Best practices for the human evaluation of automatically generated text

Chris van der Lee
Tilburg University
c.vdrlee@uvt.nl

Albert Gatt
University of Malta
albert.gatt@um.edu.mt

Emiel van Miltenburg
Tilburg University
c.w.j.vanmiltenburg@uvt.nl

Sander Wubben
Tilburg University
s.wubben@uvt.nl

Emiel Krahmer
Tilburg University
e.j.krahmer@uvt.nl

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 human evaluation is currently conducted, and presents a set of best practices, grounded in the literature. 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) (see also Swift, 1774; Rodgers, 2017), human evaluation is still an understudied aspect. Such an evaluation is crucial 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, but it can also help us 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). The importance of evaluation for NLG is itself uncontentious; what is perhaps more contentious is the way in which evaluation should be conducted. This paper provides an overview of current practices in human evaluation, showing that there is no consensus as to how NLG systems should be evaluated. As a result, it is hard to compare the results published by different groups, 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 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).

Previous studies have also provided overviews of evaluation methods. Gkatzia and Mahamood (2015) focused on NLG papers from 2005-2014; 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. 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.

Following Gkatzia and Mahamood (2015), Section 3 provides an overview of current evaluation practices, based on papers from INLG and ACL in 2018. Apart from the broad range of methods used, we also observe that evaluation practices have changed since 2015: for example, there is a significant decrease in the number of papers featuring extrinsic evaluation. This may be caused by the current focus on smaller, decontextualized tasks, which do not take users into account.

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. We hope that our recommendations can serve as a guide for newcomers in the field, and can otherwise help NLG research by standardizing the way human evaluation is carried out.

2 Automatic versus human evaluation

Automatic metrics such as BLEU, METEOR, and ROUGE are increasingly popular; Gkatzia and Mahamood's (2015) survey of NLG papers from 2005-2014 found that 38.2% used automatic met-

Proceedings of The 12th International Conference on Natural Language Generation, pages 355–368,
Tokyo, Japan, 28 Oct - 1 Nov, 2019. ©2019 Association for Computational Linguistics

rics, while our own survey (described more fully in Section 3) shows that 80% of the empirical papers presented at the ACL track on NLG or at the INLG conference in 2018 reported on 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 uninterpretable. 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 therefore requires an error analysis. Automatic metric scores can also be hard to interpret because it is unclear how stable the reported scores are. With BLEU, for instance, libraries often have their own BLEU score implementation, which may differ from one another, thus affecting the scores (this is recently addressed by Post, 2018). Reporting the scores accompanied by confidence intervals, calculated using bootstrap resampling (Koehn, 2004), may increase the stability and therefore interpretability of the results. However, such statistical tests are not straightforward to perform.

Automatic metrics do not correlate with human evaluations. This has been repeatedly observed (e.g. Belz and Reiter, 2006; Reiter and Belz, 2009; Novikova et al., 2017).[1] 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. The alternative is to perform a human evaluation.

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). We would not 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 qualitative aspects of NLG output, such as fluency or stylistic compatibility with reference texts. But for a general assessment of overall system quality, human evaluation remains the gold standard.

3 Overview of current work

This section provides an overview of current human evaluation practices, based on the papers published at INLG (N=51) and ACL (N=38) in 2018. 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. [2]

3.1 Intrinsic and extrinsic evaluation

Human evaluation of natural language generation systems can be done using intrinsic and extrinsic methods (Sparck Jones and Galliers, 1996; Belz and Reiter, 2006). Intrinsic approaches aim to evaluate properties of the system's output, for instance, by asking participants about the fluency of the system's output in a questionnaire. Extrinsic 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. While extrinsic evaluation has been argued to be more useful (Reiter and Belz, 2009), it is also rare. Only three papers (3%) in the sample of INLG and ACL papers presented an extrinsic evaluation. This is a notable decrease from Gkatzia and Mahamood (2015), who found that nearly 25% of studies contained an extrinsic evaluation. Of course, extrinsic evaluation is the most time- and cost-intensive out of all possible evaluations (Gatt and Krahmer, 2018), which might explain the rarity, but does not explain the decline in (relative) frequency. That might be because of the set-up of the tasks we see nowadays. Extrinsic evaluations require that the system is embedded in its target use context (or a suitable simulation thereof), which in turn requires that the system addresses a specific purpose. In practice, this often means the system follows the 'traditional' NLG

