to be addressed in human evaluation, transparency of data, models, and automatic evaluations is a growing topic of concern in the machine learning and natural language processing communities. Bender and Friedman (2018) proposed the usage of "data statements" for mitigating bias and increasing transparency in natural language processing and Gebru et al. (2020) proposed "datasheets for datasets" for increased data transparency and accountability. Transparency in model reporting has also been advocated for. Mitchell et al. (2019) proposed the usage of "model cards" containing model performance characteristics for transparent model reporting. Pertaining to model evaluation, there have been numerous criticisms of task leaderboards (Linzen, 2020; Rogers, 2019) which has led to calls for transparency through reporting of a more informative suite of metrics (Dodge et al., 2019; Ethayarajh and Jurafsky, 2020).

Driving the call for transparency has been the increased attention to issues of reproducibility. Crane (2018) identified a number of controllable environmental settings that are widely unreported in question answering research and demonstrated the impact they have on reproducibility of results, including whether or not a model would be considered state-of-the-art. When we consider the impact of environmental variables (Crane, 2018), computational budget including number of hyperparameter search trials (Dodge et al., 2019), and other factors that can impact results, we can draw comparisons to human evaluation design details that could similarly impact results.

We suggest that the design details of human evaluations can be thought of analogously to model hyperparameters, in that careful tuning can directly influence results. It is currently an open question as to what parameters in human evaluation could influence results, but without reporting pertinent details, we cannot begin to make comparisons across studies, or reproduce results. For example, van der Lee et al. (2019) suggested their findings pertaining to sample sizes and demographics in a survey of 89 papers using human evaluation for NLG may not reflect reality, since only 55% of the papers reported the number of participants and 18% reported demographics. An additional design parameter that we believe is largely unreported but could have an immense impact on results is that of the actual wording of questions presented to participants. More specifically, if questions are framed in ways that elicit various cognitive biases such as framing effects, response biases, or anchoring and adjustment effects, results may reflect question design rather than model performance.

Empirical Analysis To identify the extent to which question wording is unreported in the details of human evaluation for NLG, we collected a set of 81 NLG papers published in ACL ($n = 33$), EMNLP ($n = 30$), NAACL ($n = 11$), and INLG ($n = 7$) in 2019 and 2020, randomly sampled from all papers containing the keyword "generation" in the title.[1] Of these, 51 (62.96%) included human evaluation as a means to assess model performance. However, only 8 of the 51 studies (15.68%) that included human evaluation reported the actual wording and setup of the questions that were asked, either written out ($n = 4$), included as a figure displaying the prompt ($n = 3$), or both ($n = 1$). Question wording does not only have implications for increasing transparency for the purposes of comparability of results across studies, but has further implications for the validity and reproducibility of results. In the following section, we bring attention to the potential of framing effects and other cognitive biases to impact the results of human evaluation for NLG, and use this to make a case for reporting question wording as part of study design.

3 Framing Effects and Cognitive Biases

Framing (Tversky and Kahneman, 1981) refers to *how* something is asked as opposed to *what* is asked. In human evaluation for NLG, this would be reflected in the question wording or instructions provided to participants. In this section, we detail

[1]Data is available at https://github.com/stephanieschoch/framing-bias-nlg-eval

three possible framing effects and cognitive biases that could influence the results of human evalation: positive and negative framing, demand characteristics, and anchoring and adjusting. As question wording is extensively not reported in human evaluation in NLG, rather than providing empirical examples we provide hypothetical examples of the forms these effects could take when question wording is not reported.

3.1 Positive and Negative Framing

Seminal work on the influence of framing in decision-making by Tversky and Kahneman (1981) demonstrated that people are more likely to make choices that are framed positively (in terms of gains) as opposed to negatively (in terms of losses) due to the increased perceived risk associated with choosing potential losses. This effect has been extended and further demonstrated as "loss aversion" in the field of economics (Levin et al., 2002). In our context, the concept of framing based on positive or negative aspects can be extended and viewed as the framing of questions to induce positive or negative priming effects, in which participants are primed to view a choice as having more positive aspects than another, i.e. as the *better* option. For example, if fluency is the target quality in an NLG evaluation, we can consider it the positive aspect.

We demonstrate the potential for the effects of imposing positive or negative framing and priming on questions in NLG human evaluation with the following example: Suppose a researcher is evaluating sentence A from their generative model against sentence B from a baseline model. The researcher asks participants to respond to the question:

"How much more fluent is sentence A versus sentence B?"

Framing in this manner can prime participants to view sentence A as having more positive aspects, in this case, more fluency, as opposed to neutrally framed questions such as *"How do sentence A and sentence B compare in terms of fluency?"*. Positive or negative framing could therefore have a direct impact on the results of the study, in other words, the results could reflect the framing rather than the actual model performance.

