# Unveiling NLG Human-Evaluation Reproducibility: Lessons Learned and Key Insights from Participating in the ReproNLP Challenge

**Lewis Watson** and **Dimitra Gkatzia**
School of Computing, Engineering & The Built Environment
Edinburgh Napier University, UK
{l.watson, d.gkatzia}@napier.ac.uk

## Abstract

Human evaluation is crucial for NLG systems as it provides a reliable assessment of the quality, effectiveness, and utility of generated language outputs. However, concerns about the reproducibility of such evaluations have emerged, casting doubt on the reliability and generalisability of reported results. In this paper, we present the findings of a reproducibility study on a data-to-text system, conducted under two conditions: (1) replicating the original setup as closely as possible with evaluators from AMT, and (2) replicating the original human evaluation but this time, utilising evaluators with a background in academia. Our experiments show that there is a loss of statistical significance between the original and reproduction studies, i.e. the human evaluation results are not reproducible. In addition, we found that employing local participants led to more robust results. We finally discuss lessons learned, addressing the challenges and best practices for ensuring reproducibility in NLG human evaluations.

## 1 Introduction

Human evaluations have long been considered the appropriate method for reliably evaluating NLG systems, due to the shortcomings of automatic evaluation metrics (Gehrmann et al., 2022). Notwithstanding the popularity and acceptance of human evaluations in the NLG community, the reproducibility of human evaluations has not been thoroughly documented. It is widely accepted that replicating human evaluation results can be really hard due to insufficient documentation (Belz et al., 2023a,b), variability in the generated text that leads to varying assessment results due to evaluator preferences, background, or other characteristics (Gkatzia et al., 2014, 2016), confusion and diversity in defining evaluation criteria (Howcroft et al., 2020), high costs (Thomson and Reiter, 2021) or even difficulty in identifying factual errors (Thomson et al., 2023).

The *ReproGen/ReproNLP challenges* led by Belz et al. (2020) aim to address the lack of understanding surrounding the reproducibility of human evaluations in NLG research by facilitating and encouraging research on the reproducibility of current evaluation methods, and the factors leading to irreproducibility. It involves a global multi-lab shared task study aimed at assessing the reproducibility of human evaluations conducted in selected published NLG research papers. The challenge organisers initially selected studies to be reproduced and allocated them to each participating lab. Each lab is then tasked with reproducing the results of the allocated human evaluation study, allowing for an assessment of the reproducibility of human evaluations across different methods, tasks, and participant characteristics. Labs were also allowed to explore additional research questions or perform further analyses of the results.

Our lab was tasked with reproducing one of the human evaluations reported in Puduppully and Lapata (2021), which aims to identify supporting and contradictory facts in text grounded on tabular data[1], namely basketball and baseball reports generated from game statistics tables. Here, we explore two research questions: (1) whether we can reproduce a human evaluation study on Amazon Mechanical Turk (AMT) following as closely as possible the original design as presented by the authors of the study; and (2) explore the impact of using local evaluators instead of AMT evaluators for this task. Our contributions are as follows:

- We present the results from our effort to reproduce the human evaluation presented by Puduppully and Lapata (2021).

[1] The paper presents two human evaluation studies - here we only reproduce the first, another lab is tasked with reproducing the second.

- We present results from an additional study that draws evaluators from a pool of colleagues and students with experience in AI and we demonstrate that using local evaluators results in more robust results (as measured through Inter-Annotator Agreement).
- We discuss the implications of our results to NLG evaluation studies.

The rest of the paper is shaped as follows: Section 2 presents the original human evaluation study, Section 3 presents our effort to reproduce the original study, Section 4 discusses our results in comparison to the original study, and Section 5 discusses an additional study we performed with local evaluators, and finally, Section 6 discusses the results and implications for reproducibility studies in NLG.

## 2 Original Study

The original study selected for our reproduction is "Data-to-text Generation with Macro Planning" by Puduppully and Lapata (2021) which proposed a new macro planning phase for data-to-text generation. This new phase aims to enhance the structure and accuracy of the generated content by emphasising higher-level content organisation, including entities, events, and their interactions. These high-level features, termed macro plans, are learned from the provided data and are then used as inputs to guide text generation. The authors employed both automatic and human evaluations to obtain accurate assessments of their model's performance. The human evaluations conducted in the original study compared the model's performance on two datasets: MLB (MLB dataset consisting of baseball games' box line-score tables, and play-by-play tables) (Puduppully et al., 2019) and RotoWire (RotoWire dataset consisting of NBA basketball games' box and line-score boxes) (Wiseman et al., 2017). The model's performance on both datasets was compared to four different NLG systems: Gold, Template (Template-based generators from Wiseman et al. (2017)), ED+CC (encoder-decoder with attention and copy mechanism), and ENT (Entity-based model) for the MLB dataset, and RBF-2020 for the RotoWire dataset.

