Computer Speech & Language 67 (2021) 101151

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Human evaluation of automatically generated text: Current trends and best practice guidelines

Chris van der Lee[*,1,a], Albert Gatt[2,b], Emiel van Miltenburg[3,a], Emiel Krahmer[4,a]

[a] Tilburg Center for Cognition and Communication, School of Humanities and Digital Sciences, Tilburg University, Tilburg, the Netherlands
[b] Institute of Linguistics and Language Technology, University of Malta, Msida, Malta

ARTICLE INFO

Article History:
Received 13 March 2020
Revised 2 August 2020
Accepted 5 September 2020
Available online 22 November 2020

Keywords:
Natural Language Generation
Human evaluation
Recommendations
Literature review
Open science
Ethics

ABSTRACT

Currently, there is little agreement as to how Natural Language Generation (NLG) systems should be evaluated, with a particularly high degree of variation in the way that human evaluation is carried out. This paper provides an overview of how (mostly intrinsic) human evaluation is currently conducted and presents a set of best practices, grounded in the literature. These best practices are also linked to the stages that researchers go through when conducting an evaluation research (planning stage; execution and release stage), and the specific steps in these stages. With this paper, we hope to contribute to the quality and consistency of human evaluations in NLG.

1. Introduction

Even though automatic text generation has a long tradition, going back at least to Peter (1677) from the 17th century (see also Swift, 1774; Rodgers, 2017), there are ongoing debates on the best way to evaluate the output of such systems and especially, on how evaluation involving human participants should best be carried out. Consensus on these issues is important for the development of Natural Language Generation (NLG) systems. With a well-executed evaluation it is possible to assess the quality of a system and its properties, and to demonstrate the progress that has been made on a task. Valid and reliable evaluation practices involving human participants can also help to get a better understanding of the current state of the field (Mellish and Dale, 1998; Gkatzia and Mahamood, 2015; van der Lee et al., 2018).

This paper provides an overview of current practices in human evaluation (with a focus on intrinsic human evaluation), showing that there is no consensus as to how NLG systems should be evaluated. As a result, it is hard to compare results published by different authors, and it is difficult for newcomers to the field to identify which approach to take for evaluation. This paper addresses these issues by providing a set of best practices for (intrinsic) human evaluation in NLG. A further motivation for this paper's focus on human evaluation is the recent discussion on the (un)suitability of automatic measures for the evaluation of NLG systems (see Ananthakrishnan et al., 2007; Novikova et al., 2017; Sulem et al., 2018; Reiter, 2018, and the discussion in Section 2.

*Corresponding author.
E-mail addresses: c.vdrlee@uvt.nl, c.vdrlee@tilburguniversity.edu (C. van der Lee).
[1] orcid=0000-0003-3454-026X
[2] orcid=0000-0001-6388-8244
[3] orcid=0000-0002-7143-8961
[4] orcid=0000-0002-6304-7549

https://doi.org/10.1016/j.csl.2020.101151

Some previous studies have provided overviews of evaluation methods. Gkatzia and Mahamood (2015) focused on NLG papers describing end-to-end systems from 2005 to 2014; Amidei et al. (2018a) provided a 2013−2018 overview of evaluation in question generation; and Gatt and Krahmer (2018) provided a more general survey of the state-of-the-art in NLG, including evaluation. However, the aim of these papers was to give a structured overview of existing methods, rather than discuss shortcomings and best practices. Moreover, they did not focus on human evaluation. Shortcomings and best practices were discussed by Dror et al. (2018), focusing on appropriate statistical tests for common research designs in NLP; Dodge et al. (2019), focusing on how to accompany experimental results with details on the system, the dataset, and the hyperparameters; and Geiger et al. (2020), focusing on best practices for human labeling of training datasets in machine learning. The current paper can be viewed as complementary to these papers, with its focus mostly on a different human evaluation task (rating texts, rather than classifying them), on different details that should be reported on when presenting experimental results (other than details about the system, dataset, and hyperparameters), and on different considerations that come into play in the development of a human evaluation study (other than statistical test selection).

In this paper, we also initially adopt a bibliometric approach to gain a sense of the state of the art in human evaluation. Following Gkatzia and Mahamood (2015), Section 3 provides an overview of current evaluation practices, based on papers from the International Conference on Natural Language Generation (INLG) and the Annual Meeting of the Association for Computational Linguistics (ACL), in 2018 and 2019. These conferences were chosen on the grounds that INLG is the primary international event which is entirely focussed on NLG, while ACL is the premier conference in the field of Computational Linguistics and Natural Language Processing, with a substantial annual track on NLG and summarisation. We study the last two years in order to maintain a focus on trends in the evaluation practices in current NLG, which is dominated by neural methods, and where most papers are highly technical. Consider, for instance, the fact that the proportion of papers reporting neural approaches increased from 32% in 2016 to 81% in 2018 and to 90% in 2019 (see Table 1). Apart from the broad range of methods used, we observe that intrinsic evaluation is the dominant type of evaluation. This may be due to the higher time and resource investments necessary to execute an extrinsic evaluation. In any case, the emphasis on intrinsic methods adds further justification to the focus on these in the present paper, since it is clear that best practise guidelines are called for where such methods are concerned.

Building on findings from NLG, but also statistics and the behavioral sciences, Section 4 provides a set of recommendations and best practices for human evaluation in NLG, focusing on questions that may arise at every step of the design of an evaluation study.

The present paper is a considerable development of work initially presented in van der Lee et al. (2019) and seeks to make a dual contribution: On the one hand, we hope that the recommendations made can serve as a guide for newcomers to the field; on the other, as the debate on evaluation methods in NLG continues, we also hope to contribute towards further standardisation in the way human evaluation is carried out.

1.1. What is Natural Language Generation?

Before discussing the challenges of Natural Language Generation, it is useful to frame this topic. Gatt and Krahmer (2018) define NLG as: “the task of generating text or speech from non-linguistic input”, but NLG can also cover text or speech generated from linguistic input (e.g. summarization, text simplification, etc.). This is a broad definition and illustrates that NLG can include multiple subtasks and can involve different modalities. The current paper will centre on textual output (as opposed, say, to speech, or to output with mixed modalities, such as text and graphics), as most NLG systems focus on this modality at present. More specifically, the current paper concerns all NLG tasks that output human-readable phrases, sentences or texts, regardless of input (e.g., data-to-text generation, style transfer, image description generation, dialogue generation, summarization, etc.). Thus, we exclude tasks that result in single word output, or output that is not human language (e.g. reading comprehension, slot filling, meaning representation extraction, programming code generation, etc.) as they generally require a different kind of evaluation. Some practitioners might also consider Machine Translation and Summarisation to fall under the umbrella of NLG, but others might argue that they do not overlap much with NLG anymore, insofar as they have developed their own methodologies, also in terms of evaluation. While we believe that many of the recommendations may also be relevant to the evaluation of Machine Translation and Summarisation systems, we did not specifically design our recommendations with these fields in mind.

Table 1
Proportion of neural, statistical, and rule-based architectures, and proportion of automatic and human metrics per year. Based on INLG, and NLG tracks in ACL papers published between 2016−2019.

Year	% Neur.	% Stat.	% R-B	% Aut.	% Human
2016	32	46	21	71	46
2017	48	33	20	76	37
2018	81	5	13	84	51
2019	90	6	4	95	49

Although the focus and examples are constrained, we believe that all sections may hold some relevance for every researcher involved in text generation-related research (including Machine Translation, and theory-oriented NLG for which construction of practical systems is not the end goal), but also for NLP researchers, or even empirical researchers in general. All steps necessary for conducting human evaluation research on automatically generated texts (see Table 5 for a summary of these steps) can be more broadly applied to various types of research. For instance, establishing an appropriate goal of the study, choosing the right design/sample/ statistical approach, and ensuring that the study is conducted in an ethical and user-friendly way, are important issues for every empirical study with human participants.

2. Automatic versus human evaluation

Automatic metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004) are well-established nowadays. A survey of empirical papers presented at the ACL tracks on NLG and at the INLG conference from 2016 to 2019 (see Tables 1 and 2) shows that the number of papers that report on automatic metrics has been steadily increasing, from 71% in 2016 to 95% in 2019. About half of these papers (48%) evaluate their system using only automatic metrics. However, the use of these metrics for the assessment of a system's quality is controversial, and has been criticized for a variety of reasons. The two main points of criticism are:

Automatic metrics are underinformative. Text generation can go wrong in different ways while still receiving the same scores on automated metrics. Furthermore, low scores can be caused by correct, but unexpected verbalizations (Ananthakrishnan et al., 2007). Identifying what can be improved in the wake of a metric-based evaluation therefore requires an error analysis. Most automatic metric scores also have reliability issues because it is unclear how stable the reported scores are. With BLEU, for instance, libraries often have their own implementations, which may differ, thereby affecting the scores (this was recently addressed by Post, 2018). Reporting the scores accompanied by confidence intervals, calculated using bootstrap resampling (Koehn, 2004), may increase the interpretability of the results. However, such statistical tests are not straightforward to perform. Finally, one may wonder whether differences in automatic scores actually reflect observed differences by humans. Most experimental papers achieve only a relatively small increase in BLEU scores of around $1-2$ points compared to other systems. Mathur et al. (2020) recently found for Machine Translation systems that a BLEU score increase of such magnitude only corresponds with true improvements, as observed by human judges, about half the time.

Automatic metrics do not correlate well with human evaluations. This has been repeatedly observed (e.g., Belz and Reiter, 2006; Reiter and Belz, 2009; Novikova et al., 2017; Ma et al., 2019).[5][6] In light of this criticism, it has been argued that automated metrics are not suitable to assess linguistic properties (Scott and Moore, 2007), and Reiter (2018) discouraged the use of automatic metrics as a (primary) evaluation metric. There are arguably still good reasons to use automatic metrics: they are a cheap, quick and repeatable way to approximate text quality (Reiter and Belz, 2009), and they can be useful for error analysis and system development (Novikova et al., 2017). In this context, Reiter, 2020 recently made a useful terminological distinction between *scoring functions* and *evaluation metrics*. Some measures, like BLEU, may be considered a useful scoring function (for quick-and-dirty feedback), but not a proper evaluation metric.

Overall, it seems that the most popular automatic metrics are too simple and straightforward: with some variations, they mostly rely on N-gram overlap. They may therefore not reward sentences that fulfill the intended communicative purpose if these sentences also deviate lexically, semantically, or syntactically from the reference sentence. To tackle these problems, researchers have begun development on new metrics that increasingly utilize neural networks. See, for instance, the WMT Metrics Shared Task, which in recent years has been dominated by learning-based approaches (Bojar et al., 2017; Ma et al., 2018; Barrault et al., 2019). Furthermore, learning-based metrics such as RUSE (Shimanaka et al., 2018), BertScore (Zhang et al., 2020), MoverScore (Zhao et al., 2019), and BLEURT (Sellam et al., 2020) are quickly gaining traction in NLG research (e.g., Chen et al., 2019; Han et al., 2019; Wang et al., 2019; Lan et al., 2019; Ribeiro et al., 2020). While these metrics are consistently outperforming well-established metrics such as BLEU and ROUGE, their highest scores still indicate only moderate agreement with humans at the moment. Furthermore, being able to train learning-based metrics on human ratings obtained from a well-executed evaluation could help improve the quality of these metrics (see also Welty et al., 2019).

It is not the purpose of this paper to recommend using human evaluation for every step of the development process, since this would be costly and time-consuming. Furthermore, there may be automatic metrics that reliably capture some aspects of NLG output. For instance, unique n-grams and Shannon entropy (Shannon and Weaver, 1963) are established measures to assess lexical diversity (though note that complex constructs like diversity cannot be fully captured by a single, explainable metric (van Miltenburg et al., 2018). There are also situations where metrics are exactly what you need. Maximum length calculations are an essential metric for Twitter feed generation to establish if the generated texts would be valid tweets, for instance, and

[5] In theory this correlation might increase when more reference texts are used, since this allows for more variety in the generated texts. However, in contrast to what this theory would predict, both Doddington (2002) and Turian et al. (2003) report that correlations between metrics and human judgments in machine translation do not improve substantially as the number of reference texts increases. Similarly, Choshen and Abend (2018) found that reliability issues in reference-based evaluation which arise due to low-coverage reference sets cannot be overcome by increasing the number of references. The only dissonant voice (that we are aware of) is that of Vedantam et al. (2015), who show for the domain of automatic image description, that accuracy of most automatic metrics goes up until there are about 50 reference descriptions per image. At this point, saturation is reached for the average image (but we might still see improvements for images that elicit more varied responses).

[6] It should also be noted that correlations between automatic metrics and human evaluation can be the result of outlier systems whose output quality is markedly higher or lower than that of other systems (Mathur et al., 2020).

Table 2
Proportion of automatic, intrinsic human, and extrinsic human evaluation. And the proportion of automatic evaluation only, human evaluation only, or a combination of both, per architecture type. Based on INLG, and NLG tracks in ACL papers published between 2016–2019.

Method	% Aut.	% Intr. hum.	% Extr. hum.	Aut. only (%)	Hum. only (%)	Hum. + aut. (%)
Neural	96	47	3	52	3	44
Statistical	76	45	5	45	16	31
Rule-based	41	50	5	30	41	11

reporting the average runtime of an approach is essential to gauge the efficiency of the system (Dodge et al., 2019). However, most of the widely-used automatic metrics (BLEU, ROUGE, BertScore, etc.) aim to measure humanlikeness. When the goal is to assess real-world NLG applications (which is most commonly the case, and the focus in the current paper; see Section 1.1), evaluations should ultimately address the "usefulness" of the system. This is something that humanlikeness metrics are typically incapable of, and for which human evaluation remains the gold standard.

Furthermore, automatic metrics are non-viable for NLG systems developed in domains for which large-scale data sources consisting of input-output pairs are unavailable and are prohibitively expensive or impractical to construct. To take but one example, this has historically been true of a number of data-to-text systems in healthcare (Harris, 2008; Portet et al., 2009), where data with accompanying text can be scarce, or is difficult to obtain for ethical reasons. As a result, such systems are often built with scarce resources, and typically engineered on the basis of careful analysis of limited data, together with consultation with domain experts and testing on a trial-and-error basis. (Not surprisingly, this was the approach recommended in the reference text on NLG by Reiter and Dale (2000), which was written prior to the development of current data-driven approaches.)

3. Overview of current work

This section provides an overview of current human evaluation practices, based on the papers published at INLG 2018 (N=51) and 2019 (N=50), and ACL 2018 (N=68) and 2019 (N=135). Databases and scripts can be found at https://osf.io/rt3by/. We did not observe noticeable differences in evaluation practices between INLG and ACL, which is why they are merged for the discussion of the bibliometric study.[7]

3.1. Intrinsic and extrinsic evaluation

Human evaluation of NLP systems can be done using intrinsic or extrinsic methods (Belz and Reiter, 2006; Resnik and Lin, 2010; Sparck Jones and Galliers, 1996). Intrinsic approaches aim to evaluate properties of the system's output, for instance, by using a questionnaire to elicit judgments of fluency from participants. Extrinsic, or task-based, approaches aim to evaluate the impact of the system, by investigating to what degree the system achieves the overarching task for which it was developed. A further distinction can be made between intrinsic and extrinsic evaluations performed in an uncontrolled setting (using the NLG system in the intended real-world situation) vs. a controlled setting (using the system in an artificial laboratory context) (Reiter, 2011). The former case requires knowledge of and access to the intended use context and the user base the system is intended to support (cf. Sparck Jones and Galliers, 1996), which may not always be feasible[8] While extrinsic evaluation has been argued to be more useful (Reiter and Belz, 2009) than intrinsic, and a real-world setting more useful than a controlled one, they are rarely evidenced in the literature. Only eight papers (3%) in the sample of INLG and ACL papers presented an extrinsic evaluation, and all of these did so in a controlled setting. The same goes for the papers reporting on an intrinsic study, with only one exception.

Of course, extrinsic evaluation is already time- and cost-intensive (Gatt and Krahmer, 2018), and developing a system up to a point where it can be tested in a real-world scenario costs even more in terms of time and resources. This might be one reason for the rarity of such studies. However, another reason is likely to be the type of NLG tasks emphasised in recent NLG research.