[1] 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 of reference-based evaluation due to low-coverage reference sets cannot be overcome by attainably increasing references.

[2] For the ACL papers, we focused on the following tracks: Machine Translation, Summarization, Question Answering, and Generation. See Supplementary Materials for a detailed overview of the investigated papers and their evaluation characteristics

Criterion	Total	Criterion	Total
Fluency	13	Manipulation check	3
Naturalness	8	Informativeness	3
Quality	5	Correctness	3
Meaning preservation	5	Syntactic correctness	2
Relevance	5	Qualitative analysis	2
Grammaticality	5	Appropriateness	2
Overall quality	4	Non-redundancy	2
Readability	4	Semantic adequacy	2
Clarity	3	Other criteria	25

Table 1: Criteria used for human evaluation from all papers. Separate counts for ACL and INLG 2018 are in the appendix.

Scale	Count
Likert (5-point)	14
Preference	10
Likert (2-point)	6
Likert (3-point)	5
Other Likert (4,7,10-point)	5
Rank-based Magnitude Estimation	5
Free text comments	1

Table 2: Types of scales used for human evaluation

pipeline (Reiter and Dale, 2000), encompassing many of these pipeline sub-tasks to go from input data to complete output texts (Mellish et al., 2006; Gatt and Krahmer, 2018). Such systems were a mainstay of NLG literature until recently (e.g., Harris, 2008; Gatt and Portet, 2010; Reiter et al., 2003), but the field has shifted towards focusing on only one or a few of the sub-tasks from the NLG pipeline (e.g. text planning, surface realization, referring expression generation), with a concomitant focus on text output quality, for which an intrinsic evaluation may be sufficient. However, we are starting to see a swing back towards a full pipeline approach with separate neural modules handling sub-tasks (Castro Ferreira et al., 2019), which may also cause a resurgence of extrinsic evaluation.

3.2 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 4.1 for further discussion). This variety, suggesting a lack of agreement, is clear from Table 1. Except for fluency, and for naturalness and quality which were used for a shared task, most criteria are infrequent; the numerous 'other criteria' are those which are used only once. At the same time, there is probably significant overlap. For instance, naturalness is sometimes linked to fluency, and informativeness to adequacy (Novikova et al., 2018). In short, there is no standard evaluation model for NLG. Furthermore, there is significant variety in naming conventions.

3.3 Sample size and demographics

When looking at sample size, it is possible to distinguish between expert-focused and reader-focused evaluation. 14 papers (28%) used an expert-focused approach, meaning that between 1 and 4 expert annotators evaluated system output. 13 papers (26%) employed a larger-scale reader-focused method in which 10 to 60 readers judged the generated output. We found a median of 4 annotators. However, these numbers might not reflect reality: only 55% of papers specified the number of participants and an even smaller number (18%) reported the demographics of their sample. Only 12.5% of the papers with a human evaluation reported inter-annotator agreement, using 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.

3.4 Design

Apart from participant sample size, an important issue that impacts statistical power is the number of items (e.g. generated sentences) used in an evaluation. Among papers that reported these numbers, we observed a median of 100 items used for human evaluation in INLG and ACL papers. The number of items however ranged between 2 and 5,400, illustrating a sizable discrepancy. In 83% of papers that reported these figures, all annotators saw all examples. Only 12.5% of papers 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.