The original paper reports two human evaluation studies: a fact-counting study and the quality of generated summaries. Here, we reproduce the **fact-counting study**. In this study, human evaluators were asked to count the number of supporting and contradicting facts in the outputs of the NLG systems by comparing them with the input data. For the MLB dataset, the input consisted of a baseball game box, line-score, and play-by-play tables, while for the RotoWire dataset, participants were provided with an NBA basketball game box- and line-score tables.

The fact-counting study involved a total of 600 evaluations or Human Intelligence Tasks (HITs) that required human evaluators. To facilitate these evaluations, the authors utilised Amazon Mechanical Turk (MTurk) to crowdsource the completion of the HITs. Specific qualifications were set for workers to be eligible to participate, including having an MTurk approval rating greater than 98%, a minimum of 1000 previously completed HITs on MTurk, and being based in one of the following English-speaking countries: US, UK, Canada, Ireland, Australia, or New Zealand.

In the original study, the authors reported that human evaluators were not required to have prior knowledge of basketball or baseball, as they were provided with a cheatsheet explaining the semantics of the box score tables. Each summary was evaluated by three different workers, and there were a total of 131 distinct MTurk workers involved in the evaluations. The 600 HITs were divided into eight mini-batches (four per dataset), and attention checks were employed to ensure the quality of the responses. If a worker reported more than 20 total facts, their response was rejected and rerun. The agreement among the three responses for each distinct HIT was calculated using Krippendorff's alpha, resulting in 0.44 for supported facts and 0.42 for contradicting facts.

## 3 Reproduced Study

In the reproduced study, we followed the design and methodology of the original study as closely as possible, however, we only reproduced the tasks for the RotoWire dataset due to the less complex task and cheat-sheets provided, which allowed for better control over the cognitive complexity of the HITs. This decision aligned with the recommendations by Belz et al. (2023a) in controlling factors such as the number of evaluators, the cognitive complexity of the task, and the level of training/expertise of the evaluators. By narrowing down these factors, our goal was to improve the accuracy and effectiveness of reproducing the results.

We obtained the necessary model outputs and

| Original Study - Rotowire | | |
|---|---|---|
| System | #Supp | #Contra |
| Gold | 3.63 | 0.07 |
| Templ | 7.57* | 0.08 |
| ED+CC | 3.92 | 0.91* |
| RBF-2020 | 5.08* | 0.67* |
| Macro | 4.00 | 0.27 |

| Reproduction Study - Rotowire | | |
|---|---|---|
| System | #Supp | #Contra |
| Gold | 4.000 | 1.525 |
| Template | 6.3167* | 1.3583 |
| ED+CC | 5.100 | 1.9042 |
| RBF-2020 | 4.9458 | 1.7583 |
| Macro | 4.5458 | 1.5333 |

Table 1: Mean counts of supported (#Supp) and contradicting (#Contra) facts in game summaries (one-way ANOVA with posthoc Tukey HSD tests, * denotes significance with $p \leq 0.05$, when comparing each result to Macro).

Human Evaluation Datasheet (HEDs) from the original authors and filled out our own HEDs for reproduction. Then using the Amazon Mechanical Turk (MTurk) platform, we set up HITs for the fact-counting evaluation. Workers were asked to count the supporting and contradicting facts in summaries generated by different NLG systems, using the exact same UI and cheatsheet as the original study. We conducted multiple mini-batches of HITs with attention checks in between.

A total of 167 distinct workers participated in the study, and we ran 300 HITs divided into 4 batches (agreement using Krippendorff's $\alpha$ was -0.12 for supported and 0.12 for contradicting facts, as opposed to 0.44 and 0.42 respectively in the original study). We reran several batches where workers had failed the attention checks outlined in the original paper, resulting in 121 failed attention check tasks. Including paying for reruns, the total cost for the study came to $665.18. Workers received compensation at UK's Living Wage[2] level.

To analyse the results, we mirrored the original study by using the exact same code for statistical analysis using a one-way ANOVA test with post-hoc Tukey HSD test. This analysis helped us identify significant differences in the performance of the NLG systems for comparison with the proposed macro system.

## 4 Results

In this section, we compare the results from the fact-counting HIT on the RotoWire dataset for both the original and reproduced studies (see Table 1).