There are comparatively few systems developed for use in a specific, or at least, an identifiable context. Rather, a lot of research in NLG centres around tasks in which the quality of the output text is itself the research goal. By the same token, neural generation models tend to optimise for output text quality, although this raises notoriously hard problems, such as the fidelity of output to the input data, and the tendency of some NLG models to "hallucinate", that is, to generate text which is ungrounded in the input (e.g., MArchegiani and Perez-Beltrachini, 2018; Rohrbach et al., 2018). In any case, if output quality is itself the goal, then it is arguably reasonable to emphasise intrinsic evaluation with a focus on properties of text. Thus, while intrinsic evaluations are arguably artificial, insofar as they are divorced from an actual setting and involve metalinguistic judgments (Reiter, 2017), they remain the most appropriate for the NLG models currently being developed in research settings and are also

[7] For the ACL papers, we focused on the following tracks: Summarization, Question Answering, Dialog Systems, and Generation. See the Supplementary Materials at https://osf.io/rt3by/ for a detailed overview of the papers included and their evaluation characteristics.

[8] For example, because direct testing in the target context is ethically problematic, as in a medical context where output may impact decision-making for vulnerable patients; see Portet et al., 2009 as an example.

Table 3
Criteria used for human evaluation from all papers, and percentage of papers with a human intrinsic study using the criteria.

Criterion	Total	Criterion	Total
Fluency	40 (27%)	Readability	9 (6%)
Overall quality	29 (20%)	Appropriateness	7 (5%)
Informativeness	15 (10%)	Meaning preservation	6 (4%)
Relevance	15 (10%)	Clarity	5 (3%)
Grammaticality	14 (10%)	Non-reduncancy	4 (3%)
Naturalness	12 (8%)	Sentiment	4 (3%)
Coherence	10 (7%)	Consistency	4 (3%)
Accuracy	10 (7%)	Answerability	4 (3%)
Correctness	9 (6%)	Other criteria	124 (48%)*

* 71 out of 147 papers with an intrinsic evaluation used one or more other criteria.

by far the most popular methods. Furthermore, the metalinguistic judgments that are the result of an intrinsic evaluation can be useful when the goal of an NLG system is to test linguistic theories and its effects, as is the case for "theoretical NLG" (see, for instance Van Deemter, 2016; Van Deemter et al., 2017).

Therefore, the best-practices described in Section 4 will mostly be focused on this type of evaluation, although possibilities for extrinsic evaluation and evaluation in real-world context will be discussed as well. An additional reason for the focus on intrinsic evaluation is the fact that such an evaluation is more feasible for standardization. Extrinsic evaluations are oftentimes too domain-dependent and thus harder to be abstracted into a more general practice guideline.

3.2. Sample size and demographics

Sample size is an important issue in empirical studies, as it impacts statistical power (that is, the ability to find statistically reliable trends where these are present). When looking at sample size, it is possible to distinguish between small-sample (oftentimes expert-focused) evaluation on the one hand, and evaluation with larger samples (commonly representing a general audience) on the other. This division is also related to the aforementioned uncontrolled versus controlled context (Reiter, 2011), although the two are not necessarily the same: uncontrolled, real-world experiments always require a group of representative users, but such users can also be deployed in a laboratory setting, by far the most frequent scenario in the expert-focused papers included in our sample.

Of all papers in the investigated sample, 39% used a small-scale approach, meaning that between 1 and 5 annotators evaluated system output. Furthermore, 15% employed a larger-scale sample in which 10 to 670 readers judged the generated output, and 3% had a sample between 5 and 10 readers. Thus, only 57% of papers specified the number of participants. We found a median of 3 annotators.

Only 5 papers (3%) reported demographic information about their sample. 19% of the papers with a human evaluation reported inter-annotator correlation or agreement, using Pearson's correlation coefficient, Krippendorff's α, Fleiss' κ, Weighted κ or Cohen's κ. Agreement in most cases ranged from 0.3 to 0.5, but given the variety of metrics and the thresholds used to determine acceptable agreement, this range should be treated with caution.

3.3. Design

Apart from participant sample size, another important issue which impacts statistical power is the number of items (e.g. generated sentences) used in an evaluation. Among INLG and ACL papers that reported these numbers, we observed a median of 100 items. The number of items however ranged between 2 and 5,400, illustrating a sizable discrepancy. However, it should be noted that it is difficult to compare item amounts between papers, as the items can differ greatly in length (from a single phrase/sentence to a full text). In 72% of papers that reported these figures, all annotators saw all examples, although only 27% of papers report these figures. Therefore, it is hard to gauge how much time participants spent on the evaluation. Only 10 (6%) of the papers doing a human evaluation reported other aspects of evaluation study design, such as the order in which items were presented, randomisation and counterbalancing methods used (e.g. a latin square design), or whether criteria were measured at the same time or separately.

3.4. Statistics and data analysis

A minority (23%) of papers report one or more statistical analyses for their human evaluation, to investigate if findings are statistically significant. The types of statistical analyses vary greatly: there is not one single test that is the most common. Examples of tests found are Student's *t*-test, Mann-Whitney U test, and McNemar's test. Theoretically, such statistical tests should be performed to falsify a specific null hypothesis (Navarro, 2019). However, not all papers using a statistical test report their hypotheses

Table 4
Types of scales used for human evaluation, and percentage of papers with a human intrinsic study using the scales.

Scale	Total
Categorical choice	51 (35%)
Likert (5-point)	47 (32%)
Likert (3-point)	21 (14%)
Other Likert (4, 6, 7, 10, 11-point)	14 (10%)
Ranking	11 (7%)
Rank-based Magnitude Estimation	5 (3%)
Best-Worst scaling	2 (1%)
Free text comments	1 (1%)
Unknown	13 (9%)

explicitly. Conversely, some papers which state hypotheses do not perform a statistical test. Only 13% of all papers explicitly state their hypotheses or research questions, and no papers mention preregistering their hypotheses or research questions.

3.5. Properties of text quality

Many studies take some notion of 'text quality' as their primary evaluation measure, but this goal is not easy to assess, since text quality criteria differ across tasks (see Section 4.3 for further discussion). This variety, suggesting a lack of agreement, is clear from Table 3, which summarises the various quality dimensions evaluated in the papers in our sample. While there are some common criteria across all NLG tasks, most are used in only a few publications; in Table 3, 'other criteria', which combines those that appear only three times or less in the sample, is the largest grouping. At the same time, there is probably significant overlap between criteria. For instance, naturalness is sometimes linked to fluency, and informativeness to adequacy (Novikova et al., 2018). There are also papers that define fluency in terms of grammaticality and vice versa, illustrating the vagueness in terminology. Moreover, 50% of the papers did not report definitions or questions for the criteria that they investigated. Because of this vagueness and frequent lack of definitions, we decided not to merge criteria that appeared similar in the bibliometric study.

In short, there is no standard evaluation model for NLG, nor agreement in terminology, and explanatory details for the criteria are often lacking.

3.6. Number of questions and types of scales

In addition to the diversity in criteria used to measure text quality (see Section 3.5), a wide range of rating methods is used to measure them. Categorical choice ratings (e.g. preference ratings, yes/no questions, variations on the Turing Test) are the most popular rating method, but 5-point Likert scales are a close second (see Table 4), and Likert scales overall are by far the most popular method.[9] Other types of rating methods are much less common. Ranking-based methods (i.e. asking participants to rank items, Rank-based Magnitude Estimation, and Best-Worst scaling) appear much less frequently, and only one paper reported using free-text comments.

We also investigated the number of different ratings used to measure a single criterion. For instance, a paper may measure a sentence's fluency by asking participants to answer two statements: (i) "This sentence is easily readable"; (ii) "While reading this sentence, I immediately understood what it said". Only 32% of papers with a human evaluation reported the number of different ratings used to measure any single criterion. These numbers ranged from 1 to 6 ratings for a criterion, with 1 rating to measure a criterion being the most common (92% of cases). Based on the 1 rating being so common, it seems likely that the papers that omitted the exact number of ratings also used 1 rating per criterion.

4. Best practices

This section discusses what we view as best practices for carrying out and reporting human evaluation in NLG. We will separately discuss important aspects involved in setting up an evaluation study first, and draw attention to how common difficulties can be handled for each aspect. It should be noted that these steps are not intended as one-size-fits-all solutions: some steps can be skipped depending on the situation. Finally, the steps are summarized into separate stages with corresponding best practice recommendations. With that in mind, this framework should be helpful to all researchers that consider doing a human evaluation.

[9] Note that in this study, we do not distinguish between Likert and rating scales (for a distinction, see Amidei et al., 2019b).

Table 5
List of stages and steps for conducting a human evaluation research of automatically generated text, accompanied by best practices per step.

Planning stage	Recommendations
1. Determine the **goal** of the evaluation (Section 4.1)	• The goal should be summarized in an explicit research question. • Determine if you are testing hypotheses. Explicitly state these. • Choose strong, representative baselines, and baselines that contextualize the performance of the system.
2. Determine the **type** of **evaluation** (Section 4.1) (intrinsic/extrinsic; real-world/lab setting)	• The type of evaluation should depend on the goal and the constraints within which the evaluation takes place.
3. Determine the **type** of **research** (Section 4.2) (qualitative/quantitative)	• Qualitative is preferred if the goal is to improve the system. • Quantitative is preferred if the goal is to judge the merit of the system.
4. Define the **constructs** of **interest** (Section 4.3)	• Determine whether to ask implementation questions or impact questions. • Use separate criteria rather than an overall text quality construct. • Choose the criteria depending on the task and the goal of the research. • Give formal definitions of the criteria and concrete examples in the instructions before the questionnaire.
5. Determine the appropriate **scales** and **scale size** (only for quantitave research) (Section 4.4)	• Use either multiple-item 7-point Likert scales, or a (continuous) ranking task.
6. Determine the **sample** (Section 4.5)	
a. **Kind** of participants (experts/laypeople)	• Recruit a sample that reflects the target audience. • Provide a detailed description of the sample's demographics.
b. **Number** of participants	• Use large-scale rather than small-scale samples (for quantitative research). • For most designs and analyses, 100 or more participants are needed. Calculate the minimum sample size required with a tool such as G*Power (for quantitative research). • Difficult coding tasks (which most NLG evaluations are) require three or more annotators, more straight forward tasks can do with two to three.
c. **Output** sample	• Select outputs for low, medium, and high-frequent inputs.
7. Further specify the study's **design** (Section 4.6)	• If feasible, choose a within-subjects design over a between-subjects design. • Try to keep the evaluation task simple and motivating for participants • Reduce practice effects with a practice trial. • Reduce carryover effects by increasing the amount of time between presenting different conditions. • Reduce fatigue effects by shortening the task. • Reduce order effects by showing conditions in a systematically varied order. • Reduce nonresponse bias by carefully considering the survey's appearance.
8. Select a **statistical** approach (only for quantitative research) (Section 4.7)	• If exploratory research is conducted, use exploratory data analysis techniques. • When there are clear hypotheses, use statistical significance testing and report effect sizes.
9. (Optional) **Preregister** the task (Section 4.7)	• If the evaluation is confirmatory, consider preregistration.
Execution and release stage	**Recommendations**
1. Select an **evaluation platform** (Section 4.8)	• Choose an appropriate platform based on time and cost constraints.
2. Develop the **consent form** and **debriefing statement (Section 4.8)**	• Keep both consent form and debriefing statement as simple, short, and clear as possible.
3. Apply for **ethical clearance** (Section 4.8)	• Explicitly state that ethical clearance was obtained (plus Ethical Approval Code), and that the research is in compliance with the relevant data protection legislation.
4. (Optional) Conduct a **pretest** (Section 4.8)	• Consider pretesting when describing new instruments, or for manipulation checks.
5. Conduct the **evaluation study**	• Conduct the study based on the aforementioned considerations.
6. **Publish** the raw data and materials (Section 4.9)	• Try to make the raw data and all materials publicly available.

4.1. Goal definition

Research question Every evaluation should begin with a goal definition phase, during which is determined "what is being evaluated, the purpose of the evaluation, the stakeholders in the evaluation, and the constraints within which the evaluation will take place" (Mertens, 2010, p. 69). At first glance, it seems straightforward to identify what is being evaluated; in most cases, it is the developed NLG system. However, the picture can be considerably more complicated. Is the intention to test the merits of the developed system on its own, to compare the effect of various sub-components on the textual output, or compare the developed system to other systems? At this stage it is imperative to think about whether to test the system in a task-based (i.e. extrinsic) fashion, or by collecting (intrinsic) human ratings. Similarly, it is important to determine if the aim is to test the system in a real-world setting or in a controlled, artificial environment. The right research question(s) will guide these decisions regarding the research design and methods (Blaikie, 2000). The problem should be stated in a research question with enough details so that the evaluation research provides an answer to the question as unambiguously as possible (De Vaus, 2001). Therefore, it is advisable in NLG research to not only start the evaluation by asking a research question, but also to *explicitly state this research question in the paper so that it is clear what is being investigated, and why this is being investigated* (which is uncommon in NLG, see Section 3.4). Do note that research questions differ from hypotheses. Research questions serve to narrow down a topic, and are

applicable to both exploratory and confirmatory research, while hypotheses are educated guesses based on previous literature and are only suitable for confirmatory research (see Section 4.7).

Hypotheses and baselines In the behavioral sciences, it is standard to refine research questions into testable hypotheses. However, one may wonder whether standard null-hypothesis significance testing (NHST) is applicable or helpful in human NLG evaluation. Of course, researchers generally wish to test whether their system is evaluated more positively than comparison systems, but the exploratory nature of the field often means that there is no explicit empirical support to base these assumptions on. Researchers also need not necessarily assume that their system will perform better on all dependent variables. Moreover, they may have no specific hypotheses about which variant of their own system will perform best. **Still, if the research lends itself to positing explicit hypotheses, we would advise to do so (and explicitly state these)**, as they inform the design choices to make and the statistical tests to perform.

While it may be difficult to predict the performance of a developed system against baselines in the form of hypotheses, it is important to choose these baselines scrupulously. Baselines generally serve as sanity checks that the output of the developed system is of meaningful quality (Søgaard, 2017). But choosing which type of baselines to use can significantly affect the conclusions regarding this quality. An experimental approach may seem overall much better when compared against a simple, vanilla baseline model, but a comparison against baselines that are similar to the introduced system may reveal the subtle interactions, the strengths, and the weaknesses of various techniques (Denkowski and Neubig, 2017). Similarly, a neural system may seem competitive against other neural systems, but perhaps the task is still so challenging that competitive results can be obtained with straightforward fill-in-the-blank template, random, or majority vote baselines (neural data-to-text generation is one such example, see Castro Ferreira et al. (2019)). **Therefore, we advise for most types of system-focused NLG research to choose baselines that are strong, representative of those deployed in real-world cases, and are as similar as possible to the system being evaluated, so as to enable a form of ablation testing** (Denkowski and Neubig, 2017; Søgaard, 2017). **And additionally it is good to have simple (e.g. fill-in-the-blank template, random, majority) baselines to enable contextualization of the introduced system.** In any case, it is good practice to think carefully about which baselines to choose and provide an explicit argument for the baseline choice(s) in the paper.

4.2. Research type

If the purpose of the evaluation is clear, a sensible next step would be to outline the questions that are needed to find an answer to the research question. In other words: which type of measure should be considered for the evaluation study. Important here is whether the purpose of the evaluation study is *formative* or *summative* (Scriven, 1991). A formative evaluation is conducted when the system's development cycle is not yet complete and serves to improve the system. A summative evaluation is conducted when the system's development is complete and serves to provide judgment about the merit of the system.

Qualitative research **If the goal of the evaluation is to improve the system, it would be most beneficial to focus on qualitative research** (i.e. empirical research where the gathered data is not numerical (Punch, 1998, p.4). This is because such data can give explicit indications of which linguistic choices made by a system are negatively evaluated by users. To the extent that such choices can be traced back to the point where they are made, which is of course heavily dependent on the architecture of the system in question, they can be rectified. Alternatively, they can motivate a consideration of the input data or training regime used to train the models.

Examples of qualitative research that could be used in the context of NLG are open-ended questionnaires (or: free-text comments), semi-structured or unstructured interviews, participant observation, and qualitative text analysis methods (see Lacity and Janson, 1994). Of these, perhaps the most common in the NLG literature are questionnaires and free-text comments, and qualitative analysis. Both methods have been defended as potentially useful diagnostic evaluation tools (Reiter and Belz, 2009; Sambaraju et al., 2011). Sambaraju et al. (2011) report on a content and discourse analysis exercise carried out on texts generated in the medical domain, arguing that they shed light on limitations in generated text, including mismatches between the intended meaning of words or phrases and the way humans actually interpret them. In another, large-scale evaluation of a medical NLG system carried out on-ward (Hunter et al., 2012), a questionnaire was used to collect users' impressions of the understandability, accuracy and helpfulness (for decision support) of the system's outputs. The questionnaire also included a free-text component on which the authors performed a limited form of content analysis which suggested, among other things, that a number of problems were localised in the system's analysis and interpretation of the input data, rather than, say, the microplanning or realisation modules of the system. More recently, Hommes et al. (2019) used semi-structured interviews with experts to gather information about content that is currently missing from their system.