3.5 Number of questions and types of scales

In addition to the diversity in criteria used to measure text quality (see Section 3.2), there is a wide range of rating methods that are used to measure those criteria. Do note that Likert and rating scales are treated indistinctly here (for a distinction, see Amidei et al., 2019). The 5-point Likert scale is the most popular option, but preference ratings are a close second (see Table 2). Other types of rating methods are much less common. Rank-based Magnitude Estimation, a continuous metric, was only found among shared task papers, and only one paper reported using free-text comments.

We also investigated the number of ratings used to measure a single criterion (e.g. a paper may use two ratings to measure two different aspects of fluency). Only 34% of papers with a human evaluation reported the number of ratings to measure a criterion. These numbers ranged from 1 to 4 ratings for a criterion, with 1 rating being the most common.

3.6 Statistics and data analysis

A minority (33%) 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 test a specific hypothesis (Navarro, 2019). However, not all papers using a statistical test report their hypotheses. And conversely, some papers reporting hypotheses do not perform a statistical test. 19% of all papers explicitly state their hypotheses or research questions.

4 Best practices

This section provides best practices for carrying out and reporting human evaluation in NLG. We (mostly) restrict ourselves to intrinsic evaluation.

4.1 Text quality and criteria

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.

If successfully conveying communicative intention is taken to be the main 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). 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.

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 success beyond a 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 result in more convoluted text which is harder to understand).

Ultimately, the criteria to be considered will depend on the task. For example, in style transfer, manipulation checks are important to determine whether the style has been transferred correctly, while also ensuring meaning preservation. These criteria are not necessarily important for a system that generates weather reports from numerical data, where accuracy, fluency, coherence and genre compatibility might be more prominent concerns. By contrast, 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, 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 more objective measures, similar to existing readability measures (e.g., Ambati et al., 2016; Kincaid et al., 1975; Pitler and Nenkova, 2008; Vajjala and Meurers, 2014), where objective text metrics (e.g. average word length, average parse tree height, average number of nouns) are used in a formula, or as features in a regression model, to obtain a score for a text criterion. Similarly, it may be possible to use separate subjective criteria as features in a regression model to calculate overall text quality scores. This would also provide information about the importance of the subjective criteria for overall text quality judgments. However, such research on the relationship between subjective criteria and objective measures is currently lacking for NLG.

One obstacle to addressing the difficulties identified in this section is the lack of 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 different criteria may overlap or be inter-definable. As Gatt and Belz (2010) and Hastie and Belz (2014) suggest, common and shared evaluation guidelines should be developed for each task, and efforts should be made to standardise criteria and naming conventions. In the absence of such guidelines, care should be taken to explicitly define the criteria measured and highlight possible overlaps between them.

4.2 Sample size, demographics and agreement

Expert- versus reader-focused Section 3.3 made a distinction between expert-focused and reader-focused evaluation. With an expert-focused design, a small number of expert annotators is recruited to judge aspects of the NLG system. A reader-focused 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 a more qualitative analysis is used, whereas most expert-focused evaluations in our sample of papers used closed-ended questions with Likert scales.

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 non-experts. This is corroborated by Lentz and De Jong (1997), who found that expert judgments only predict the outcomes of reader-focused evaluation to a limited extent. Experts are also susceptible to considerable variance, so that automatic metrics are sometimes more reliable (Belz and Reiter, 2006). Thus, the conclusion of Belz and Reiter (2006) in favour of large-scale reader-focused studies, rather than expert-focused ones, seems well-taken.

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).

Evaluator agreement The varying opinions of judges are also reflected in low Inter-Annotator Agreement (IAA), where adequate thresholds also 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 a more precise IAA score) 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 (§ 3.4).

When the goal of an evaluation is to identify potential problems with output texts, a low IAA, indicating variety 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 statistics, it is advisable to also report percentage agreement, which is easily interpretable (McHugh, 2012).