In the original study, the template-based generator (Template) showed a statistically significant higher number of reported supporting facts (7.57) when compared to the Macro system (4.00). Similarly, the RBF-2020 system showed a statistically significant increase in supporting facts (5.08). In terms of contradicting facts, ED+CC and RBF-2020 reported higher numbers, with statistical significance underlined at 0.91 and 0.67 respectively.

The reproduced study results demonstrated a different pattern. The Template system once again recorded a statistically significant higher number of supporting facts (when compared to the macro system) at 6.3167, but the differences in the contradicting facts were less pronounced across the systems, without statistical significance against macro.

### 4.1 Comparative Analysis

Comparing both studies, it is evident that there are inconsistencies between the original and reproduced results. While the Template system consistently showed a higher number of supporting facts in both studies, the magnitude of this difference was reduced in the reproduction. The number of contradicting facts, in particular, exhibited a notable increase in the reproduced study across all systems.

## 5 Additional study with local evaluators

In addition to the above reproduction study, we conducted a supplementary study **on a smaller scale** with a selected pool of academic evaluators. This study aimed to provide further insights into the reproducibility of human evaluations as well as the impact of sampling participants with different characteristics.

To carry out this additional study, we adapted the HTML task interface to work locally without relying on the MTurk platform. The interface was modified to allow participants to save their answers to a JSON file, which they would then email back to us. For each of the five NLG systems, two tasks were randomly selected from both the RotoWire

[2] https://www.livingwage.org.uk/what-real-living-wage

and MLB datasets (total of 20 HITS). Unlike the large-scale reproduction, each task in the additional study was completed by two distinct participants instead of three.

To gather respondents for the study, we allocated the individual HIT HTML files to participants and asked them to submit the JSON file once they completed the task(s). The participants were instructed to count the number of supporting and contradicting facts in the summary, following the same approach as the original and large-scale reproduction studies.

A total of 17 distinct evaluators, including 4 non-native English speakers (who however live and work/study in the UK), participated in the additional study (agreement using Krippendorff's $\alpha$ was 0.65 for supported and 0.56 for contradicting facts). We ran 40 HITs in total, with each task having two participants. There were no failed attention checks in this study. The collected responses from the additional study were analysed using the same statistical analysis Python script used for the original and large-scale reproduction studies. This allowed us to compare the performance of the different NLG systems in the additional study as well.

## 5.1 Results

### 5.1.1 RotoWire Dataset

The results of the additional study on RotoWire are shown in Table 2.

| Additional Study - RotoWire | | |
|---|---|---|
| System | #Supp | #Contra |
| Gold | 6.9375* | 0.0625 |
| Template | 4.125 | 0.25 |
| ED+CC | 5.0625* | 0.5625 |
| RBF-2020 | 5.5* | 0.0 |
| Macro | 2.625 | 0.125 |

Table 2: Mean counts of supported (#Supp) and contradicting (#Contra) facts in game summaries (one-way ANOVA with posthoc Tukey HSD tests, * denotes significance with $p \leq 0.05$, when comparing each result to Macro).

Compared with the original study shown in Table 1, the Gold standard reported statistically higher supporting facts and maintained low contradicting facts. The Template system showed a decrease in supporting facts however, lost statistical significance in the additional study. Compared to the original study, ED+CC displayed an increase in supporting facts and gained statistical significance compared to the macro system. The RBF-2020 system maintained statistical significance against the macro system found in the original study. **We should note, however, that this additional study evaluates a smaller pool of system outputs and therefore there is an expected natural discrepancy in the number of supporting and contradictory facts. As such, the results should not be interpreted as definitive indicators of individual system performance.** However, when looking at the Inter-Annotator Agreement, we see that the participants in the local study score higher than the AMT participants, indicating that the results are more robust.

### 5.1.2 MLB Dataset

In addition to the RotoWire dataset, the supplementary study also evaluated task agreeability using the MLB dataset. The results are summarised in Table 3.

The Gold system exhibited an increase in supporting facts and a higher number of contradicting facts. The Template system reported a decrease in both supporting and contradicting facts, while ED+CC showed an increase in supporting facts with lower contradicting facts. The ENT system displayed lower supporting facts but higher contradicting facts, whereas the Macro system maintained similar levels. Similarly to the previous experiment, the outputs evaluated here are a subset of the ones used in the original study.