As these examples show, qualitative analyses can help in a diagnostic evaluation and overcome potential blind spots of purely quantitative analyses. At the same time, making sense of free-text comments or other unstructured data involves substantial analytical effort and introduces issues of reliability. While these can often be addressed using familiar measures for inter-annotator agreement, for methods such as discourse analysis, subjectivity seems inherent to the method itself. In such cases, such as the study of Sambaraju et al. (2011), any conclusions would need to be complemented by data from larger-scale, quantitative studies.

Regardless of their value, qualitative evaluation methods have become rare in NLG (and NLP more generally), though it has become more common to find 'error analysis' sections in papers reporting on neural models. These are often invaluable in shedding light on the types of cases on which models perform well or not, and can give rise to interesting insights as to the possible reasons for such observations. However, such analyses are usually conducted on a small scale, and typically by the authors

themselves, rather than garnering insights from independent participants. It could benefit many studies to elicit and report free-text comments or conduct error analysis systematically, with independent participants. These could also serve to channel future research efforts to address commonly observed shortcomings. As noted above, one of the benefits of qualitative data is that, rather than comparing models or giving an 'aggregate' view of a model's performance (as many metrics do), it can yield specific insights with a diagnostic value.

Quantitative research **In most cases for NLG, the study has a summative purpose, for which quantitative (human) research is appropriate** (i.e. empirical research which gathers numerical data; this data can be aggregated into categories, put in rank order, or measured in appropriate units (McLeod, 2019). The most common quantitative method in NLG evaluation is the questionnaire (with closed questions), although experiments (e.g., Di Eugenio et al., 2002) are possible as well. Quantitative research is oftentimes viewed as scientifically objective and rational (Carr, 1994; Denscombe, 2010),[10] and is the most well-suited type of research to test hypotheses. Furthermore, analysis of quantitative research is usually quick and relatively effortless compared to an analysis of qualitative research. This is especially the case when large-scale studies are involved.

However, large sample sizes are needed for an accurate analysis of quantitative research, while qualitative research can make do with smaller samples (for further discussion, see Section 4.5). Constructing a valid quantitative research design can also take as much work as doing a qualitative analysis. On the other hand, repeating a previously constructed design for a quantitative study is generally much faster than doing a qualitative analysis, provided that the assumptions of the previous design continue to hold.

Besides time considerations, it is most important to choose a type of analysis that suits the research goal. Purely quantitative experiments do not provide an opportunity for participants to explain their answers, or explain how they interpreted the questions (Carr, 1994). If this is considered desirable, then qualitative questions are needed.

4.3. Question selection

Question types After deciding on the main features of the research, an important next step is to think about which questions to ask. A first distinction that can be made, based on Mertens (2010) is whether to ask *implementation questions* or *impact questions*. Implementation questions are those that help to verify whether system components have been implemented as intended, and are therefore mostly relevant when the goal is to evaluate components of a system. van der Lee et al. (2018), for instance, asked participants if texts were tailored towards specific groups, which assessed whether the tailoring component of their system was successfully implemented. Similarly, for style transfer, manipulation checks (e.g. "what is the sentiment of the text") are important to determine whether the style has been transferred correctly, while also ensuring meaning preservation (but see Section 4.8 for discussion on manipulation checks). However, for many other NLG tasks, this type of question is uncommon. Most research involves impact questions, which sheds light on whether the system achieves its intended purposes, or at least produces output which readers evaluate positively. For most NLG systems which produce textual output, this will mean that the questions should indicate if the system manages to output texts of high quality.

Text quality Renkema (2012, p. 37) defines text quality in terms of whether the writer (or: NLG system) succeeds in conveying their intentions to the reader. He outlines three requirements for this to be achieved: (i) the writer needs to achieve their goal while meeting the reader's expectations; (ii) linguistic choices need to match the goal; and (iii) the text needs to be free of errors (broadly conceived to encompass all types of errors).[11]

If successfully conveying communicative intention is taken to be the overarching criterion for quality, then two possibilities arise. One could treat quality as a primitive, as it were, evaluating it directly with users. Alternatively—and more in line with current NLG evaluation practices—one could take text quality to be contingent on individual dimensions or criteria (for various studies of such criteria, see Dell'Orletta et al. (2011); Falkenjack et al., 2013; Nenkova et al., 2010; Pitler and Nenkova, 2008, inter alia). The choice between these two options turns out to be a point of contention. Highly correlated scores on different quality criteria suggest that human annotators find them hard to distinguish (Novikova et al., 2017) (although careful design could potentially address this ambiguity; see Section 4.6). For this reason, some researchers directly measure the overall quality of a text. However, Hastie and Belz (2014) note that an overall communicative goal is often too abstract a construct to measure directly. They argue against this practice and in favour of identifying separate criteria, weighted according to their importance in contributing to the overall goal. In line with their suggestion, **we recommend the use of separate criteria.**

The position taken by Hastie and Belz (2014) implies that, to the extent that valid and agreed-upon definitions exist for specific quality criteria, these should be systematically related to overall communicative success. Yet, this relationship need not be monotonic or linear. For example, two texts might convey the underlying intention (including the intention to inform) equally successfully, while varying in fluency, perhaps as long as some minimal level of fluency is satisfied by both. In that case, the relationship would not be monotonic (higher fluency may not guarantee higher evaluation of communicative success beyond a certain point). A further question is how the various criteria interact. For instance, it is conceivable that under certain conditions (e.g. summarising high-volume, heterogeneous data in a short span of text), readability and adequacy are mutually conflicting goals beyond a certain point (e.g. because adequately conveying all information will inevitably result in more convoluted text which is harder to understand).

[10] Although "objective" numbers can still be misleading in various ways (for a discussion, see MacDonald, 2018).
[11] See also Gabriel (1988); Kukich (1988) for text quality definitions.

Criteria **While certain quality criteria may be relevant for most, if not all, NLG evaluations, the type of text being generated, or the goal of such text, should also be taken into account.** For instance, for a system that generates weather reports from numerical data (e.g., Goldberg et al., 1994; Reiter et al., 2005; Ramos-Soto et al., 2015), accuracy, fluency, coherence, and genre compatibility will be of foremost importance. By contrast, accuracy is of less concern for generation tasks such as poetry generation and story ending generation, as their output cannot usually be judged by fidelity to an identifiable, external input. Similarly, coherence and fluency would not be important criteria for the PARRY chatbot (Colby et al., 1971) which attempts to simulate the speech of a person with paranoid schizophrenia.

As we have shown in Section 3.5, the criteria used for NLG evaluation are usually treated as subjective (as in the case of judgments of fluency, adequacy and the like). It is also conceivable that these criteria can be assessed using other, possibly more objective measures. Examples include using reading time to measure fluency; or carrying out a structured and detailed error analysis where (amongst others) the number of lemmatization errors, lexicalization errors, and repetitions are counted and reported (see, for instance Popel and Žabokrtsk₹2009; Dušek and Jurčíček, 2019), which could provide a different idea regarding a system's fluency, correctness, accuracy, etc.

One obstacle to addressing the difficulties identified in this section is the lack of standardized questionnaires, and a standardised nomenclature for different text quality criteria. This presents a practical problem, in that it is hard to compare evaluation results to previously reported work; but it also presents a theoretical problem, in that the employed questionnaires might not be valid, and different criteria may overlap or be inter-definable. As Hastie and Belz (2014) suggest, common and shared evaluation guidelines should be developed for each task, similar to what Williams and Scheutz (2017) have done for Referring Expression Generation. Furthermore, efforts should be made to standardise criteria and naming conventions. Some work on this has been done in the domain of dialogue systems (Morrissey and Kirakowski, 2013; Radziwill and Benton, 2017), and we greatly encourage more research on this topic. **In the absence of such guidelines, care should be taken to report the questions asked, and also to explicitly define the criteria measured and highlight possible overlaps between them.** Similarly, many text quality criteria can be confusing to participants, thereby harming criterion validity. **For participants, formal definitions and concrete examples in the instructions before the questionnaire help to understand how the researchers have operationalized the relevant criteria** (Geiger et al., 2020).

4.4. Scale selection

Likert scales As shown in Section 3.6, Likert scales are the prevalent rating method for NLG evaluation, 5-point scales being the most popular, followed by 3-point, and 4-point scales. While the most appropriate number of response points may depend on the task itself, 7-point scales (with clear verbal anchoring) seem best for most tasks. In the experimental literature, there is some consensus that 7-point scales maximise reliability, validity and discriminative power (e.g., Miller, 1956; Green and Rao, 1970; Jones, 1968; Cicchetti et al., 1985; Lissitz and Green, 1975; Preston and Colman, 2000). These studies discourage smaller scales, and adding more response points than 7 also does not increase reliability according to these studies.

It should be noted that we use Likert scales and rating scales synonymously in this paper, as they are often used interchangeably in literature, although differences between the two exist. Likert scales usually express a matter of agreement/disagreement with a statement, while rating scales measure the attitudes, feelings, or opinions of participants. Furthermore, it is important to note that Likert scales are considered ordinal scales by some and interval scales by others. This is an important distinction to make, as it affects the statistical tests that should be applied to the collected data. Therefore, it is good practice to justify the scale interpretation when reporting statistical testing on Likert scales (see Amidei et al., 2019b for a further discussion on this topic).

Ranking While Likert scales are the most popular scale in NLG (and probably in many other domains), the use of this scale has been receiving more and more criticism. Recent studies have found that participant ratings are more reliable, consistent, and are less prone to order effects when they involve ranking rather than Likert scales (Martinez et al., 2014; Yannakakis and Martínez, 2015; Yannakakis and Hallam, 2011). Empirical studies have also found that best-worst scaling (i.e., a method where participants only indicate which item they found the best and which one the worst) was more reliable than Likert scales (Kiritchenko and Mohammad, 2017), and that (rank-based) magnitude estimation (i.e. a method where participants indicate on a scale how sentences compare to a reference sentence) was more reliable and consistent (Novikova et al., 2018). In contrast, an extensive study by Langsford et al. (2018)—comparing categorical choice, Likert, and Magnitude Estimation measures on a sentence grammaticality rating task—found Likert scales to be reliable, stable, and the overall preferred option.

While some of these studies suggest a preference for ranking-based methods, there are two critical remarks to be made in this regard. Firstly, a drawback of ranking-based methods is that the number of judgments increases substantially as more systems are compared. To mitigate this, Novikova et al. (2018) use the TrueSkill™ algorithm (Herbrich et al., 2007). This algorithm uses binary comparisons to reliably rank systems, which greatly reduces the amount of data needed for multiple-system comparisons as systems do not have to be compared in different pairs.

Multiple items Another point of criticism is that studies comparing Likert scales to other research instruments mostly look at single-rating constructs, that is, experiments where a single judgment is elicited on a given criterion. While constructs measured with one rating are also the most common in NLG research, this practice has been criticized. It is unlikely that a complex concept (e.g. fluency or adequacy) can be captured in a single rating (McIver and Carmines, 1981). Furthermore, a single Likert scale often does not provide enough points of discrimination: a single 7-point Likert question has only 7 points to discriminate on, while five 7-point Likert questions have 5 * 7 = 35 points of discrimination. A practical objection against single-item scales is that no reliability measure for internal consistency (i.e., how different items correlate on the same test, commonly measured using

Cronbach's alpha) can be calculated for a single item. At least two items or more are necessary for this. In light of these concerns, Diamantopoulos et al. (2012) advocate great caution in the use of single-item scales, unless the construct in question is very simple, clear and one-dimensional. Under most conditions, multi-item scales have much higher predictive validity. Using multiple items may well place the reliability of Likert scales on a par with that of ranking tasks; this, however, has not been empirically tested. Furthermore, it is important to not ask redundant questions when using multiple items. Grice's Maxims Grice (1975) for conversational norms should also apply to survey questions (Krosnick, 2018b). Also, note that the use of multiple-item scales versus single-item scales affects the type of statistical testing needed (for an overview and explanation, see Amidei et al., 2019b).

In sum, we recommend **either multiple-item 7-point Likert scales, or a (continuous) ranking task when doing a quantitative study.** This ranking task can be done in combination with TrueSkill™ or Best-Worst scaling when multiple systems are compared, so as to avoid having to obtain too many ratings. Using multiple-item scales rather than single-item scales to measure a criterion will of course increase the time it takes for participants to judge an example, thereby making it difficult to have participants judge the same number of examples as they would with single-item scales. However, it is better to have a smaller number of accurately rated examples rather than a larger number of inaccurately rated examples.

Implicit measures It should be noted that the quantitative measures discussed above are all self-report measures. However, self-report measures have received criticism as well. Participants could see self-reporting as a form of self-presentation (Bainbridge, 1990), meaning that participants may want to make a good impression, which could distort results. For instance, participants could give a high rating on the understandability of a text, even though they do not fully understand it, in order to appear clever. Or they might give the answers that they think the researcher desires (Lentz and De Jong, 1997). A questionnaire also triggers participants to actively engage in introspection, whereas regular perusal of the output of an NLG system would mostly be guided by automatic and unconscious processes (Frith and Frith, 2012). Self-report metrics might not catch such implicitly-formed impressions.[12] Employing a task-based (i.e., extrinsic) evaluation study would avoid these self-report issues, but alternative measures for an intrinsic study exist as well.

As an example, Zarrieß et al. (2015) used a mouse contingent reading paradigm in an evaluation study of generated text, finding that features recorded using this paradigm (e.g. reading time) provided valuable information to gauge text quality levels. Heyselaar and Bosse (2019) used a referential communication game to assess whether participants perceived a dialogue agent to have their own mental state, which could be a proxy for a dialogue system's output quality. It should also be noted that most metrics used in NLG are reader-focused. However, there are many scenarios where NLG is used in tandem with human authorship,[13] especially in 'creative writing' NLG applications (e.g. story generation, song lyric generation, poetry generation), where humans act as the editor of the system's output (e.g. Maher, 2012; Manjavacas et al., 2017), or where active co-creation is embedded in the design of the system (e.g. Kreminski et al., 2020; Perrone and Edwards, 2019). In such collaborative settings, it makes sense to also investigate writer-focused methods. For example, having participants edit generated texts, then processing these edits using post-editing distance measures like Translation Edit Rate (Snover et al., 2006), might be a viable method to investigate the time and cost associated with using a system. While more commonly seen in Machine Translation, different authors have explored the use of such methods in NLG (Bernhard et al., 2012; Han et al., 2017; Sripada et al., 2005).

Implementing implicit measures may incur higher costs in terms of time and effort. **However, implicit measures are an effective way to assess the—usually hidden—implicit opinions about a system's output.** An additional reason to consider using implicit measures is the recent discussion on the importance of "triangulation" in science: addressing one research question using various approaches (Munafò and Smith, 2018). When an evaluation is carried out using implicit measures as well as self-report measures, and agreement is found between the two measures, this would mean that the findings are less likely to be artifacts.

4.5. Sample selection

Sampling, the method used to select a number of people from a population (Mertens, 2010) is an important decision in the evaluation design process, as the method chosen not only affects the quality of the data, but also the inferences that can reliably be made from this data.

Small-scale (expert) vs. large-scale (non-expert) Section 3.2 made a distinction between small-scale evaluation (with 1–5 participants), and large(r)-scale evaluation, where 10 or more participants are recruited. Generally, the small-scale evaluation studies employ expert annotators to judge aspects of the NLG system, while the larger-scale design entails a typically larger sample of non-expert participants. Lentz and De Jong (1997) found that these two methods can be complementary: expert problem detection may highlight textual problems that are missed by general readers. However, this strength is mostly applicable when an error analysis or a qualitative analysis is used for the small-scale study, whereas most small-scale evaluations in our sample of papers used quantitative analyses.

Evidence suggests that expert readers approach evaluation differently from general readers, injecting their own opinions and biases (Amidei et al., 2018b). This might be troublesome if a system is meant for the general population, as expert opinions and biases might not be representative for those of the target (non-expert) population. This is corroborated by Lentz and

[12] These issues with self-report measures do not mean that we suggest abandoning their use. Overall, a meta-analysis across various domains showed a strong agreement between self-reports and objective records of the same phenomena (Krosnick, 2018a). However, it is good to be aware that self-report measures might not reflect the whole truth about a system's output.

[13] See Edwards et al. (2020) for an overview of levels of collaboration.

De Jong (1997), who found that expert judgments only predict the outcomes of non-expert evaluation to a limited extent. Having only a few participants also risks giving rise to considerable variance, so that automatic metrics are sometimes more reliable (Belz and Reiter, 2006). Similarly, external validity becomes a concern: it is difficult to argue that the results from 1 to 5 participants can be extrapolated to the whole population (Norman, 2010). Thus, the conclusion of Belz and Reiter (2006) in favour of **large-scale samples, rather than small-scale ones**—at least for quantitative studies—seems well-taken.