Sample size For expert-focused evaluations, good advice is provided by Van Enschot et al. (2017): 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. In the case of large-scale studies, Brysbaert (2019) recently stated that most studies with less than 50 participants are underpowered and that for most designs and analyses 100 or more participants are needed. With the introduction of crowdsourcing such numbers are obtainable, at least for widely-spoken languages (though see van Miltenburg et al. 2017 for a counterexample). Furthermore, the number of participants necessary can be decreased by having multiple observations per condition per participant (i.e., having participants perform more judgments).

Whatever the sample size, a minimum good practice guideline is to always report participant numbers, with relevant demographic data (i.e., gender, nationality, age, fluency in the target language, academic background, etc), in order to enhance replicability and enable readers to gauge the meaningfulness of the results.

4.3 Number of questions and types of scales

As shown in Section 3.5, Likert scales are the prevalent rating method for NLG evaluation, 5-point scales being the most popular, followed by 2-point, and 3-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. Most of the experimental literature's findings found that 7-point scales maximise reliability, validity and discriminative power (for instance, 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.

While Likert scales are the most popular scale within the NLG domain (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). Similarly, for the development of an automatic metric for NLG, Chaganty et al. (2018) found that annotator variance decreased significantly when using post-edits as a metric instead of a Likert scale survey. Finally, Novikova et al. (2018) compared Likert scales for NLG system evaluation to two continuous scales: a vanilla magnitude estimation measure and a rank-based magnitude estimation measure. The researchers found that both magnitude estimation scales delivered more reliable and consistent text evaluation scores.

All these studies seem to suggest that ranking-based methods (combined with continuous scales) are the preferred method. However, there are two critical remarks to be made on this. 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) illustrated that the TrueSkill$^{\text{TM}}$ algorithm (Herbrich et al., 2007) can be implemented. This algorithm uses binary comparisons to reliably rank systems, which greatly reduces the amount of data needed for multiple-system comparisons.

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 5 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 (e.g., 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 make the reliability of Likert scales on a par with that of ranking tasks; this, however, has not been empirically tested. Also, do 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., 2019).

In sum, we advise to use either multiple-item 7-point Likert scales, or a (continuous) ranking task. The latter should be used in combination with TrueSkill$^{\text{TM}}$ when multiple systems are compared. As Aroyo and Welty (2014) note, disagreement in the responses can be due to three factors: the item, the worker, and the task. Therefore, it is necessary to pilot the rating task before deploying it more widely, and to analyze disagreement on the annotator level, to see whether individual annotators are causing discrepancies in the ratings for different items.

Alternative evaluation instruments should not be ruled out either. Ever since a pilot in 2016 (Bojar et al., 2016a), recent editions of the Conference on Machine Translation (WMT), have used Direct Assessment, whereby participants compare an output to a reference text on a continuous (0-100) scale (Graham et al., 2017; Bojar et al., 2016b), similar to Magnitude Estimation (Bard et al., 1996). 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. It should also be noted that most metrics used in NLG are reader-focused. However, in many real-world scenarios, especially 'creative' NLG applications, NLG systems and human writers work alongside each other in some way (see Maher, 2012; Manjavacas et al., 2017). With such a collaboration in mind, it makes sense to also investigate writer-focused methods. 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, authors have explored the use of such metrics in NLG (Bernhard et al., 2012; Han et al., 2017; Sripada et al., 2005).

Finally, some remarks on qualitative evaluation methods are in order. Reiter and Belz (2009) note that free-text comments can be beneficial to diagnose potential problems of an NLG system. Furthermore, Sambaraju et al. (2011) argue the added value of content analysis and discourse analysis for evaluation. Such qualitative analyses can find potential blind spots of quantitative analyses. At the same time, the subjectivity that is often inherent in studies based on discourse analysis, such as Sambaraju et al. (2011) would need to be offset by data from larger-scale, quantitative studies.