3.2 Demand Characteristics

Demand characteristics are response biases that refer to cues in a study design that may reveal a researcher's hypothesis to the participants, resulting in adjusting responses to meet the expectations of the researcher (Orne, 1962). Dell et al. (2012) demonstrated participant response bias due to interviewer demand characteristics in evaluating human-computer interactive systems. Specifically, when participants knew which artifact was developed by an interviewer, they were consistently more likely to report preference for it, even when it was inferior. For human evaluation in NLG, if questions are framed in a way that cues the evaluators as to which output corresponds to the researcher's system, it is probable that similar response bias could be elicited. As an example, in the context of NLG, this could take form as follows:

A researcher has developed style transfer model A to generate formal sentences, and is evaluating sentence A from their generative model against sentence B from a baseline model. Unconsciously aware of model A's artifacts, in this case, as a system that only uses "." as end punctuation, the researcher states 'We consider sentences that end with "." as more formal than sentences that end with "!"' in the task description.

Framing the question in this manner subjects the responses to demand characteristics as the participants are aware of the researcher's expectations that they will rank sentences ending with "." as more formal than sentences with alternative end punctuation. Due to the fact that most studies are conducted via crowdsourcing platforms in which annotators receive compensation for responses, this adds an additional incentive to perform in accordance with the researcher's expectations.

3.3 Anchoring and Adjusting

Anchoring and adjusting is a cognitive bias in which participants anchor their perceptions based on an initial value and adjust subsequent evaluations accordingly (Tversky and Kahneman, 1974). Gehlbach and Barge (2012) demonstrated anchoring and adjustment effects on attitude-opinion questionnaires in which participants insufficiently adjusted responses on adjacent questionnaire items measuring similar constructs, which affected scale reliability. In the context of human evaluation for NLG, we present the following scenario in which we extend the concept of framing to include framing of task description and instructions displayed alongside questions to elicit advantageous anchoring effects:

A researcher has developed style transfer model A to generate formal sentences. As model design is an iterative process, the researcher has seen model A's outputs throughout the model design process. When selecting example formal sentences to include in the evaluation task description and instructions displayed to participants, the researcher inadvertently selects sentences that look similar to the types of outputs generated by model A. These examples become an anchor for participants in evaluating sentences generated by model A and model B.

By unintentionally framing the question instructions in a way that introduces an advantageous anchor, the results could reflect the overall framing and bias that is introduced rather than the objective model performance differences.

4 Human Evaluation Design Statements

Throughout the previous sections, we have provided examples demonstrating the potential question framing that could elicit human evaluation results for NLG that are biased in favor of a particular model. While these examples may at first glance seem implausible and only possible in cases of conscious (explicit) researcher bias in favor of a particular model, it is important to take into consideration the potential for researchers to possess unconscious (implicit) bias whether due to underlying expectations for a model's performance or due to influences of publication bias. During the peer review process reviewers may default to heuristics to simplify the task of review, including rejecting papers where models do not achieve SOTA results (Rogers and Augenstein, 2020). This can implicitly motivate and incentivize researchers to show their model performs best on the gold standard of evaluation for NLG: human evaluation. We use this example to demonstrate the potential for the current lack of evaluation design details, in particular question wording, to leave the door open for results that have been subject to framing effects and bias which threatens the validity of the results.

We draw attention to these effects in an effort to both increase researcher awareness to their own evaluation study design, decrease the potential for questions framed in ways in which results reflect question framing rather than actual model performance, and increase the amount of transparency in human evaluation to aid in study replicability and comparability. We also suggest that the results for studies which do not include exact question wording should be viewed through a skeptical lens *as though they could contain researcher imparted bias that could significantly impact results*. Further, we use our demonstration of the potential for framing effects and biases in question wording as support for a call for transparency in human evaluation for NLG through the inclusion of study design details, which can aid in the development of more robust human evaluation guidelines.

When guidelines exist that can reduce the complexity and time required to design human evaluation studies, they are used. For the evaluation of paraphrase generation, Li et al. (2018) included the human evaluation guidelines they used as an appendix, which have since been adopted by other studies (Qian et al., 2019). This example shows that guidelines for human evaluation have value: guidelines make life easier and people often adopt those that are available. As such, we make the case for increased transparency in human evaluation with respect to design details that could potentially influence results. In an effort to take preliminary steps towards human evaluation guidelines, we propose the concept of "human evaluation design statements" akin to data statements (Bender and Friedman, 2018; Gebru et al., 2020) or model cards (Mitchell et al., 2019). Determining what should be included on such statements will require additional input, perspectives, and empirical evidence. As a preliminary effort, we provide a list of design parameters that we believe could influence results and should therefore be included when describing human evaluation design setup:

Question Design: *Types, Scales, Wording* Basic inclusions pertaining to question design are question type and corresponding scales due to the variability that can arise based on these design decisions (Novikova et al., 2018). Further, as we demonstrated in this paper, question wording also has the potential to influence results. Because of the potential for empirical differences due to how questions are framed, it is imperative to report question wording as part of design details, especially in studies where researchers use human evaluation to claim state-of-the-art performance.