It is important to note that this recommendation is based on the fact that most papers in our bibliometric study did not outline the intended audience for the system. Therefore, we assume that most systems have been developed for a general audience. However, the most important consideration when deciding to recruit an expert or non-expert sample is to **choose a sample that reflects the audience for which the system was developed.** For instance, systems trained on the SumTime corpus (Reiter et al., 2005), which contains weather forecasts for offshore oil rigs, are best evaluated by participants who are experienced with reading these texts (Reiter and Belz, 2009). Similarly, if the evaluation of a poetry generator aims to identify the extent to which the output contains the features expected of the genre and their aesthetic qualities (rather than, say, the impact poems have on general readers), then poetry experts are probably the ideal population from which to sample evaluators. In many cases, it is important to verify that participants have the requisite skills, knowledge or expertise to perform the evaluation task. This can be checked, for instance, by screening potential participants using a representative task and rejecting participants that fail this task. This is especially important in crowdsourcing contexts where it is difficult to vet people beforehand.

Sampling bias An additional factor to consider is the types of 'general' or 'expert' populations that are accessible to NLG researchers. It is not untypical for evaluations to be carried out with students, or fellow researchers (recruited, for instance, via SIGGEN or other mailing lists). This may introduce sampling biases of the kind that have been critiqued in psychology in recent years, where experimental results based on samples of WEIRD (Western, Educated, Industrialised, Rich and Developed) populations may well have given rise to biased models (see, for example (Henrich et al., 2010). The rise of crowdsourcing in recent years makes it relatively easy to obtain large-scale samples from all over the world (though see van Miltenburg et al., 2017 for a counterexample). However, crowdsourced samples face bias issues as well (see, for instance Fulgoni, 2014).

It is generally difficult, if not impossible, to recruit a participant sample that is free of any type of bias, which would mean that the results of the study cannot necessarily be generalized to the target population (in most cases: the general population). What researchers should always do is **provide a detailed description of the sample**, meaning participant numbers with relevant demographic data (i.e., gender, nationality, age, fluency in the target language, academic background, relevant expertise, etc), which is common practice in the psychological and medical fields. This lets readers know the composition of the sample, and the setting in which the sample was collected, which in turn empowers the reader to judge the generalizability of the sample (Stake, 2000; Lincoln and Guba, 2000), and enhances replicability of the evaluation study.

Evaluator agreement Besides being non-representative due to sampling bias, results can also be difficult to generalize because of excessive variability between raters, reflected by a low Inter-Annotator Agreement (IAA). Low IAA is common in NLG studies (see also Amidei et al., 2019a), although thresholds to determine what counts as high or low tend to be open to interpretation (Artstein and Poesio, 2008). Amidei et al. (2018b) argue that, given the variable nature of natural language, it is undesirable to use restrictive thresholds, since an ostensibly low IAA score could be due to a host of factors, including personal bias. The authors therefore suggest reporting IAA statistics with confidence intervals. However, narrower confidence intervals (suggesting that the "true" IAA score, of which the IAA score obtained in a study is an estimate, is subject to less variability) would normally be expected with large samples (e.g., 1000 or more comparisons McHugh, 2012), which are well beyond most sizes reported in our overview (Section 3.3).

When the goal of an evaluation is to identify potential problems with output texts, a low IAA, indicating variance among annotators, can be highly informative (Amidei et al., 2018b). On the other hand, low IAA in evaluations of text quality can also suggest that results should not be extrapolated to a broader reader population. An additional consideration is that some statistics (such as κ; see McHugh, 2012) make overly restrictive assumptions, though they have the advantage of accounting for chance agreement. **Thus, apart from reporting such agreement metrics, it is advisable to also report percentage agreement, which is easily interpretable** (McHugh, 2012), **or a correlation statistic** (Amidei et al., 2019a). This combination of statistics could give a more complete overview regarding the reliability of annotator judgments. There is, unfortunately, no such thing as a perfect IAA metric, and choosing a metric might come down to personal considerations regarding which IAA metric characteristics are important. Therefore, we encourage publishing the annotator data on a digital repository, or as supplementary material to the paper, so that others may check the agreement using other metrics (see also Section 4.9).

Sample size When doing a qualitative study or an error-analysis study, which can be done with a small sample, good advice is provided by Van Enschot et al.: **difficult coding tasks (which most NLG evaluations are) require three or more annotators (though preferably more; see** Potter and Levine-Donnerstein, 1999); **more straightforward tasks can do with two to three.** This is corroborated by Snow et al. (2008) who investigated the reliability of annotator judgments on a variety of NLP labelling tasks. The authors found that for most tasks (e.g. temporal ordering, word sense disambiguation) 2 to 3 annotations per item resulted in reliable results. Affect recognition, a more difficult task, required around 4 labels per item.

In the case of large-scale studies, Brysbaert (2019) recently found that **most studies with less than 50 participants are underpowered and that for most designs and analyses 100 or more participants are needed** (for a discussion on the effects of underpowered samples, see Button et al., 2013; Makin and Orban de Xivry, 2019). The number of participants necessary can be decreased by having multiple observations per condition per participant (i.e., having participants perform more judgments). It is also possible to calculate the minimum number of participants needed based on the properties of the study (e.g. the statistical test that will be used, number of groups, etc) using a program such as G*Power (https://www.gpower.hhu.de/) (Faul et al., 2007; 2009).

Sampling items to judge Finally, some of the concerns listed above also hold for the items that the participants are asked to judge. If we want the judgments to be reflective of system performance, we need to make sure that the selected outputs are representative. But what does it mean for a sample of texts to be representative? Word frequencies typically follow a Zipfian distribution (Corral et al., 2015; Van Heuven et al., 2014; Zipf, 1949), meaning that a small part of the vocabulary accounts for the largest part of the data. Similarly, training data may be biased towards specific entities, while providing very little (if any) information about others (Ilievski et al., 2016). These properties of the data typically lead to NLP systems performing much better on the head, than on the tail of the distribution (e.g. Postma et al., 2016). The quality and the range of possible outputs will be heavily dependent on the range of inputs provided to the system. Based on these observations, **we recommend to evaluate NLG systems using a stratified sample of the output, consisting of examples that were generated using low, medium, and high-frequent inputs.** One might also consider categorizing different outputs along different frequency bands, to determine the performance of the system for words and phrases that are more or less frequent in the training data (cf. van Miltenburg et al., 2018). Or consider looking at edge cases, which can be especially relevant for a qualitative analysis to determine when and where a system fails. Finally, there should also be a sufficient number of outputs, so that a couple of particularly good or bad items do not skew the results too much. However, the number of items to evaluate, depends heavily on the diversity of the sample, so we cannot give any specific recommendations here.

4.6. Design

Satisficing As stated above, the research design is tightly interconnected with the research question, as the function of a research design is to ensure that the obtained results provide an answer to the research question. A carefully thought-out design also results in the most reliable data. For every evaluation question, participants have to retrieve relevant information from memory, and summarize this information into a single judgment (Krosnick, 2018b). When participants do this in a cursory fashion, or skip the retrieval process entirely, thereby engaging in a form of "satisficing" behaviour, results could be biased. Satisficing is primarily affected by (i) respondent ability, (ii) respondent motivation, and (iii) task difficulty (Krosnick, 2018b). Thus, it is always important to carefully monitor these aspects and **try to keep the evaluation task as simple and motivating for participants as possible.** We will now describe more detailed ways in which design choices could minimize bias.

Order effects Few papers report exact details of the design of their human evaluation experiments, although most indicate that multiple systems were compared and annotators were shown all examples. This suggests that within-subjects designs are a common practice.

Within-subjects designs are susceptible to order effects: over the course of an experiment, annotators can change their responses (that is, respond differently to items which are identical, or which are exemplars of identical conditions), and start to employ satisficing strategies (e.g. selecting "I don't know" or neutral answers, answering using mental coin-flipping, selecting the first sensible response) due to fatigue, practice, carryover effects or other (external) factors. If the order of presentation of the outputs of systems is fixed, differences found between systems may be due to order effects rather than differences in the output itself. To mitigate this, researchers can implement measures in the task design. **Practice effects can be reduced with a practice trial** in which examples of both very good (fluent, accurate, grammatical) and very bad (disfluent, inaccurate, ungrammatical) outputs are provided before the actual rating task. This allows for the participants to calibrate their responses, before starting with the actual task. **Carryover effects can be reduced by increasing the amount of time between presenting different conditions** (Shaughnessy et al., 2006). Alternatively, a filler task or filler questions can be employed in between conditions, although this may introduce biases if participants perceive the information obtained from the filler as relevant to performing the main task (de Quidt et al., 2019). **Fatigue effects can be reduced by shortening the task**, although this also means more participants are necessary since fewer observations per condition per participant means less statistical power (Brysbaert, 2019). This issue can be alleviated by structuring the data collection longitudinally: having the same group of participants perform the evaluation more than once, with a set amount of time between the evaluation sessions. Another way to tackle fatigue effects sometimes seen in research is to remove all entries with missing data, or to remove participants that failed 'attention checks' (or related checks e.g. instructional manipulation checks, or trap questions) from the sample. However, the use of attention checks is subject to debate, with some researchers pointing out that after such elimination procedures, the remaining cases may be a biased subsample of the total sample, thus biasing the results (Anduiza and Galais, 2016; Bennett, 2001; Berinsky et al., 2016). Experiments show that excluding participants who failed attention checks introduces a demographic bias, and attention checks actually induce low-effort responses or socially desirable responses (Clifford and Jerit, 2015; Vannette, 2016).

Order effects can also be reduced by presenting conditions in a systematically varied order. Counterbalancing is one such measure. With counterbalancing, all examples are presented in every possible order. While such a design is the best way to reduce order effects, it quickly becomes expensive. When annotators judge 4 examples, 4! = 24 different orders should be investigated. This issue can be partially mitigated by using incomplete counterbalancing, such as Latin Square. With a Latin Square, participants and items are rotated such that each participant is exposed to each condition, possibly with different items, an equal number of times (Dean and Voss, 1999). However, such a design can still induce order effects (with rotation, you would find that condition B follows condition A, and precedes condition C, most of the times), which is why the "Balanced Latin Square" (rotation happens via a more thorough formula, see (Kantowitz et al., 2008) is generally the preferred counterbalancing method, as it lowers the risk of these order effects occurring (Shuttleworth, 2009). In most cases, however, randomising the order of examples should be sufficient.

Another possibility is to use a between-subjects design, in which the subjects only judge the (randomly ordered) outputs of one system. **When order effects are expected and a large number of conditions are investigated, a between-subjects design is preferable** (Shaughnessy et al., 2006). Note, however, that between-subjects designs typically require larger samples, to obtain the same number of observations as with within-subjects designs. Also, despite the risks of order effects, **it is usually preferable to use a within-subject over a between-design (if this design is feasible)**. Direct comparisons done with a within-subjects design generally give rise to less variation than comparisons with a between-subjects design (Normand, 2016). Furthermore, if two or more NLG systems are compared that have systematically different behavior, seeing texts from all systems helps prevent that participants anchor their answers based only on the behavior of one of the systems, without considering the behaviors of the other systems.[14]

Web design With an online questionnaire—now the most commonly used mode of dissemination—it is often possible to make certain choices in terms of the appearance and layout of the online survey (hereafter: web design). These choices should be considered as well, as they can greatly affect the response outcomes of the questionnaire. Especially if the goal is to make the questionnaire accessible for a broad group of people, the appearance and layout matters (Goegan et al., 2018). Three main effects of web design choices on response outcomes can be observed: item response rates, dropout rates, and quality of responses. An example of web design choices affecting item response rates are studies comparing a questionnaire presented on a single, long scrolling page, to a questionnaire broken down into multiple pages. Multiple studies found that item non-response became higher when the questionnaire was presented on a single, long scrolling page (Manfreda et al., 2002; Tourangeau et al., 2004; Peytchev et al., 2006).

For dropout rates, higher dropout rates were found for long surveys that included an always visible survey progress indicator (Crawford et al., 2001), and for surveys that forced participants to respond to every question (giving participants an error page when they skipped certain questions) (Stieger et al., 2007).

Regarding the quality of responses, Novikova et al. (2018) found that when evaluating text criteria, answers to questions about different criteria tend to correlate when they are presented simultaneously on a single page. When participants are shown an item multiple times and questioned about each text criterion on separate pages, this correlation is reduced. Another way to reduce correlation between text criteria is to show or hide the aligned input when rating the textual output, depending on the criterion being judged. When rating correctness, for example, the input representation is necessary, while this input could be a distraction when rating fluency. Note that this is only relevant for NLG evaluation tasks where showing the input might be relevant.

Therefore it is useful to take a careful look at the web design of the online survey to check if it minimizes the possibility for nonresponse bias (see also Vicente and Reis, 2010).

4.7. Statistics and data analysis

Statistics turn quantitative data into useful information that help to answer the research question or hypotheses. Research design and statistics are strongly connected, as a good research design facilitates the collection of valid data, and statistics helps to interpret this data (Whitley and Kite, 2013). A distinction can be made between descriptive statistics and inferential statistics. Descriptive statistics (e.g. means, frequencies) can be used to summarize data, while inferential statistics identify important patterns, relationships, and differences between groups (McLeod, 2019). It is outside the scope of this paper to provide a complete overview of statistical tests that can be used for quantitative evaluation research (and Dror et al., 2018 already provides a comprehensive overview for the NLP domain). We will therefore take a more general focus on the applicability of standard null-hypothesis significance testing (NHST).

Comparing systems A majority of NLG papers do not report inferential statistics (see Section 3.4), only reporting descriptive statistics from their evaluation study. However, a common scenario involves NLG researchers comparing their own novel system against one or more ('state-of-the-art' or 'baseline') systems. In such a scenario, a Paired Student's *t*-test (if compared against one system), or Analysis of Variance test (if compared against multiple systems) would be used.[15] However, many researchers would extend this by testing, for instance, various versions of their own novel system (e.g. with or without output variation, or relying on different word embedding models) to compare them to each other, to some other ('state-of-the-art') systems, and/or with respect to one or more baselines. Notice that this quickly gives rise to a rather complex statistical design with multiple factors and multiple levels. Ironically, with every system or baseline that is added to the evaluation, the comparison becomes more interesting but the statistical model becomes more complex, and power issues become more pressing (Cohen, 1988; Button et al., 2013). However, statistical power—the probability that the statistical test will reject the null hypothesis (H0) when the alternative hypothesis (H1, e.g., that your new NLG system is the best) is true—is seldom (if ever) discussed in the NLG literature. These complexities in terms of design could quickly lead to errors, and there is a growing awareness that statistical tests are often conducted incorrectly, both in NLP (Dror et al., 2018) and in behavioral sciences more generally (e.g., Wagenmakers et al., 2011).

In fact, in the scenario sketched above there may be multiple (implicit) hypotheses: new system better than state-of-the-art, new system better than baseline, etcetera. When testing multiple hypotheses, the probability of making at least one false claim (incorrectly rejecting H0) increases (such errors are known as false positives or Type I errors). Various remedies for this particular

[14] See Field and Hole (2002) for a more detailed guide on design.

[15] Or non-parametric variants if the assumption of normality is violated, although it has been argued that *t*-tests and ANOVAs are robust for highly skewed non-normal distributions and small sample sizes (Norman, 2010).

problem exist, one being an application of the simple Bonferroni correction, which amounts to lowering the significance threshold (or α-level)—commonly $p < 0.05$, but see for example Benjamin et al. (2018) and Lakens et al. (2018)—to α/m, where m is the number of hypotheses tested. This procedure is not systematically applied in NLG, although the awareness of the issues with multiple comparisons is increasing.

The significance of a result is sometimes confused with its (practical) importance, but arguably this is a misconception. Especially with larger samples, small differences in performance between systems can quickly become statistically significant, but significance in itself is not particularly informative about how much better one system is compared to another. This information is captured by effect size estimates. Effect sizes (e.g., η^2, Cohen's d) can also add meaningful information about the desirability of adopting a new system, by highlighting what the trade-off is between performance gains and computational resources (Søgaard, 2013). Therefore, NLP researchers have stressed the importance of reporting effect sizes (e.g., Köhn, 2020). It should be noted that effect sizes can only be reliably captured for large sample sizes. For small sample sizes, only large effects can be detected. Meaning that with small sample sizes, there is high uncertainty about the true effect size, and that the actual effect size is likely overestimated (Button et al., 2013).