4.4 Design

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 due to fatigue, practice, carryover effects or other (external) factors. If the order in which the output of systems are presented 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). 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). 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 that 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 the 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, however, can be partially mitigated by grouping items randomly into sets, and counterbalancing the order of sets rather than individual items). In most cases, 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, such a design is preferable (Shaughnessy et al., 2006).

Novikova et al. (2018) found that the presentation of questions matters. When evaluating text criteria, answers to questions about different criteria tend to correlate when they are presented simultaneously for a given item. When participants are shown an item multiple times and questioned about each text criterion separately, this correlation is reduced.

4.5 Statistics and data analysis

Within behavioral sciences, it is standard to evaluate hypotheses based on whether findings are statistically significant or not (typically, in published papers, they are), although a majority of NLG papers do not report statistical tests (see Section 3.6). However, 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). Moreover, one may wonder whether standard null-hypothesis significance testing (NHST) is applicable or helpful in human NLG evaluation.

In a common scenario, NLG researchers may want to compare various versions of their own novel system (e.g. with or without output variation, or relying on different word embedding models, to give just two more or less random examples) 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—are seldom (if ever) discussed in the NLG literature.

A related issue is that clear hypotheses are often not stated (see Section 3.6). Of course, researchers generally assume that their system will be rated higher than the comparison systems. But they will not necessarily assume that they will perform better on all dependent variables. Moreover, they may have no specific hypotheses about which variant of their own system will perform best.

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 a H0) increases (such errors are known as false positives or Type I errors). Various remedies for this particular problem exist, one being an application of the simple Bonferroni correction, which amounts to lowering the significance threshold α—commonly .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.

Finally, 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),

Topic	Best practice
General	Always conduct a human evaluation (if possible).
Criteria	Use separate criteria rather than an overall quality assessment. Properly define the criteria that are used in the evaluation.
Sampling	Preferably use a (large-scale) reader-focused design rather than a (small-scale) expert-focused design. Always recruit sufficiently many participants. Report (and motivate) the sample size and the demographics.
Annotation	For a qualitative analysis, recruit multiple annotators (at least 2, more is better) Report the Inter-Annotator Agreement score with confidence intervals, plus a percentage agreement.
Measurement	For a quantitative study, use multiple item 7-point (preferably) Likert scales, or (continuous) ranking.
Design	Reduce order- and learning effects by counterbalancing/random ordering, and properly report this.
Statistics	If the evaluation study is exploratory, only report exploratory data analysis. If the study is confirmatory, consider preregistering and conduct appropriate statistical analyses.

Table 3: List of best practices for human evaluation of automatically generated text.

if only because different systems may rely on the same training data. In view of these issues, some have argued that NHST should be abandoned (Koplenig, 2017; McShane et al., 2019).

In our opinion, the distinction between exploratory and confirmative (hypothesis) testing should be taken more seriously within NLG. Much human evaluation of NLG could better be approached from an exploratory perspective, and instead of full-fledged hypothesis testing it would be more appropriate to analyse findings with exploratory data analysis techniques (Tukey, 1980; Cumming, 2013). When researchers do have clear hypotheses, statistical significance testing can be a powerful tool (assuming it is applied correctly). In these cases, we recommend preregistering the hypotheses and analysis plans before conducting the actual evaluation.[3]

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?), etc. By being explicit beforehand (i.e., by preregistering), any flexibility in the analysis (be it intentional or not) is removed. Preregistration is increasingly common in medical and psychological science, and even though it is not perfect (Claesen et al., 2019) at least it has made research more transparent and controllable, which has a positive impact on the possibilities to replicate earlier findings.

Finally, alternative statistical models deserve more attention within NLG. For example, within psycholinguistics it is common to look both at participant and item effects (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 item aspects very well (e.g., Barr et al., 2013), and deserve more attention in NLG. Finally, Bayesian models are worth exploring, because they are less sensitive to the aforementioned problems with NHST (e.g., Gelman et al., 2006; Wagenmakers, 2007).