## 5.2 Feedback Insights

Feedback received from participants unveiled other critical aspects that might have impacted the studies. Many disagreed with the notion that prior knowledge of basketball or baseball was unnecessary, leading to confusion and the need to look up specific phrases. Some suggested layout changes to minimise scrolling, while others were unclear about what qualified as a "fact." Interestingly, unnecessary feedback was common in the larger study, possibly due to different incentives for paid workers trying to quickly fill out tasks versus unpaid student and academic participants - this is supported by the absence of failed attention checks in the additional study. The smaller sample size in the local study could also be argued as an explanation for the absence of such feedback, although it's unlikely to be the sole reason.

| Original Study - MLB | | |
|---|---|---|
| System | #Supp | #Contra |
| Gold | 3.59 | 0.14 |
| Templ | 4.21 | 0.04 |
| ED+CC | 3.42 | 0.72* |
| ENT | 3.71 | 0.73* |
| Macro | 3.76 | 0.25 |

| Additional Study - MLB | | |
|---|---|---|
| System | #Supp | #Contra |
| Gold | 4.375 | 0.75 |
| Template | 2.6875 | 0.5 |
| ED+CC | 4.875 | 0.25 |
| ENT | 2.875 | 0.875 |
| Macro | 3.0 | 0.8125 |

Table 3: Mean counts of supported (#Supp) and contradicting (#Contra) facts in game summaries (one-way ANOVA with posthoc Tukey HSD tests, * denotes significance with $p \leq 0.05$, when comparing each result to Macro).

## 6 Discussion & Conclusions

The attempt to reproduce the results of the original study yielded mixed outcomes, with substantial differences observed in the reproduction studies. While the original study showcased certain statistical significance in the reported performance of the systems against the macro system, this significance was often lost in reproduction studies, particularly concerning the number of contradicting facts.

The full, large-scale reproduction exhibited a noticeable increase in the number of contradicting facts across various systems, and the alignment between the original and reproduced studies was limited. Strikingly, the local study displayed more consistency with the original study but also brought forth its unique variations. As expected the local study resulted in higher annotator agreement than the AMT study.

Across different NLG systems, there was a clear fluctuation in the number of reported supporting and contradicting facts. This variation, although intriguing, added to the complexity of drawing definitive conclusions regarding the reproducibility of human evaluations in these contexts.

### 6.1 Contributing Factors

Several factors emerged as potential contributors to the observed discrepancies between the studies. Differences in evaluator opinions, missing information, and the evaluators' understanding of the task likely played significant roles in the outcomes. Additionally, inconsistencies in the evaluation criteria, the make-up of the evaluator pool, biases in the evaluation process, and the inherent subjectivity of human judgement cannot be overlooked as influencing factors.

The local study, being conducted in a more controlled environment, and with an evaluator pool where incentives are better aligned and not tied to financial gain, may have mitigated some of these confounding variables, showing more consistency with the original study. However, the human-centric nature of the evaluations leaves room for unpredictable variations.

### 6.2 Moving Forward

The findings of this research underscore the intricate nature of human evaluations and the challenges in reproducing such studies. While the reproduction attempt was not entirely successful, the insights gleaned from the process are invaluable.

Future work should aim to incorporate these insights, focusing on minimising biases, clarifying evaluation criteria, and possibly developing standardised protocols for human evaluations. The collaboration between AI and human judgement must be tuned, recognising the complex interaction between objectivity and subjectivity, to advance the field in a meaningful and responsible manner.

## Acknowledgements

Dimitra Gkatzia's work is supported under the EPSRC projects NLG for low-resource domains (EP/T024917/1) and CiViL (EP/T014598/1).

## References

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2020. ReproGen: Proposal for a shared task on reproducibility of human evaluations in NLG. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 232–236, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Jackie Cheung, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Huiyuan Lai, Chris van der Lee, Emiel van Miltenburg, Yiru Li, Saad Mahamood, Margot Mieskes, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Pablo Mosteiro Romero, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023a. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in nlp.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023b. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *ArXiv*, abs/2202.06935.

Dimitra Gkatzia, Helen Hastie, and Oliver Lemon. 2014. Finding middle ground? multi-objective natural language generation from time-series data. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 210–214, Gothenburg, Sweden. Association for Computational Linguistics.

Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. 2016. Natural language generation enhances human decision-making with uncertain information. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 264–268, Berlin, Germany. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2023–2035, Florence, Italy. Association for Computational Linguistics.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *CoRR*, abs/2102.02723.

Craig Thomson and Ehud Reiter. 2021. Generation challenges: Results of the accuracy evaluation shared task. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 240–248, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech  Language*, 80:101482.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.