Assumptions Statistical tests are associated with assumptions about their applicability. One is the independence assumption (especially relevant for t-tests and ANOVAs, for example), which amounts to assuming that the value of one observation obtained in the experiment is unaffected by the value of other observations. This assumption is difficult to guarantee in NLP research (Dror et al., 2018), if only because different systems may rely on the same training data. In view of these issues, some have even argued that NHST should be abandoned (Koplenig, 2017; McShane et al., 2019).

Exploratory versus confirmatory We do believe there is a place for NHST, but in our opinion, the distinction between exploratory (hypothesis generating) and confirmatory (hypothesis testing) research should be taken more seriously. Much human evaluation of NLG could better be approached from an exploratory perspective. NLG researchers might fail to observe phenomena because of the focus on system comparison (which is compatible with a hypothesis-testing orientation), rather than on hypothesis generation; an example of confirmation bias. **If exploratory research is conducted, it would also be more appropriate to analyse findings with exploratory data analysis techniques** (Tukey, 1980; Cumming, 2013). **When researchers do have clear hypotheses, statistical significance testing (in combination with effect size estimates) can be a powerful tool.**

Finally, alternative statistical models deserve more attention within NLG. For example, within psycholinguistics it is common to look both at participant and item effects as potential sources of random variation (Clark, 1973). This would make a lot of sense in human NLG evaluations as well, because it might well be that a new NLG system works well for one kind of generated item (short active sentences, say) and less well for another kind (complex sentences with relative clauses). Mixed effects models capture such potential random effects very well (e.g., Barr et al., 2013), and deserve more attention in NLG.[16] Meteyard and Davies (2020) offer a recent survey of best practices for conducting and reporting mixed effect models. Furthermore, Bayesian models are worth exploring, because they are less sensitive to the aforementioned problems with NHST (e.g., Gelman et al., 2006; Wagenmakers, 2007).

Preregistration and HARKing If the goal, hypotheses, and research questions have been clearly delineated beforehand, it is good practice to **consider preregistration.** Preregistration is still uncommon in NLG and other fields of AI (with a few notable exceptions, like for instance Vogt et al., 2019), but it addresses an important issue with human evaluations. Conducting and analysing a human experiment is like entering a garden of forking paths (Gelman and Loken, 2013): along the way researchers have many choices to make, and even though each choice may be small and seemingly innocuous, collectively they can have a substantial effect on the outcome of the statistical analyses, to the extent that it becomes possible to present virtually every finding as statistically significant (Simmons et al., 2011; Wicherts et al., 2016). In human NLG evaluation, choices may include for instance, termination criteria (when does the data collection stop?), exclusion criteria (when is a participant removed from the analysis?), reporting of variables (which dependent variables are reported?), etcetera. By being explicit beforehand (i.e., by preregistering), any flexibility in the analysis (be it intentional or not) is removed.

In the past, hypothesizing after results are known—also known as 'HARKing'—was widespread practice (even recommended practice by some, see Bem et al., 1987). Researchers would, for instance, tweak their measures and samples when analyzing data so that they can report the outcomes in such a way that it best supports their story ('cherry-picking'), or use several different constructs, criteria, scale types, etc. and only report those that provide interesting results ('question trolling') (Lindsay et al., 2016). Nowadays, most researchers disapprove of such HARKing behaviors, as it can bias the results and have a substantial impact on the conclusions of a study (Murphy and Aguinis, 2019), not to mention that the impact of testing multiple hypotheses is to increase the likelihood of Type I errors, as noted above. By preregistering the study's method and design beforehand, it becomes more difficult to indulge in such behaviors.

Preregistration is increasingly common in medical and psychological science, and even though it is not perfect (Claesen et al., 2019; Lindsay et al., 2016) at least it has made research more transparent and controllable, which has a positive impact on the possibilities to replicate earlier findings. Platforms that offer online preregistration forms include the Open Science Framework (https://osf.io) and AsPredicted (https://aspredicted.org).

[16] Mixed effects models may also be used to study variation in elicitation tasks that are used to construct NLG datasets (van Miltenburg et al., 2019).

4.8. Practical matters

Evaluation platform After choices have been made regarding the aspects mentioned in the paragraphs above, it is time to prepare the evaluation study for publishing. A first consideration would be to pick an evaluation platform. The most important factor for this choice is the time and cost constraints associated with the evaluation study. Online crowdsourcing platforms (e.g. Qualtrics, MTurk, Figure8, Prolific) may offer the opportunity to get sufficient responses quickly, but participants on these platforms need to receive monetary compensation for participation (at least minimum wage compensation, ideally (Silberman et al., 2018), and there is the risk of inadvertently recruiting bots or participants that want to immediately engage in satisficing behavior so that they get paid for as little work as possible. Alternatively, one may recruit participants from the university, or recruit participants in public spaces (e.g. in the library, in the hospital waiting room, on the train) which could be entirely voluntary (although monetary incentives do increase response rates (Yu et al., 2017), but would mean that recruitment takes more time. Similarly, the choice for pre-developed online survey software (e.g. Qualtrics, SurveyMonkey, Google forms) or a custom developed online survey website should be made with regards to time and cost constraints, as well as the design challenges discussed in Section 4.6.

Consent form Consent forms are documents providing information to participants regarding the study. This information empowers participants to agree to the study knowing the roles and responsibilities of themselves and the researcher. As such, consent forms enable morally integrous research and have become mandatory for many research institutes. When signing the consent form, participants indicate that they (i) understand that they will participate in the evaluation study; (ii) understand the rights they have regarding the study; (iii) understand the risks and potential benefits associated with the evaluation study. The consent form also provides evidence that the participants were made aware of their rights, and the associated risks and benefits (Sterling, 2018). They should be developed carefully, as previous research found that consent forms are often too complex and too long, and often do not provide sufficient information about the decisions that participants are about to make (see Flory and Emanuel, 2004; Marshall, 2006; Paasche-Orlow et al., 2003; Sachs et al., 2003; Stunkel et al., 2010). **Therefore, it is imperative that the consent form is kept as simple, short, and clear as possible, and that lay language is used as much as possible.**

The same advice goes for the debriefing statement at the end of the survey, where participants are thanked for their participation. This statement should also provide (i) the research question, and/or hypotheses together with background on why and how this research question and/or hypotheses are being studied; (ii) information for participants on withdrawal procedures and on the opportunity to withdraw their data from the study; (iii) information on the opportunity to be informed of the study's results; and (iv) additional resources, such as contact information for the IRB (in case of ethical concerns), contact information of the researchers, and relevant research references.

Ethical clearance and data protection Researchers have a duty of care towards human participants. To ensure that this duty is taken seriously, many research institutions nowadays require formal ethical clearance before allowing researchers to carry out their studies. **It is important to obtain ethical clearance from the institution that the researcher works in (as well as clearance from the institution where the research is done, if applicable), and to state this explicitly in publications**, so as to ensure that the study is in accordance with the ethical requirements of the institution. If your organisation does not have an institutional review board, Rice (2008) notes that there are three options: start an IRB at your institution or organisation, submit your research proposal to an external, commercial IRB, or collaborate with partners who already have access to an IRB. Ethical rules and guidelines differ per country, so in cases where there is no IRB at your institution or organization, approval should be sought from an ethics board which understands the ethical rules and guidelines for your country.

Ethical clearance forms, as well as consent forms for participants, typically include parts meant to ensure that the handling of results complies with data protection legislation. Legislation regarding data protection varies for different parts of the world. For instance, research institutes in the European Union have to adhere to the General Data Protection Regulation (GDPR).[17] What is important is to make choices regarding data collection and storage transparent: the research institute, as well as the participants, should be aware of what will happen with the collected data. **Information should be given about how the data will be stored, in which circumstances the data can be shared and which data can be shared, and how these data sharing and storage decisions take the privacy of the participant into account** (Howitt, 2016).

Research institutes usually have data protection officers, who should be consulted if there are questions about adherence to the data protection legislation. All organisations should have someone who monitors GDPR compliance (Wolford). It is recommended to have this person check the research proposal as well. Note that a check for GDPR compliance may already be part of the IRB process. **Above all, it is recommended to explicitly state when publishing the findings that ethical clearance was obtained, and that the research is in compliance with the relevant data protection legislation.** Often, research institutes give an "Ethical Approval Code". If this code is reported, together with the Institutional Review Board from which the code was obtained, other researchers could acquire information about the ethical clearance without having to consult the paper's authors.

Beyond these practical recommendations, it is important to consider the broader ethical implications of NLG research and human evaluations in particular. These are discussed further below, in Section 4.10.

Pretesting After obtaining ethical clearance, researchers may start with their research. But **it is advisable to pilot the evaluation task before deploying it more widely**. In the case of survey research, pretesting helps to ensure that measures are as valid as possible, while minimizing measurement error (Colbert et al., 2019). In other words, pretesting could help eliminate

[17] https://gdpr.eu/

misinterpretation of a question, participant confusion, or clarity issues regarding the meaning of a question (Willis, 2004). A pretest is also the best moment to do a manipulation check when experimental research is performed (i.e. to check if the independent variables were successfully manipulated; see Ejelöv and Luke (2020) for recommendations on when and how to use manipulation checks) (Hauser et al., 2018).

There are various ways that a questionnaire can be pre-tested (e.g. cognitive interviews, expert panels, questionnaire appraisal tools; see Colbert et al., 2019, for an overview). While all these forms of pretesting have their limitations, using a combination of two or more pretesting methods can overcome these (Blake, 2014). Constraints regarding time and costs are also to be considered when choosing to pretest or choosing which method of pretesting to use. Furthermore, pretesting questionnaires is mostly recommended when describing new instruments. For pre-existing instruments, pretesting is not as important.

4.9. Transparency and replicability

Within the field of Machine Learning, awareness about the importance of replicability and transparency is starting to grow (see, for instance, (Emmery et al., 2019), for a review on current issues that are stalling the progress towards open science). While Machine Learning, and NLG research as well, are at the forefront of open science in some ways (e.g. publicly available datasets, community-driven shared tasks, commonly sharing the full research code, tendency towards open-access publications (Munafò et al., 2017), they still have some challenges to overcome. After an evaluation study is conducted, it is good practice to save all the raw data and materials used for the study (e.g., the questionnaire, consent form, annotator instruction manual), and make this available to other researchers (Ruggles et al., 2003), but this is rarely done. Sometimes, researchers state that all data is "available upon request", but these requests are not always fulfilled (Krawczyk and Reuben, 2012). Alternatively, data and materials can be shared using online platforms such as Zenodo (https://zenodo.org/), the Open Science Framework (https://osf.io) and Figshare (https://figshare.com/), or as supplementary materials in the ACL Anthology. This advice also extends to code that is published on GitHub, as the mentioned online platforms and the ACL Anthology focus on preserving data for the long term, which is a focus that GitHub currently lacks. But while the practice of publishing participant response data and evaluation materials is becoming increasingly common in the behavioral sciences, it is not seen as often in NLG.

One worry might be ethical clearance and ethical management of data issues, in that it might be ethically difficult to share this information, and also to publish demographic details about the participants. However, generally this can be done as long as participants consent to the possibility that data collected from them can be shared. Furthermore, in most cases, if personal identifiers (any information which can be identified as being about a particular individual) are removed from the data, then the data protection legislation no longer applies (Howitt, 2016). However, it should be noted that this advice is based on the state of affairs in late 2020 and may be subject to change. Therefore, consult with data protection experts to find out best practice and legal constraints regarding management of data. **In any case, we recommend trying to make the raw data, as well as all materials related to the evaluation research, publicly available.** At the same time, we acknowledge that it is good practice to be cautious about the anonymization measures. Removing too much data could remove important contextual information, meaning that useful information becomes unavailable for other researchers (Nespor, 2000). But sometimes not enough information is removed in the anonymization process, making it possible to identify participants (Ayers et al., 2018). "Differential privacy" has emerged as a paradigm that allows researchers or companies to access sensitive data without breaching the privacy of individuals (Yang et al., 2012). Individuals' information is protected by applying an algorithm that injects random noise into the sensitive parts of the dataset, such as Laplace mechanism (Dwork, 2006). This provides individuals with plausible deniability, since the private information found in their record may be injected random noise.

4.10. Ethical considerations

The foregoing discussion made some practical recommendations on steps to obtain ethical clearance and ensure compliance with data protection and privacy regulations. That discussion focused entirely on issues related to evaluation studies involving human participants. The recommendations made above ultimately boil down to three main principles, namely Transparency, Protection of Privacy and Non- maleficence. It is worth noting that all three emerged as salient in a recent review of AI ethics guidelines published worldwide by different entities in the private and public sectors (Jobin et al., 2019).[18]

The purpose of this section is to discuss the implications of these issues not only in terms of their relevance for evaluation of developed systems, but also for NLG research in general. This seems timely in view of ongoing debates on ethics in AI, and more specifically in NLP, as reflected, for example, by the establishment of the ACL Workshop on Ethics in NLP, which at the time of writing has gone through two editions.

Where evaluation studies are concerned, ethical considerations become more acute with systems intended to be tested (or even deployed) in real settings, with participants sampled from a specific target population. Considerations that arise in these contexts include (a) whether the demands made on participants are reasonable; (b) whether special measures need to be taken to minimise the possibility of harm. As an example, consider an NLG system intended for use in a fault-critical decision-support

[18] Broadly, *transparency* refers to understandability and/or explainability of systems or procedures; *non-maleficence* refers to the avoidance of harm; while *privacy* relates to the use of personal data, which in our case is data obtained from human participants with their consent. It is worth noting that non-maleficence is more frequently cited in the guidelines under review by Jobin et al. (2019) than *beneficence*, which implies benefits or enhanced well-being to humans, something which most studies are typically not equipped to guarantee.

context. In an experimental evaluation (that is, prior to actual deployment in production), such a system may well result in faulty decisions due to errors or ambiguities. These can have negative consequences even though the setting is experimental, necessitating special steps to supervise the evaluation and make its experimental nature transparent to participants. Such steps were taken in the evaluation of the BT-Nurse system (Hunter et al., 2012), during its period of deployment on-ward: nurses were asked to interact with the system and give feedback under the supervision of an experienced research nurse. It was also made clear that the output was intended for them to judge, and not for them to use in patient treatment.

As for the broader ethical implications of NLG systems, perhaps the very question of whether a system should be developed in the first place constitutes a form of evaluation in its own right, albeit not so much of the system itself as of its underlying purpose and guiding ideas. For this purpose, focus groups could be very useful. It is important to acknowledge that some systems just should not be built (Baumer and Silberman, 2011), and the advantage of ethical analysis is that it can be carried out before development. Obviously, whether or not a system or model is likely to be misused is hard to ascertain in advance. However, the relevance of this question is becoming clearer as large-scale NLG models are developed which, by virtue of their high adaptability to multiple tasks and low cost for non-expert users, can easily be deployed to generate high-volume content easily and cheaply for nefarious purposes, such as spreading propaganda or misinformation (see Mcguffie and Newhouse (2020) for a recent study on possible misuse of the GPT-3 model (Brown et al., 2020).

In this light, it is worth considering recent guidelines dealing with ethics in information systems. Smiley et al. (2016) proposed an 'ethics checklist for NLG systems'. Although a checklist may be somewhat reductionist, it does provide a useful guideline for NLG system development. A more general approach is *Value-Sensitive Design*, which asks developers to consider the implications of technology for both direct and indirect stakeholders (Friedman et al., 2013).[19]

Therefore, before development begins, it is important to consider the ethical implications of the proposed system, with reference to ethics checklists for NLG and a consideration of potential impact in stakeholders.

5. General discussion and conclusion

We have provided an overview of the current state of human evaluation in NLG, and presented a set of best practices. For convenience, Table 5 provides a summary of the main recommendations arising from the foregoing discussion. This is a broad topic, but we have attempted to be as complete as possible in our overview. While some of the suggested practices are already common, others are still rarely attested in NLG. Do note that the majority of recommendations pertain to the most common human evaluation research in NLG (that is: an intrinsic, controlled, and quantitative research). Other types of NLG research (extrinsic, uncontrolled, qualitative research) have not received as much attention in the literature thus far, and deserve to be explored further in future work.

Readers that have come this far in the paper will not be surprised to hear that we find human evaluation of NLG output important, and believe it should be done more frequently and more systematically than currently is often the case. Partially, this is because good automatic evaluation metrics are still hard to come by. Sometimes human evaluations can be seen to function as a stand-in for evaluation dimensions which we cannot yet explain properly. Questions such as "is this fluent?" or "is this accurate?" are frequently asked to participants (although such questions are discouraged, see Section 4.3), knowing that these are vague, ill-defined terms, but assuming that people will have a reasonable understanding of these notions, which are not properly captured in automatic evaluation metrics. In some sense, with such human evaluations we may be kicking the can of explanation further down the line, until a better, possibly automated insight in these dimensions arrives. However, even when better automatic measures will become available, human evaluations will remain essential both to validate the metrics and to provide independent evidence of their usefulness in different application domains.