5 Conclusion

We have provided an overview of the current state of human evaluation in NLG, and presented a set of best practices, summarized in Table 3. This is a broad topic, and for reasons of space we were not able to cover all aspects of human evaluation in detail. Nevertheless, we hope that this overview will serve as a useful reference for NLG practitioners, and in future work we aim to provide a more extensive set of best practices for carrying out human evaluations in Natural Language Generation.

[3] For example at `osf.io` or `aspredicted.org`

Acknowledgements

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. We would also like to thank the anonymous reviewers for their valuable and insightful comments.

References

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental CCG parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1051–1057, San Diego, California, USA. Association for Computational Linguistics.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018a. Evaluation methodologies in Automatic Question Generation 2013-2018. *INLG 2018*, page 307.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018b. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 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*, Tokyo, Japan. Association for Computational Linguistics.

R Ananthakrishnan, Pushpak Bhattacharyya, M Sasikumar, and Ritesh M Shah. 2007. Some issues in automatic evaluation of English-Hindi MT: More blues for BLEU. *ICON*.

Eva Anduiza and Carol Galais. 2016. Answering without reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, 29(3):497–519.

Lora Aroyo and Chris Welty. 2014. The three sides of crowdtruth. *Journal of Human Computation*, 1:31–34.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Ellen Gurman Bard, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, pages 32–68.

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320. Association for Computational Linguistics.

Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. 2018. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6.

Derrick A Bennett. 2001. How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25(5):464–469.

Adam J Berinsky, Michele F Margolis, and Michael W Sances. 2016. Can we turn shirkers into workers? *Journal of Experimental Social Psychology*, 66:20–28.

Delphine Bernhard, Louis De Viron, Véronique Moriceau, and Xavier Tannier. 2012. Question generation for french: Collating parsers and paraphrasing questions. *Dialogue & Discourse*, 3(2):43–74.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016a. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016b. Ten years of WMT evaluation campaigns: Lessons learnt. In *Proceedings of the LREC 2016 Workshop Translation Evaluation From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 27–34.

Marc Brysbaert. 2019. How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, 2(1):1–38.

Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 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*, Hong Kong, SAR. Association for Computational Linguistics.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evaluation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653.

Leshem Choshen and Omeri Abend. 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*, pages 632–642, Melbourne, Australia. Association for Computational Linguistics.

Domenic V Cicchetti, Donald Shoinralter, and Peter J Tyrer. 1985. The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement*, 9(1):31–36.

Aline Claesen, Sara Lucia Brazuna Tavares Gomes, Francis Tuerlinckx, et al. 2019. Preregistration: Comparing dream to reality. *PsyArXiv*.

Herbert H Clark. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior*, 12(4):335–359.

Scott Clifford and Jennifer Jerit. 2015. Do attempts to improve respondent attention increase social desirability bias? *Public Opinion Quarterly*, 79(3):790–802.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. Routledge.

Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. 1971. Artificial paranoia. *Artificial Intelligence*, 2(1):1–25.

Geoff Cumming. 2013. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 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*, pages 73–83. Association for Computational Linguistics.

Adamantios Diamantopoulos, Marko Sarstedt, Christoph Fuchs, Petra Wilczynski, and Sebastian Kaiser. 2012. Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science*, 40(3):434–449.

George Doddington. 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*, pages 138–145. Morgan Kaufmann Publishers Inc.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhikers 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)*, pages 1383–1392.

Johan Falkenjack, Katarina Heimann Mühlenbock, and Arne Jönsson. 2013. Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 27–40.

Albert Gatt and Anja Belz. 2010. Introducing shared tasks to NLG: The TUNA shared task evaluation challenges. In *Empirical Methods in Natural Language Generation*, pages 264–293. Springer.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Albert Gatt and François Portet. 2010. Textual properties and task based evaluation: Investigating the role of surface properties, structure and content. In *Proceedings of the 6th International Natural Language Generation Conference*, pages 57–65. Association for Computational Linguistics.