Question Presentation: *Ordering, Questions per Annotator* Ordering effects are influences on results that occur based on the order in which a sequence of questions is presented (Strack, 1992).

As such, reporting question presentation order or balancing increases transparency as well as study comparability and reproducibility. In addition to ordering effects, response fatigue can occur when the quality and integrity of evaluations degrades as participants tire of a task (Lavrakas, 2008). Due to the possibility of response fatigue effects, statistics regarding the number of questions per annotator should be reported to increase design transparency in terms of potential influences on variability in results.

Target Criteria: *Definitions* It makes intuitive sense that what is actually being measured in human evaluation would influence results, and further that measuring the same or different target criteria in different studies would impact the comparability of the results. However, naming conventions and definitions are inconsistent and may exhibit significant overlap, such as with naturalness, grammaticality, and fluency (Mir et al., 2019; Novikova et al., 2018). As such, what is being measured should be compared across studies based on definition and the resulting participant understanding of the task, rather than simply based on naming convention: studies may measure the same aspect under different names or different aspects under the same name. Studies consistently reporting this detail in human evaluation is also a preliminary step towards agreed upon task definitions.

Annotators: *Demographics, Background, Recruitment, Compensation* Understanding and reporting the details of the *human* factor in human evaluation is intuitively one of the most important sets of details to include in terms of transparency and potential influence on results. Inclusions involve who annotators are in terms of demographics and background, how they were recruited, and whether or not annotators received fair compensation (Silberman et al., 2018). As an example impact, annotator familiarity with the target language for a task might largely influence judgments towards biases, fluency, or grammatical correctness. The human factor in human evaluation, our annotators, is central to and interacts with every other detail of study design, and is therefore vital to report.

While this list is not comprehensive, we believe these design details could have influences on evaluation results, and as such, are important details to consider and include.

5 Other Considerations

One of the factors that could limit the potential for widespread adoption of human evaluation statements that include human evaluation design details is the page limits imposed for many journal and conference papers. One approach to combat this is to include the details of human evaluation in Supplementary Material that accompanies papers. However, we suggest that many details in human evaluation design are central to understanding the meaningfulness of results, and further suggest that there will need to be community agreed upon guidelines for what details must be included within main papers. We further suggest that a complementary strategy would be the eventual development of comprehensive, agreed upon human evaluation guidelines that could operate similarly to "long-form" data statements (Bender and Friedman, 2018). In this scenario, guidelines could be referenced, summarized briefly, and appended with pertinent additional study details as was proposed with "short-form" data statements (Bender and Friedman, 2018).

6 Conclusion

In this paper, we demonstrate the extent to which including the details of human evaluation is limited in natural language generation. We further demonstrate the need for including design details such as question wording using existing work in psychology and human-computer interaction on framing and cognitive biases, and cite the recent push for transparency with datasets and model details, such as details of hyperparameter tuning, as support for similar efforts to increase transparency in human evaluation. Based on these observations, we propose working towards human evaluation statements and make several suggested inclusions, while noting the future need for additional perspectives and direct empirical support.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 307–317, Tilburg University, The Netherlands. Association for Computational Linguistics.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. The use of rating and Likert scales in natural lan-

guage generation human evaluation tasks: A review and some recommendations. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 397–402, Tokyo, Japan. Association for Computational Linguistics.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Matt Crane. 2018. Questionable answers in question answering research: Reproducibility and variability of published results. *Transactions of the Association for Computational Linguistics*, 6:241–252.

Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. 2012. "yours is better!": Participant response bias in hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1321–1330, New York, NY, USA. Association for Computing Machinery.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. *arXiv preprint arXiv:2009.13888*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2020. Datasheets for datasets.

Hunter Gehlbach and Scott Barge. 2012. Anchoring and adjusting in questionnaire responses. *Basic and Applied Social Psychology*, 34(5):417–433.

Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005 - 2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 57–60, Brighton, UK. Association for Computational Linguistics.

Paul J Lavrakas. 2008. *Encyclopedia of survey research methods*. Sage Publications.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Irwin P Levin, Judy Schreiber, Marco Lauriola, and Gary J Gaeth. 2002. A tale of two pizzas: Building up from a basic product versus scaling down from a fully-loaded product. *Marketing Letters*, 13(4):335–344.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018. Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878, Brussels, Belgium. Association for Computational Linguistics.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural

language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Martin T Orne. 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist*, 17(11):776.

Lihua Qian, Lin Qiu, Weinan Zhang, Xin Jiang, and Yong Yu. 2019. Exploring diverse expressions for paraphrase generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3173–3182, Hong Kong, China. Association for Computational Linguistics.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Anna Rogers. 2019. How the transformers broke nlp leaderboards.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in nlp?

M Six Silberman, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar. 2018. Responsible research with crowds: pay crowdworkers at least minimum wage. *Communications of the ACM*, 61(3):39–41.

Fritz Strack. 1992. "order effects" in survey research: Activation and information functions of preceding questions. In *Context effects in social and psychological research*, pages 23–34. Springer.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.