Conducting a proper human evaluation is time consuming and expensive, which may not be a welcome message in a fast-moving field such as ours. Indeed, we have been told things like "With improvements being as rapid as they are, it seems like a bad idea to put so much time into evaluation, because then by the time I can publish my results, my system is already obsolete." Of course, we sympathise with this sentiment, but at the same time we strongly believe that even in these cases proper human evaluations are important. How, after all, could we decide otherwise whether a seemingly small improvement over the state-of-the-art is actually meaningful for human readers? Of course, here it also depends on the precise goal of your algorithm or study, on the availability of evaluation data and/or other constraints (e.g., to evaluate the speed of your new algorithm no human participants are needed). But in general, it is fair to say that doing proper experimental studies with human participants is hard (and certainly harder than many people think). It could even be argued that we do too many of them. It is arguably better to conduct less experiments, but reliable, preregistered, and suitably powered ones, than more experiments which may be underpowered or unreliable and hence of less value.

We hope that NLG researchers will become more aware of the positive effects of high quality evaluations. A well-executed human evaluation, as well as methodological pluralism (intrinsic as well as extrinsic; qualitative and quantitative; exploratory and confirmatory, etc.), could help to broaden the understanding of the systems we develop (Lacity and Janson, 1994). If the task-related issues are clear, they open up new avenues for further development, which will benefit the progress in the field as

[19] For brainstorming about potential uses and abuses of NLG, one might use *envisioning cards* (Friedman and Hendry, 2012) or the *Tarot cards of Tech* (Artefact Group). Extending this idea even further, researchers in Human-Computer Interaction have recently started publishing *design fictions*: fictional scenarios to explore the ethical implications of different kinds of technology e.g., Lindley and Sharma, 2016; Baumer et al., 2018.

well. Therefore we hope that this overview will serve as a useful reference for NLG researchers, and help with carrying out and improving the quality of human evaluations in Natural Language Generation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We received support from RAAK-PRO SIA (2014-01-51PRO) and The Netherlands Organization for Scientific Research (NWO 360-89-050), which is gratefully acknowledged. Furthermore, we want to extend our gratitude towards the anonymous reviewers and also towards Leshem Choshen, Ondřej Dušek, Kees van Deemter, Dimitra Gkatzia, David Howcroft, Ehud Reiter, and Sander Wubben for their valuable comments on the paper.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.csl.2020.101151

Appendix A. Supplementary materials

Supplementary Data S1. Supplementary Raw Research Data. This is open data under the CC BY license http://creativecommons.org/licenses/by/4.0/

References

Amidei, J., Piwek, P., Willis, A., 2018. Evaluation methodologies in Automatic Question Generation 2013–2018. INLG 2018 307. https://doi.org/10.18653/v1/W18-6537.

Amidei, J., Piwek, P., Willis, A., 2018. Rethinking the agreement in human evaluation tasks. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3318–3329.

Amidei, J., Piwek, P., Willis, A., 2019. Agreement is overrated: a plea for correlation to assess human evaluation reliability. In: Proceedings of the 12th International Conference on Natural Language Generation. Association for Computational Linguistics, Tokyo, Japan, pp. 344–354. https://doi.org/10.18653/v1/W19-8642.

Amidei, J., Piwek, P., Willis, A., 2019. The use of rating and Likert scales in Natural Language Generation human evaluation tasks: a review and some recommendations. In: Proceedings of the 12th International Conference on Natural Language Generation. Association for Computational Linguistics, Tokyo, Japan, pp. 397–402. https://doi.org/10.18653/v1/W19-8648.

Ananthakrishnan, R., Bhattacharyya, P., Sasikumar, M., Shah, R.M., 2007. Some Issues in Automatic Evaluation of English-Hindi MT: More Blues for BLEU. ICON.

Anduiza, E., Galais, C., 2016. Answering without reading: IMCs and strong satisficing in online surveys. Int. J. Public Opin. Res. 29 (3), 497–519. https://doi.org/10.1093/ijpor/edw007.

Artefact Group,. The tarot cards of tech: Discover the power of predicting impact. URL: https://www.artefactgroup.com/case-studies/the-tarot-cards-of-tech/, retrieved: July 6, 2020.

Artstein, R., Poesio, M., 2008. Inter-coder agreement for computational linguistics. Comput. Linguist. 34 (4), 555–596. https://doi.org/10.1162/coli.07-034-R2.

Ayers, J.W., Caputi, T.L., Nebeker, C., Dredze, M., 2018. Don't quote me: reverse identification of research participants in social media studies. NPJ Digit. Med. 1 (1), 1–2. https://doi.org/10.1038/s41746-018-0036-2.

Bainbridge, L., 1990. Verbal protocol analysis. In: Corlett, J.R.W.E.N. (Ed.), Evaluation of Human Work. A Practical Ergonomics Methodology. Taylor and Francis, London, UK, pp. 161–179.

Banerjee, S., Lavie, A., 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Association for Computational Linguistics, Ann Arbor, Michigan, USA, pp. 65–72.

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. J. Mem. Lang. 68 (3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Barrault, L., Bojar, O., Costa-jussà, M.R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., Zampieri, M., 2019. Findings of the 2019 conference on machine translation (WMT19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). Association for Computational Linguistics, Florence, Italy, pp. 1–61. https://doi.org/10.18653/v1/W19-5301.

Baumer, E.P., Berrill, T., Botwinick, S.C., Gonzales, J.L., Ho, K., Kundrik, A., Kwon, L., LaRowe, T., Nguyen, C.P., Ramirez, F., Schaedler, P., Ulrich, W., Wallace, A., Wan, Y., Weinfeld, B., 2018. What would you do? Design fiction and ethics. In: Proceedings of the 2018 ACM Conference on Supporting Groupwork. Association for Computing Machinery, New York, NY, USA, pp. 244–256. https://doi.org/10.1145/3148330.3149405.

Baumer, E.P., Silberman, M.S., 2011. When the implication is not to design (technology). In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 2271–2274. https://doi.org/10.1145/1978942.1979275.

Belz, A., Reiter, E., 2006. Comparing automatic and human evaluation of NLG systems. In: Proceedings of the11th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 313–320.

Bem, D.J., Zanna, M., Darley, J., 1987. Writing the empirical journal. The Compleat Academic: a Practical Guide for the Beginning Social Scientist, pp. 171–201.

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.-J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., et al., 2018. Redefine statistical significance. Nat. Hum. Behav. 2 (1), 6. https://doi.org/10.1038/s41562-017-0189-z.

Bennett, D.A., 2001. How can i deal with missing data in my study? Aust. N. Z. J. Public Health 25 (5), 464–469. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x.

Berinsky, A.J., Margolis, M.F., Sances, M.W., 2016. Can we turn shirkers into workers? J. Exp. Soc. Psychol. 66, 20–28. https://doi.org/10.1016/j.jesp.2015.09.010.

Bernhard, D., De Viron, L., Moriceau, V., Tannier, X., 2012. Question generation for french: collating parsers and paraphrasing questions. Dialogue Discourse 3 (2), 43–74. https://doi.org/10.5087/dad.2012.203.

Blaikie, N., 2000. Designing Social Research: The Logic of Anticipation. Wiley, New York, NY, USA.

Blake, M., 2014. Other pretesting methods. In: Collins, D. (Ed.), Cognitive Interviewing Practice. Sage Publications, Los Angeles, CA, USA, pp. 28–56. https://doi.org/10.4135/9781473910102.n2.

Bojar, O., Graham, Y., Kamran, A., 2017. Results of the WMT17 Metrics Shared Task. In: Proceedings of the Second Conference on Machine Translation. Association for Computational Linguistics, Copenhagen, Denmark, pp. 489–513. https://doi.org/10.18653/v1/W17-4755.
Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language Models are Few-Shot Learners. ArXiv. arXiv:2005.14165.
Brysbaert, M., 2019. How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. J. Cogn. 2 (1), 1–38. https://doi.org/10.5334/joc.72.
Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14 (5), 365. https://doi.org/10.1038/nrn3475.
Carr, L.T., 1994. The strengths and weaknesses of quantitative and qualitative research: what method for nursing? J. Adv. Nurs. 20 (4), 716–721. https://doi.org/10.1046/j.1365-2648.1994.20040716.x.
Castro Ferreira, T., van der Lee, C., van Miltenburg, E., Krahmer, E., 2019. Neural data-to-text generation: a comparison between pipeline and end-to-end architectures. 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). Association for Computational Linguistics, Hong Kong, China, pp. 552–562. https://doi.org/10.18653/v1/D19-1052.
Chen, A., Stanovsky, G., Singh, S., Gardner, M., 2019. Evaluating question answering evaluation. In: Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Association for Computational Linguistics, Hong Kong, China, pp. 119–124. https://doi.org/10.18653/v1/D19-5817.
Choshen, L., Abend, O., 2018. Inherent biases in reference-based evaluation for grammatical error correction and text simplification. In: Proceedings of 56th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Melbourne, Australia, pp. 632–642. https://doi.org/10.18653/v1/P18-1059.
Cicchetti, D.V., Shoinralter, D., Tyrer, P.J., 1985. The effect of number of rating scale categories on levels of interrater reliability: a Monte Carlo investigation. Appl. Psychol. Meas. 9 (1), 31–36. https://doi.org/10.1177/014662168500900103.
Claesen, A., Gomes, S.L.B.T., Tuerlinckx, F., et al., 2019. Preregistration: comparing dream to reality. PsyArXiv. https://doi.org/10.31234/osf.io/d8wex.
Clark, H.H., 1973. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. J. Verb. Learn. Verb. Behav. 12 (4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3.
Clifford, S., Jerit, J., 2015. Do attempts to improve respondent attention increase social desirability bias? Public Opin. Q 79 (3), 790–802. https://doi.org/10.1093/poq/nfv027.
Cohen, J., 1988. Statistical power analysis for the behavioral sciences. Routledge. https://doi.org/10.4324/9780203771587.
Colbert, C.Y., French, J.C., Arroliga, A.C., Bierer, S.B., 2019. Best practice versus actual practice: an audit of survey pretesting practices reported in a sample of medical education journals. Med. Educ. Online 24 (1), 1–11. https://doi.org/10.1080/10872981.2019.1673596.
Colby, K.M., Weber, S., Hilf, F.D., 1971. Artificial paranoia. Artif. Intell. 2 (1), 1–25. https://doi.org/10.1016/0004-3702(71)90002-6.
Corral, Á., Boleda, G., Ferrer-i Cancho, R., 2015. Zipf's law for word frequencies: word forms versus lemmas in long texts. PloS One 10 (7), e0129031.
Crawford, S.D., Couper, M.P., Lamias, M.J., 2001. Web surveys: perceptions of burden. Soc. Sci. Comput. Rev. 19 (2), 146–162.
Cumming, G., 2013. Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. Routledge. https://doi.org/10.4324/9780203807002.
De Vaus, D., 2001. Research Design in Social Research. Sage Publications, Thousand Oaks, CA, USA.
Dean, A., Voss, D., 1999. Design and analysis of experiments. 1. Springer, New York, NY, USA.
Dell'Orletta, F., Montemagni, S., Venturi, G., 2011. READ-IT: assessing readability of Italian texts with a view to text simplification. In: Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies. Association for Computational Linguistics, pp. 73–83.
Denkowski, M., Neubig, G., 2017. Stronger baselines for trustable results in neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation. Association for Computational Linguistics, Vancouver, pp. 18–27. https://doi.org/10.18653/v1/W17-3203.
Denscombe, M., 2010. The Good Research Guide: For Small-Scale Social Research Projects. McGraw-Hill Education (UK).
Di Eugenio, B., Glass, M., Trolio, M., 2002. The DIAG experiments: natural language generation for intelligent tutoring systems. In: Proceedings of the International Natural Language Generation Conference. Association for Computational Linguistics, Harriman, New York, USA, pp. 120–127.
Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., Kaiser, S., 2012. Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective. J. Acad. Mark. Sci. 40 (3), 434–449. https://doi.org/10.1007/s11747-011-0300-3.
Doddington, G., 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc., pp. 138–145. https://doi.org/10.3115/1289189.1289273.
Dodge, J., Gururangan, S., Card, D., Schwartz, R., Smith, N.A., 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). Association for Computational Linguistics, Hong Kong, China, pp. 2185–2194. https://doi.org/10.18653/v1/D19-1224.
Dror, R., Baumer, G., Shlomov, S., Reichart, R., 2018. The Hitchhiker's guide to testing statistical significance in natural language processing. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1383–1392. https://doi.org/10.18653/v1/P18-1128.
Dušek, O., Jurčíček, F., 2019. Neural generation for Czech: data and baselines. In: Proceedings of the 12th International Conference on Natural Language Generation. Association for Computational Linguistics, Tokyo, Japan, pp. 563–574. https://doi.org/10.18653/v1/W19-8670.
Dwork, C., 2006. Differential privacy. In: Proceedings of the 33rd International Colloquium on Automata, Languages and Programming. Springer, Venice, Italy, pp. 1–12. https://doi.org/10.1007/11787006_1.
Edwards, J., Perrone, A., Doyle, P.R., 2020. Transparency in language generation: levels of automation. In: Proceedings of the 2nd International Conference on Conversational User Interface. ACM, Bilbao, Spain, pp. 1–2. https://doi.org/10.1145/3342775.3342799.
Ejelöv, E., Luke, T.J., 2020. "rarely safe to assume": evaluating the use and interpretation of manipulation checks in experimental social psychology. J. Exp. Soc. Psychol. 87, 1–13. https://doi.org/10.1016/j.jesp.2019.103937.
Emmery, C., Kádár, Á., Wiltshire, T.J., Hendrickson, A.T., 2019. Towards replication in computational cognitive modeling: a machine learning perspective. Comput. Brain Behav. 2 (3–4), 242–246. https://doi.org/10.1007/s42113-019-00055-w.
Falkenjack, J., Mühlenbock, K.H., Jönsson, A., 2013. Features indicating readability in Swedish text. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), pp. 27–40.
Faul, F., Erdfelder, E., Buchner, A., Lang, A.-G., 2009. Statistical power analyses using g*power 3.1: tests for correlation and regression analyses. Behav. Res. Methods 41 (4), 1149–1160.
Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39 (2), 175–191.
Field, A., Hole, G., 2002. How to design and report experiments. Sage.
Flory, J., Emanuel, E., 2004. Interventions to improve research participants' understanding in informed consent for research: a systematic review. J. Am. Med. Assoc. 292 (13), 1593–1601. https://doi.org/10.1001/jama.292.13.1593.
Friedman, B., Hendry, D., 2012. The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1145–1148.
Friedman, B., Kahn Jr, P.H., Borning, A., Huldtgren, A., 2013. Value sensitive design and information systems. Early Engagement and New Technologies: Opening up the Laboratory. Springer, pp. 55–95.
Frith, C.D., Frith, U., 2012. Mechanisms of social cognition. Ann. Rev. Psychol. 63, 287–313.
Fulgoni, G., 2014. Uses and misuses of online-survey panels in digital research. J. Advert Res. 54 (2), 133–137.
Gabriel, R.P., 1988. Deliberate writing. In: McDonald, D.D., Bolc, L. (Eds.), Natural Language Generation Systems. Springer-Verlag, pp. 1–46.
Gatt, A., Krahmer, E., 2018. Survey of the state of the art in natural language generation: core tasks, applications and evaluation. J. Artif. Intell. Res. 61, 65–170. https://doi.org/10.1613/jair.5477.