Andrew Gelman and Eric Loken. 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*.

Andrew Gelman et al. 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3):515–534.

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

Yvette Graham, Timothy Baldwin, Alistair Moffast, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):330.

Paul E Green and Vithala R Rao. 1970. Rating scales and information recovery: How many scales and response categories to use? *Journal of Marketing*, 34(3):33–39.

Bo Han, Will Radford, Anaïs Cadilhac, Art Harol, Andrew Chisholm, and Ben Hachey. 2017. Post-edit analysis of collective biography generation. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 791–792, Perth, Australia. International World Wide Web Conferences Steering Committee.

Mary Dee Harris. 2008. Building a large-scale commercial NLG system for an EMR. In *Proceedings of the Fifth International Natural Language Generation Conference (INLG '08)*, pages 157–160, Morristown, NJ, USA. Association for Computational Linguistics.

Helen Hastie and Anja Belz. 2014. A comparative evaluation methodology for NLG in interactive systems. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *The Behavioral and Brain Sciences*, 23:61–83; discussion 83–135.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueSkill$^{\mathrm{TM}}$: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems*, pages 569–576.

Richard R Jones. 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*, volume 3, pages 247–248. American Psychological Association Washington, DC.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, FOG count and Flesch reading ease formula) for navy enlisted personnel. *Research Branch Report*, 8(75).

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics, Association for Computational Linguistics.

Alexander Koplenig. 2017. Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory*.

Daniel Lakens, Federico G Adolfi, Casper J Albers, Farid Anvari, Matthew AJ Apps, Shlomo E Argamon, Thom Baguley, Raymond B Becker, Stephen D Benning, Daniel E Bradford, et al. 2018. Justify your alpha. *Nature Human Behaviour*, 2(3):168.

Chris van der Lee, Bart Verduijn, Emiel Krahmer, and Sander Wubben. 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*, pages 962–972.

Leo Lentz and Menno De Jong. 1997. The evaluation of text quality: Expert-focused and reader-focused methods compared. *IEEE transactions on professional communication*, 40(3):224–234.

Robert W Lissitz and Samuel B Green. 1975. Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology*, 60(1):10.

Mary Lou Maher. 2012. Computational and collective creativity: Who's being creative? In *Proceedings of the 3rd International Conference on Computer Creativity*, pages 67–71, Dublin, Ireland. Association for Computational Linguistics.

Enrique Manjavacas, Folgert Karsdorp, Ben Burtenshaw, and Mike Kestemont. 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)*, pages 29–37, Santiago de Compostela, Spain. Association for Computational Linguistics.

Hector P Martinez, Georgios N Yannakakis, and John Hallam. 2014. Don't classify ratings of affect; rank them! *IEEE transactions on affective computing*, 5(3):314–326.

Mary L McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282.

John McIver and Edward G Carmines. 1981. *Unidimensional scaling*. 24. Sage.

Blakeley B McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L Tackett. 2019. Abandon statistical significance. *The American Statistician*, 73(sup1):235–245.

Chris Mellish and Robert Dale. 1998. Evaluation in the context of natural language generation. *Computer Speech & Language*, 12(4):349–373.

Chris Mellish, Donia Scott, Lynne Cahill, Daniel Paiva, Roger Evans, and Mike Reape. 2006. A reference architecture for natural language generation systems. *Natural Language Engineering*, 12(01):1–34.

George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 21–30, Santiago de Compostela, Spain. Association for Computational Linguistics.

Daniel Navarro. 2019. *Learning statistics with R: A tutorial for psychology students and other beginners: Version 0.6.1*. University of Adelaide.

Ani Nenkova, Jieun Chae, Annie Louis, and Emily Pitler. 2010. Structural features for predicting the linguistic quality of text. In *Empirical Methods in Natural Language Generation*, pages 222–241. Springer.

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

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for Natural Language Generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78.