Geiger, R.S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., Huang, J., 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, pp. 325–336. https://doi.org/10.1145/3351095.3372862.
Gelman, A., Loken, E., 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. Unpublished Manuscript.
Gelman, A., et al., 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Anal. 1 (3), 515–534. https://doi.org/10.1214/06-BA117C.
Gkatzia, D., Mahamood, S., 2015. A snapshot of NLG evaluation practices 2005−2014. In: Proceedings of the 15th European Workshop on Natural Language Generation (ENLG). Association for Computational Linguistics, pp. 57–60. https://doi.org/10.18653/v1/W15-4708.
Goegan, L.D., Radil, A.I., Daniels, L.M., 2018. Accessibility in questionnaire research: integrating universal design to increase the participation of individuals with learning disabilities.. Learn. Disabil.: Contemp. J. 16 (2), 177–190.
Goldberg, E., Driedger, N., Kittredge, R.I., 1994. Using natural language processing to produce weather forecasts. IEEE Expert 2, 45–53.
Green, P.E., Rao, V.R., 1970. Rating scales and information recovery: how many scales and response categories to use? J. Mark. 34 (3), 33–39. https://doi.org/10.1177/002224297003400307.
Grice, H.P., 1975. Logic and conversation. In: Morgan, P.C.,J. (Ed.), Syntax and Semantics: Speech Acts. 3, Brill, New York, NY, USA, pp. 43–58.
Han, B., Radford, W., Cadilhac, A., Harol, A., Chisholm, A., Hachey, B., 2017. Post-edit analysis of collective biography generation. In: Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, Perth, Australia, pp. 791–792. https://doi.org/10.1145/3041021.3054264.
Han, S., Lin, X., Joty, S., . Resurrecting submodularity in neural abstractive summarization. . arXiv preprint arXiv:1911.03014.
Harris, M.D., 2008. Building a large-scale commercial NLG system for an EMR. In: Proceedings of the Fifth International Natural Language Generation Conference (INLG ’08). Association for Computational Linguistics, Morristown, NJ, USA, pp. 157–160. https://doi.org/10.3115/1708322.1708351.
Hastie, H., Belz, A., 2014. A comparative evaluation methodology for NLG in interactive systems. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), pp. 4004–4011. https://doi.org/10.1017/CBO9780511844492.013.
Hauser, D.J., Ellsworth, P.C., Gonzalez, R., 2018. Are manipulation checks necessary? Front. Psychol. 9, 1–10. https://doi.org/10.3389/fpsyg.2018.00998.
Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? Behav. Brain Sci. 23, 61–83. https://doi.org/10.1017/S0140525X0999152X.discussion 83−135
Herbrich, R., Minka, T., Graepel, T., 2007. TrueSkill™: A Bayesian skill rating system. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 569–576.
Heyselaar, E., Bosse, T., 2019. Using theory of mind to assess users’ sense of agency in social chatbots. In: Proceedings of the Third International Workshop on Chatbot Research and Design (Conversations 2019). Springer, Amsterdam, The Netherlands, pp. 158–169. https://doi.org/10.1007/978-3-030-39540-7_11.
Hommes, S., van der Lee, C., Clouth, F., Vermunt, J., Verbeek, X., Krahmer, E., 2019. A personalized data-to-text support tool for cancer patients. In: Proceedings of the 12th International Conference on Natural Language Generation. Association for Computational Linguistics, Tokyo, Japan, pp. 443–452. https://doi.org/10.18653/v1/W19-8656.
Howitt, D., 2016. Introduction to Qualitative Research Methods in Psychology. 3 Pearson, Harlow, UK.
Hunter, J., Freer, Y., Gatt, A., Reiter, E., Sripada, S., Sykes, C., 2012. Automatic generation of natural language nursing shift summaries in neonatal intensive care: Bt-nurse. Artif. Intell. Med. 56 (3), 157–172.
Ilievski, F., Postma, M., Vossen, P., 2016. Semantic overfitting: what ‘world’ do we consider when evaluating disambiguation of text? In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. The COLING 2016 Organizing Committee, Osaka, Japan, pp. 1180–1191.
Jobin, A., Ienca, M., Vayena, E., 2019. The global landscape of AI ethics guidelines. Nat. Mach. Intell. 1 (9), 389–399. https://doi.org/10.1038/s42256-019-0088-2.
Jones, R.R., 1968. Differences in response consistency and subjects’ preferences for three personality inventory response formats. In: Proceedings of the 76th Annual Convention of the American Psychological Association, 3. American Psychological Association Washington, DC, pp. 247–248.
Kantowitz, B., Roediger, H., Elmes, D., 2008. Conditioning and learning. Experimental Psychology. Cengage Learning, pp. 227–260.International student edition
Kiritchenko, S., Mohammad, S., 2017. Best-worst scaling more reliable than rating scales: a case study on sentiment intensity annotation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Vancouver, Canada, pp. 465–470. https://doi.org/10.18653/v1/P17-2074.
Koehn, P., 2004. Statistical significance tests for machine translation evaluation. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Barcelona, Spain, pp. 388–395.
Köhn, A., 2020. We need to talk about significance tests. Retrieved from https://arne.chark.eu/2019/we-need-to-talk-about-significance-tests/ on March 10, 2020.
Koplenig, A., 2017. Against statistical significance testing in corpus linguistics. Corpus Linguist. Linguist. Theory. https://doi.org/10.1515/cllt.2005.1.2.277.
Krawczyk, M., Reuben, E., 2012. (Un) available upon request: field experiment on researchers' willingness to share supplementary materials. Account. Res. 19 (3), 175–186. https://doi.org/10.1080/08989621.2012.678688.
Kreminski, M., Dickinson, M., Mateas, M., Wardrip-Fruin, N., 2020. Why are we like this?: The AI architecture of a co-creative storytelling game. In: Proceedings of the 15th International Conference on the Foundations of Digital Games. ACM, Bugibba, Malta, pp. 1–4. https://doi.org/10.1145/3402942.3402953.
Krosnick, J.A., 2018. Assessing the accuracy of survey research. In: Krosnick, D.V.,J. (Ed.), The Palgrave Handbook of Survey Research. Palgrave Macmillan, Cham, Switzerland, pp. 3–5. https://doi.org/10.1007/978-3-319-54395-6_.
Krosnick, J.A., 2018. Improving question design to maximize reliability and validity. In: Krosnick, D.V.,J. (Ed.), The Palgrave Handbook of Survey Research. Palgrave Macmillan, Cham, Switzerland, pp. 95–101. https://doi.org/10.1007/978-3-319-54395-6_.
Kukich, K., 1988. Fluency in natural language reports. In: McDonald, D.D., Bolc, L. (Eds.), Natural Language Generation Systems. Springer-Verlag, pp. 280–305.
Lacity, M.C., Janson, M.A., 1994. Understanding qualitative data: a framework of text analysis methods. J. Manag. Inf. Syst. 11 (2), 137–155. https://doi.org/10.1080/07421222.1994.11518043.
Lakens, D., Adolfi, F.G., Albers, C.J., Anvari, F., Apps, M.A., Argamon, S.E., Baguley, T., Becker, R.B., Benning, S.D., Bradford, D.E., et al., 2018. Justify your alpha. Nat. Hum. Behav. 2 (3), 168. https://doi.org/10.17605/OSF.IO/9S3Y6.
Lan, T., Mao, X., Huang, H., Wei, W., . When to talk: Chatbot controls the timing of talking during multi-turn open-domain dialogue generation. . arXiv preprint arXiv:1912.09879.
Langsford, S., Perfors, A., Hendrickson, A.T., Kennedy, L.A., Navarro, D.J., 2018. Quantifying sentence acceptability measures: reliability, bias, and variability. Glossa: J. Gen. Linguist. 3 (1), 1–34. https://doi.org/10.5334/gjgl.396.
van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., Krahmer, E., 2019. Best practices for the human evaluation of automatically generated text. In: Proceedings of the 12th International Conference on Natural Language Generation. Association for Computational Linguistics, Tokyo, Japan, pp. 355–368. https://doi.org/10.18653/v1/W19-8643.
van der Lee, C., Verduijn, B., Krahmer, E., Wubben, S., 2018. Evaluating the text quality, human likeness and tailoring component of PASS: a Dutch data-to-text system for soccer. In: Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 962–972.
Lentz, L., De Jong, M., 1997. The evaluation of text quality: expert-focused and reader-focused methods compared. IEEE Trans.Prof. Commun. 40 (3), 224–234. https://doi.org/10.1109/47.649557.
Lin, C.-Y., 2004. ROUGE: A package for automatic evaluation of summaries. Text Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain, pp. 74–81.
Lincoln, Y.S., Guba, E.G., 2000. Paradigmatic controversies, contradictions, and emerging confluences. In: Lincoln, Y.S., Denzin, N.K. (Eds.), Handbook of Qualitative Research. Sage, Thousand Oaks, CA, USA, pp. 163–188.

Lindley, J., Sharma, D., 2016. Operationalising design fiction for ethical computing. SIGCAS Comput. Soc. 45 (3), 79–83. https://doi.org/10.1145/2874239.2874251.
Lindsay, D. S., Simons, D. J., Lilienfeld, S. O., 2016. Research preregistration 101. Retrieved from https://www.psychologicalscience.org/observer/research-preregistration-101/on September 21, 2020.
Lissitz, R.W., Green, S.B., 1975. Effect of the number of scale points on reliability: a Monte Carlo approach.. J. Appl. Psychol. 60 (1), 10. https://doi.org/10.1037/h0076268.
Ma, Q., Bojar, O., Graham, Y., 2018. Results of the WMT18 metrics shared task: both characters and embeddings achieve good performance. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. Association for Computational Linguistics, Belgium, Brussels, pp. 671–688. https://doi.org/10.18653/v1/W18-6450.
Ma, Q., Wei, J., Bojar, O., Graham, Y., 2019. Results of the WMT19 metrics shared task: segment-level and strong MT systems pose big challenges. In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). Association for Computational Linguistics, Florence, Italy, pp. 62–90. https://doi.org/10.18653/v1/W19-5302.
MacDonald, H., 2018. Numbers. Truth: How the Many Sides to Every Story Shape Our Reality. Random House, pp. 82–102.chapter 4
Maher, M.L., 2012. Computational and collective creativity: who's being creative? In: Proceedings of the Third International Conference on Computer Creativity. Association for Computational Linguistics, Dublin, Ireland, pp. 67–71.
Makin, T.R., Orban de Xivry, J.-J., 2019. Science forum: ten common statistical mistakes to watch out for when writing or reviewing a manuscript. eLife 8, 1–13. https://doi.org/10.7554/eLife.48175.
Manfreda, K.L., Batagelj, Z., Vehovar, V., 2002. Design of web survey questionnaires: three basic experiments. J. Comput.-Mediated Commun. 7 (3), 1–28. https://doi.org/10.1111/j.1083-6101.2002.tb00149.x.
Manjavacas, E., Karsdorp, F., Burtenshaw, B., Kestemont, M., 2017. Synthetic literature: writing science fiction in a co-creative process. In: Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017). Association for Computational Linguistics, Santiago de Compostela, Spain, pp. 29–37. https://doi.org/10.18653/v1/W17-3904.
MArchegiani, D., Perez-Beltrachini, L., 2018. Deep graph convolutional encoders for structured data to text generation. In: Proceedings of the 11th International Natural Language Generation Conference (INLG'18). Association for Computational Linguistics, Tilburg, The Netherlands, pp. 1–9.
Marshall, P.A., 2006. Informed consent in international health research. J. Empir. Res. Hum. Res. Ethics 1 (1), 25–42. https://doi.org/10.1525/jer.2006.1.1.25.
Martinez, H.P., Yannakakis, G.N., Hallam, J., 2014. Don't classify ratings of affect; rank them!. IEEE Trans. Affect. Comput. 5 (3), 314–326. https://doi.org/10.1109/TAFFC.2014.2352268.
Mathur, N., Baldwin, T., Cohn, T., 2020. Tangled up in BLEU: reevaluating the evaluation of automatic machine translation evaluation metrics. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 4984–4997. https://doi.org/10.18653/v1/2020.acl-main.448.
Mcguffie, K., Newhouse, A., 2020. The radicalization risks of GPT-3 and advanced neural language models, Monterey, CA. Technical Report.
McHugh, M.L., 2012. Interrater reliability: the Kappa statistic. Biochem. Med. 22 (3), 276–282.
McIver, J., Carmines, E.G., 1981. Unidimensional Scaling. Sage Publications, Thousand Oaks, CA, USA. https://doi.org/10.1002/9781118445112.stat06462.pub2.24
McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L., 2019. Abandon statistical significance. Am. Stat. 73 (sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253.
Mellish, C., Dale, R., 1998. Evaluation in the context of natural language generation. Comput. Speech Lang. 12 (4), 349–373. https://doi.org/10.1006/csla.1998.0106.
Mertens, D.M., 2010. Research and Evaluation in Education and Psychology: Integrating Diversity with Quantitative, Qualitative, and Mixed Methods. 3 Sage Publications, Thousand Oaks, CA, USA.
Meteyard, L., Davies, R.A., 2020. Best practice guidance for linear mixed-effects models in psychological science. J. Mem. Lang. 112, 104092. https://doi.org/10.1016/j.jml.2020.104092.
Miller, G.A., 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information.. Psychol. Rev. 63 (2), 81. https://doi.org/10.1037/h0043158.
van Miltenburg, E., Elliott, D., Vossen, P., 2017. Cross-linguistic differences and similarities in image descriptions. In: Proceedings of the 10th International Conference on Natural Language Generation. Association for Computational Linguistics, Santiago de Compostela, Spain, pp. 21–30. https://doi.org/10.18653/v1/W17-3503.
van Miltenburg, E., Elliott, D., Vossen, P., 2018. Measuring the diversity of automatic image descriptions. In: Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 1730–1741.
van Miltenburg, E., van de Kerkhof, M., Koolen, R., Goudbeek, M., Krahmer, E., 2019. On task effects in NLG corpus elicitation: a replication study using mixed effects modeling. In: Proceedings of the 12th International Conference on Natural Language Generation. Association for Computational Linguistics, Tokyo, Japan, pp. 403–408. https://doi.org/10.18653/v1/W19-8649.
Morrissey, K., Kirakowski, J., 2013. 'realness' in chatbots: establishing quantifiable criteria. In: Proceedings of the 15th International Conference on Human-Computer Interaction. Springer, pp. 87–96. https://doi.org/10.1007/978-3-642-39330-3_10.
Munafò, M.R., Nosek, B.A., Bishop, D.V., Button, K.S., Chambers, C.D., Du Sert, N.P., Simonsohn, U., Wagenmakers, E.-J., Ware, J.J., Ioannidis, J.P., 2017. A manifesto for reproducible science. Nat. Hum. Behav. 1 (1), 1–9. https://doi.org/10.1038/s41562-016-0021.
Munafò, M.R., Smith, G.D., 2018. Robust research needs many lines of evidence. Nature 553, 399–401. https://doi.org/10.1038/d41586-018-01023-3.
Murphy, K.R., Aguinis, H., 2019. HARKing: how badly can cherry-picking and question trolling produce bias in published results? J. Bus. Psychol. 34 (1), 1–17. https://doi.org/10.1007/s10869-017-9524-7.
Navarro, D., 2019. Learning Statistics with R: A Tutorial for Psychology Students and Other Beginners: Version 0.6.1. University of Adelaide.
Nenkova, A., Chae, J., Louis, A., Pitler, E., 2010. Structural features for predicting the linguistic quality of text. Empirical Methods in Natural Language Generation. Springer, pp. 222–241. https://doi.org/10.1007/978-3-642-15573-4_12.
Nespor, J., 2000. Anonymity and place in qualitative inquiry. Qualit. Inq. 6 (4), 546–569. https://doi.org/10.1177/107780040000600408.
Norman, G., 2010. Likert scales, levels of measurement and the "laws" of statistics. Adv. Health Sci. Educ. 15 (5), 625–632. https://doi.org/10.1007/s10459-010-9222-y.
Normand, M.P., 2016. Less is more: psychologists can learn more by studying fewer people. Front. Psychol. 7, 934. https://doi.org/10.3389/fpsyg.2016.00934.
Novikova, J., Dušek, O., Curry, A.C., Rieser, V., 2017. Why we need new evaluation metrics for NLG. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2241–2252. https://doi.org/10.18653/v1/D17-1237.
Novikova, J., Dušek, O., Rieser, V., 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), pp. 72–78. https://doi.org/10.18653/v1/N18-2012.
Paasche-Orlow, M.K., Taylor, H.A., Brancati, F.L., 2003. Readability standards for informed-consent forms as compared with actual readability. New Engl. J. mMedicine 348 (8), 721–726. https://doi.org/10.1056/NEJMsa021212.
Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318. https://doi.org/10.3115/1073083.1073135.
Perrone, A., Edwards, J., 2019. Chatbots as unwitting actors. In: Proceedings of the First International Conference on Conversational User Interface. ACM, Dublin, Ireland, pp. 1–2. https://doi.org/10.1145/3342775.3342799.
Peter, J., 1677. Artificial Versifying, or the Schoolboy's Recreation. John Sims, London, UK.
Peytchev, A., Couper, M.P., McCabe, S.E., Crawford, S.D., 2006. Web survey design: paging versus scrolling. Int. J. Public Opin. Q. 70 (4), 596–607.

Pitler, E., Nenkova, A., 2008. Revisiting readability: A unified framework for predicting text quality. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 186–195. https://doi.org/10.3115/1613715.1613742.
Popel, M., Žabokrtsk₹ Z., 2009. Improving english-czech tectogrammatical MT. Prague Bull. Math. Linguist. 92, 115–134. https://doi.org/10.2478/v10108-009-0025-3.
Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., Sykes, C., 2009. Automatic generation of textual summaries from neonatal intensive care data. Artif. Intell. 173 (7−8), 789–816. https://doi.org/10.1016/j.artint.2008.12.002.
Post, M., 2018. A call for clarity in reporting BLEU scores. In: Proceedings of the Third Conference on Machine Translation: Research Papers. Association for Computational Linguistics, Belgium, Brussels, pp. 186–191.
Postma, M., Izquierdo, R., Agirre, E., Rigau, G., Vossen, P., 2016. Addressing the MFS bias in WSD systems. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA), Portorož, Slovenia, pp. 1695–1700.
Potter, W.J., Levine-Donnerstein, D., 1999. Rethinking validity and reliability in content analysis. J. Appl. Commun. Res. 27, 258–284. https://doi.org/10.1080/00909889909365539.
Preston, C.C., Colman, A.M., 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Psychol. 104 (1), 1–15. https://doi.org/10.1016/s0001-6918(99)00050-5.
Punch, K., 1998. Introduction to Social Research: Quantitative and Qualitative Approaches. Sage Publications, London, UK.
de Quidt, J., Vesterlund, L., Wilson, A.J., 2019. Experimenter demand effects. In: Schram, A., Ule, A. (Eds.), Handbook of Research Methods and Applications in Experimental Economics. Edward Elgar Publishing, pp. 384–400. https://doi.org/10.4337/9781788110563.
Radziwill, N.M., Benton, M.C., 2017. Evaluating quality of chatbots and intelligent conversational agents. CoRR. arXiv:1303.5076.
Ramos-Soto, A., Bugarin, A.J., Barro, S., Taboada, J., 2015. Linguistic descriptions for automatic generation of textual short-Term weather forecasts on real prediction data. IEEE Trans. Fuzzy Syst. 23 (1), 44–57. https://doi.org/10.1109/TFUZZ.2014.2328011.
Reiter, E., 2011. Task-based evaluation of NLG systems: control vs real-world context. In: Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop. Association for Computational Linguistics, Edinburgh, Scotland, pp. 28–32.
McLeod, S., 2019. Qualitative vs. quantitative research. Retrieved from https://www.simplypsychology.org/qualitative-quantitative.html on March 2, 2020.
Reiter, E., 2017. Types of NLG evaluation: which is right for me?
Reiter, E., 2018. A structured review of the validity of BLEU. Comput. Linguist. 1–12. https://doi.org/10.1162/coli_a_00322.
Reiter, E., Belz, A., 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Comput. Linguist. 35 (4), 529–558. https://doi.org/10.1162/coli.2009.35.4.35405.
Reiter, E., Dale, R., 2000. Building Natural Language Generation Systems. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511519857.
Reiter, E., Sripada, S., Hunter, J., Yu, J., Davy, I., 2005. Choosing words in computer-generated weather forecasts. Artif. Intell. 167 (1−2), 137–169. https://doi.org/10.1016/j.artint.2005.06.006.
Renkema, J., 2012. Schrijfwijzer. 5 SDU Uitgevers, Den Haag, The Netherlands.
Resnik, P., Lin, J., 2010. Evaluation of NLP systems. In: Clark, A., Fox, C., Lappin, S. (Eds.), The Handbook of Computational Linguistics and Natural Language Processing. Wiley-Blackwell, pp. 271–295.
Ribeiro, L.F.R., Schmitt, M., Schütze, H., Gurevych, I., . Investigating pretrained language models for graph-to-text generation. . arXiv:2007.08426.
Rice, T.W., 2008. How to do human-subjects research if you do not have an institutional review board. Respir. Care 53 (10), 1362–1367.
Rodgers, J., 2017. The genealogy of an image, or, what does literature (not) have to do with the history of computing?Tracing the sources and reception of gUlliver's " knowledge engine". Humanities 6 (4), 85. https://doi.org/10.3390/h6040085.
Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K., 2018. Object hallucination in image captioning. In: Proceedings ofthe 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP'18). Association for Computational Linguistics, pp. 4035–4045. https://doi.org/10.18653/v1/d18-1437.
Ruggles, S., Sobek, M., King, M.L., Liebler, C., Fitch, C.A., 2003. IPUMS redesign. Histor. Methods: J. Quant. Interdiscip. Hist. 36 (1), 9–19. https://doi.org/10.1080/01615440309601210.
Sachs, G.A., Hougham, G.W., Sugarman, J., Agre, P., Broome, M.E., Geller, G., Kass, N., Kodish, E., Mintz, J., Roberts, L.W., et al., 2003. Conducting empirical research on informed consent: challenges and questions. IRB: Ethics Hum. Res. 25 (5), 4–10. https://doi.org/10.2307/3564116.
Sambaraju, R., Reiter, E., Logie, R., McKinlay, A., McVittie, C., Gatt, A., Sykes, C., 2011. What is in a text and what does it do: qualitative evaluations of an NLG system −the BT-Nurse− using content analysis and discourse analysis. In: Proceedings of the 13th European Workshop on Natural Language Generation. Association for Computational Linguistics, pp. 22–31.
Scott, D., Moore, J., 2007. An NLG evaluation competition? Eight reasons to be cautious. In: Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, pp. 22–23.
Scriven, M., 1991. Beyond formative and summative evaluation. In: Phillips, M.W.M.D.D. (Ed.), Evaluation and Education: At Quarter Century. University of Chicago Press, Chicago, IL, USA, pp. 19–64.
Sellam, T., Das, D., Parikh, A., 2020. BLEURT: learning robust metrics for text generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 7881–7892.
Shannon, C.E., Weaver, W., 1963. A Mathematical Theory of Communication. University of Illinois Press, Champaign, IL, USA.
Shaughnessy, J., Zechmeister, E., Zechmeister, J., 2006. Research Methods in Psychology. McGraw-Hill.
Shimanaka, H., Kajiwara, T., Komachi, M., 2018. RUSE: regressor using sentence embeddings for automatic machine translation evaluation. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. Association for Computational Linguistics, Belgium, Brussels, pp. 751–758. https://doi.org/10.18653/v1/W18-6456.
Reiter, E., 2020. Why do we still use 18-year old BLEU? Retrieved from https://ehudreiter.com/2020/03/02/why-use-18-year-old-bleu/ on March 7, 2020.
Shuttleworth, M., 2009. Counterbalanced measures design. Retrieved from https://explorable.com/counterbalanced-measures-design/ on July 27, 2020.
Silberman, M.S., Tomlinson, B., LaPlante, R., Ross, J., Irani, L., Zaldivar, A., 2018. Responsible research with crowds: pay crowdworkers at least minimum wage. Commun. ACM 61 (3), 39–41.
Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22 (11), 1359–1366. https://doi.org/10.1177/0956797611417632.
Smiley, C., Plachouras, V., Schilder, F., Bretz, H., Leidner, J., Song, D., 2016. When to plummet and when to soar: corpus based verb selection for Natural Language Generation. In: Proceedings of the 9th International Natural Language Generation conference. Association for Computational Linguistics, Edinburgh, UK, pp. 36–39. https://doi.org/10.18653/v1/W16-6606.
Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J., 2006. A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas. Association for Machine Translation in the Americas, Cambridge, MA, USA, pp. 223–231.
Snow, R., O'Connor, B., Jurafsky, D., Ng, A., 2008. Cheap and fast − but is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Honolulu, Hawaii, USA, pp. 254–263. https://doi.org/10.5555/1613715.1613751.
Søgaard, A., 2013. Estimating effect size across datasets. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Atlanta, Georgia, pp. 607–611.
Søgaard, A., 2017. Evaluation in natural language processing (and tennis rackets in a world with no gravity). Retrieved from https://medium.com/@soegaarducph/yoavs-recent-blog-post-sparked-a-lot-of-interest-across-different-communities-and-many-have-5b6a6c794887/on July 27, 2020.
Sparck Jones, K., Galliers, J.R., 1996. Evaluating Natural Language Processing Systems: An Analysis and Review. Springer, Berlin and Heidelberg.
Sripada, S., Reiter, E., Hawizy, L., 2005. Evaluation of an NLG system using post-edit data: lessons learnt. In: Proceedings of the 10th European Workshop on Natural Language Generation. Association for Computational Linguistics, Aberdeen, Scotland, pp. 133–139.

Stake, R.E., 2000. Case studies. In: Lincoln, N.K.D..Y.S. (Ed.), Handbook of Qualitative Research. Sage, Thousand Oaks, CA, USA, pp. 435–454.
Sterling, S., 2018. Investigating the complexity of consent forms in ESL research. J. Res. Desi. Stat. Linguist. Commun. Sci. 4 (2), 156–175. https://doi.org/10.1558/jrds.35702.
Stieger, S., Reips, U.-D., Voracek, M., 2007. Forced-response in online surveys: Bias from reactance and an increase in sex-specific dropout. J. Am. Soc. Inf. Sci. Technol. 58 (11), 1653–1660.
Stunkel, L., Benson, M., McLellan, L., Sinaii, N., Bedarida, G., Emanuel, E., Grady, C., 2010. Comprehension and informed consent: assessing the effect of a short consent form. IRB: Ethics Hum. Res. 32 (4), 1–9.
Sulem, E., Abend, O., Rappoport, A., BLEU is not suitable for the evaluation of text simplification. http://arxiv.org/abs/1810.05995arXiv preprint arXiv:1810.05995 Accepted for publication as a short paper at EMNLP 2018. 10.18653/v1/D18-1081
Swift, J., 1774. Travels Into Several Remote Nations of the World: In Four Parts. By Lemuel Gulliver. First a Surgeon, and Then a Captain of Several Ships..., 1. Benjamin Motte, London, UK.
Tourangeau, R., Couper, M.P., Galesic, M., Givens, J., 2004. A comparison of two web-based surveys: static versus dynamic versions of the NAMCS questionnaire. In: Proceedings of the RC33 6th International Conference on Social Science Methodology: Recent Developments and Applications in Social Research Methodology, Amsterdam, The Netherlands, pp. 1–8.
Tukey, J.W., 1980. We need both exploratory and confirmatory. Am. Stat. 34 (1), 23–25. https://doi.org/10.1080/00031305.1980.10482706.
Turian, J.P., Shen, L., Melamed, I.D., 2003. Evaluation of machine translation and its evaluation. In: Proceedings of MT Summit IX, pp. 1–9.
Van Deemter, K., 2016. Computational Models of Referring: A Study in Cognitive Science. The MIT Press. https://doi.org/10.7551/mitpress/9082.001.0001.
Van Deemter, K., Sun, L., Sybesma, R., Li, X., Chen, B., Yang, M., 2017. Investigating the content and form of referring expressions in Mandarin: introducing the mtuna corpus. In: Proceedings of the 10th International Conference on Natural Language Generation. Association for Computational Linguistics, Santiago de Compostela, Spain, pp. 213–217. https://doi.org/10.18653/v1/W17-3532.
Van Enschot, R., Spooren, W., van den Bosch, A., Burgers, C., Degand, L., Evers-Vermeul, J., Kunneman, F., Liebrecht, C., Linders, Y., Maes, A., 2017. Taming our wild data: on intercoder reliability in discourse research. Unpublished Manuscript Submitted for publication.
Van Heuven, W.J., Mandera, P., Keuleers, E., Brysbaert, M., 2014. Subtlex-UK: a new and improved word frequency database for british english. Q. J. Exp. Psychol. 67 (6), 1176–1190.
Vannette, D.L., 2016. Testing the effects of different types of attention interventions on data quality in web surveys. experimental evidence from a 14 country study. In: Proceedings of the 71st Annual Conference of the American Association for Public Opinion Research, pp. 1–4.
Vedantam, R., Lawrence Zitnick, C., Parikh, D., 2015. Cider: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4566–4575.
Vicente, P., Reis, E., 2010. Using questionnaire design to fight nonresponse bias in web surveys. Soc. Sci. Comput. Rev. 28 (2), 251–267. https://doi.org/10.1177/0894439309340751.
Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., Montanier, J.-M., Oranc, C., Oudgenoeg-Paz, O., Garcia, D.H., Papadopoulos, F., Schodde, T., Verhagen, J., Wallbridge, C., Willemsen, B., de Wit, J., Belpaeme, T., Göksun, T., Kopp, S., Krahmer, E., Küntay, A., Leseman, P., Kumar Pandey, A., 2019. Second language tutoring using social robots: a large-scale study. In: Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, pp. 497–505. https://doi.org/10.1109/HRI.2019.8673077.
Wagenmakers, E.-J., 2007. A practical solution to the pervasive problems of p values. Psychon. Bull. Rev. 14 (5), 779–804. https://doi.org/10.3758/BF03194105.
Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H., 2011. Why psychologists must change the way they analyze their data: the case of PSI: comment on bEM (2011).. J. Pers. Soc. Psychol. 100 (3), 426–432. https://doi.org/10.1037/a0022790.
Wang, L., Qin, S., Xu, M., Zhang, R., Qi, L., Zhang, W., 2019. From quick-draw to story: a story generation system for kids' robot. In: Proceedings of the 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO). IEEE, Dali, China, pp. 1941–1946. https://doi.org/10.1109/ROBIO49542.2019.8961449.
Welty, C., Paritosh, P., Aroyo, L., . Metrology for AI: from benchmarks to instruments. . arXiv preprint arXiv:1911.01875.
Whitley, B.E., Kite, M.E., 2013. Principles of Research in Behavioral Science. 3 Routledge, New York, NY, USA.
Wicherts, J.M., Veldkamp, C.L., Augusteijn, H.E., Bakker, M., Van Aert, R., Van Assen, M.A., 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. Front. Psychol. 7, 1832. https://doi.org/10.3389/fpsyg.2016.01832.
Williams, T., Scheutz, M., 2017. Referring expression generation under uncertainty: algorithm and evaluation framework. In: Proceedings of the 10th International Conference on Natural Language Generation. Association for Computational Linguistics, Santiago de Compostela, Spain, pp. 75–84. https://doi.org/10.18653/v1/W17-3511.
Willis, G.B., 2004. Cognitive Interviewing: A Tool for Improving Questionnaire Design. Sage Publications, Thousand Oaks, CA, USA.
Wolford, B.,. Everything you need to know about the GPDR Data Protection Officer (DPO). GDPR.eu. https://gdpr.eu/data-protection-officer/, last accessed June 5, 2020.
Yang, Y., Zhang, Z., Miklau, G., Winslett, M., Xiao, X., 2012. Differential privacy in data publication and analysis. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, Scottsdale, Arizona, USA, pp. 601–606. https://doi.org/10.1145/2213836.2213910.
Yannakakis, G.N., Hallam, J., 2011. Ranking vs. preference: a comparative study of self-reporting. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction. Springer, pp. 437–446. https://doi.org/10.1007/978-3-642-24600-5_65.
Yannakakis, G.N., Martínez, H.P., 2015. Ratings are overrated!. Front. ICT 2, 13. https://doi.org/10.3389/fict.2015.00013.
Yu, S., Alper, H.E., Nguyen, A.-M., Brackbill, R.M., Turner, L., Walker, D.J., Maslow, C.B., Zweig, K.C., 2017. The effectiveness of a monetary incentive offer on survey response rates and response completeness in a longitudinal study. BMC Med. Res. Methodol. 17 (1), 77–86. https://doi.org/10.1186/s12874-017-0353-1.
Zarrieß, S., Loth, S., Schlangen, D., 2015. Reading times predict the quality of generated text above and beyond human ratings. In: Proceedings of the 15th European Workshop on Natural Language Generation. Association for Computational Linguistics, Brighton, UK, pp. 38–47. https://doi.org/10.18653/v1/W15-4705.
Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2020. BERTScore: evaluating text generation with BERT. In: Proceedings of the Eighth International Conference on Learning Representations. OpenReview.net, Ethiopia, Addis Ababa, pp. 1–43.
Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C.M., Eger, S., 2019. MoverScore: text generation evaluating with contextualized embeddings and earth mover distance. 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). Association for Computational Linguistics, Hong Kong, China, pp. 563–578. https://doi.org/10.18653/v1/D19-1053.
Zipf, G.K., 1949. Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology.. Addison-Wesley, Cambridge, MA.

Chris van der Lee is a Ph.D. student at the Tilburg center for Cognition and Communication. His research mainly focuses on data-to-text generation, dialogue systems, automated journalism, and text classification.

Albert Gatt is a Senior Lecturer and Director of the Institute of Linguistic and Language Technology, University of Malta. His research interests are in data-to-text generation, the Vision-Language interface, and NLP evaluation.

Emiel van Miltenburg is an assistant professor at the Tilburg center for Cognition and Communication, Tilburg University. His research interests include (multimodal) NLG, evaluation, accessibility, and ethics in NLP.

Emiel Krahmer is a full professor at the Tilburg center for Cognition and Communication, Tilburg University. He studies how people communicate with each other and how computers can be taught to do the same. Specific research interests include natural language generation, human-robot interaction, health communication and evaluation.