John Peter. 1677. *Artificial Versifying, or the Schoolboys Recreation*. John Sims, London, UK.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*, pages 186–195. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

W James Potter and Deborah Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, 27:258–284.

Carolyn C Preston and Andrew M Colman. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104(1):1–15.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, pages 1–12.

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

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, UK.

Ehud Reiter, Roma Robertson, and Liesl M Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1-2):41–58.

Jan Renkema. 2012. *Schrijfwijzer*, 5 edition. SDU Uitgevers, Den Haag, The Netherlands.

Johannah Rodgers. 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 Gullivers 'Knowledge Engine. *Humanities*, 6(4):85.

Rahul Sambaraju, Ehud Reiter, Robert Logie, Andy McKinlay, Chris McVittie, Albert Gatt, and Cindy Sykes. 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*, pages 22–31. Association for Computational Linguistics.

Donia Scott and Johanna Moore. 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*, pages 22–23.

JJ Shaughnessy, EB Zechmeister, and JS Zechmeister. 2006. *Research methods in psychology*. McGraw-Hill.

Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, USA. Association for Machine Translation in the Americas.

Karen Sparck Jones and Julia R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer, Berlin and Heidelberg.

Somayajulu Sripada, Ehud Reiter, and Lezan Hawizy. 2005. Evaluation of an NLG system using post-edit data: Lessons learnt. In *Proceedings of the 10th European Workshop on Natural Language Generation*, pages 133–139, Aberdeen, Scotland. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*. Accepted for publication as a short paper at EMNLP 2018.

Jonathan Swift. 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...*, volume 1. Benjamin Motte, London, UK.

John W Tukey. 1980. We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25.

Joseph P Turian, Luke Shen, and I Dan Melamed. 2003. Evaluation of machine translation and its evaluation. In *Proceedings of MT Summit IX*.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 288–297, Gothenburg, Sweden. Association for Computational Linguistics.

Renske Van Enschot, Wilbert Spooren, Antal van den Bosch, Christian Burgers, Liesbeth Degand, Jacqueline Evers-Vermeul, Florian Kunneman, Christine Liebrecht, Yvette Linders, and Alfons Maes. 2017. Taming our wild data: On intercoder reliability in discourse research. *Unpublished Manuscript*, pages 1–18.

David L Vannette. 2016. Testing the effects of different types of attention interventions on data quality in web surveys. Experimental evidence from a 14 country study. In *71st Annual Conference of the American Association for Public Opinion Research*.

Paul Vogt, Rianne van den Berghe, Mirjam de Haas, Laura Hoffman, Junko Kanero, Ezgi Mamus, Jean-Marc Montanier, Cansu Oranc, Ora Oudgenoeg-Paz, Daniel Hernandez Garcia, , Fotios Papadopoulos, Thorsten Schodde, Josje Verhagen, Christopher Wallbridge, Bram Willemsen, Jan de Wit, Tony Belpaeme, Tilbe Göksun, Stefan Kopp, Emiel Krahmer, Aylin Küntay, Paul Leseman, and Amit Kumar Pandey. 2019. Second language tutoring using social robots: A large-scale study. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE.

Eric-Jan Wagenmakers. 2007. A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, 14(5):779–804.

Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, and Han van der Maas. 2011. Why psychologists must change the way they analyze their data: the case of psi: Comment on Bem (2011). *Journal of personality and social psychology*, 100(3):426.

Jelte M Wicherts, Coosje LS Veldkamp, Hilde EM Augusteijn, Marjan Bakker, Robbie Van Aert, and Marcel ALM Van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology*, 7:1832.

Georgios N Yannakakis and John Hallam. 2011. Ranking vs. preference: A comparative study of self-reporting. In *International Conference on Affective Computing and Intelligent Interaction*, pages 437–446. Springer.

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

Sina Zarrieß, Sebastian Loth, and David Schlangen. 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*, pages 38–47, Brighton, UK. Association for Computational Linguistics.