How to Do Human Evaluation: Best Practices for User Studies in NLP

Anonymous ACL submission

Abstract

Many research topics in natural language processing (NLP), such as explanation generation, dialog modeling or machine translation, require evaluation that goes beyond standard metrics like accuracy or F_1 score toward a more human-centered approach. Therefore, understanding how to design user studies becomes increasingly important. However, few comprehensive resources exist on planning, conducting and evaluating user studies for NLP, making it hard to get started for researchers without prior experience in the field of human evaluation. In this paper, we summarize the most important aspects of user studies and their design and evaluation, providing direct links to NLP tasks and NLP specific challenges where appropriate. We (i) outline general study design, ethical considerations, and factors to consider for crowdsourcing, (ii) discuss the particularities of user studies in NLP and provide starting points to select questionnaires, experimental designs and evaluation methods that are tailored to the specific NLP tasks. Additionally, we offer examples with accompanying statistical evaluation code in R throughout, to bridge the gap between theoretical guidelines and practical applications.[1]

1 Introduction

Over the past years, the natural language processing (NLP) community has increasingly expressed the need for and the importance of human evaluation to complement automatic evaluation (Belz and Reiter, 2006). While human evaluation has received much attention in the context of natural language generation (Belz and Reiter, 2006; Novikova et al., 2018; van der Lee et al., 2019), it can benefit many additional fields within NLP.

Tasks, such as machine translation (Graham et al., 2013), explanation generation (Nguyen, 2018; Narang et al., 2020; Clinciu et al., 2021), text-to-speech generation (Cardoso et al., 2015; Clark et al., 2019), question answering (Chen et al., 2019; Schuff et al., 2020), and automatic summarization (Owczarzak et al., 2012; Paulus et al., 2017) benefit from human evaluation as automatic evaluation scores, such as BLEU or F_1, are limited in reflecting how humans perceive a system's qualities (Callison-Burch et al., 2006; Liu et al., 2016; Schuff et al., 2020; Clinciu et al., 2021).

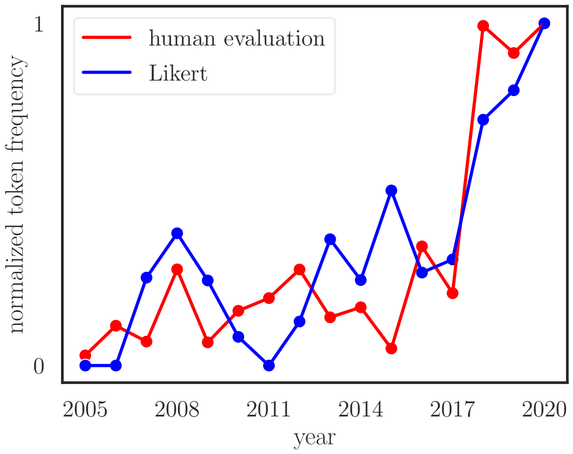


Figure 1: Normalized frequencies of "human evaluation" and "Likert" (as in the Likert scale questionnaire type) in the ACL anthology from 2005 to 2020 showing the growing attention on human evaluation.

As more and more systems are deployed into real-world applications, human evaluation has also gained more attention in the NLP community (Figure 1). On the one hand, there are numerous text books on human evaluation, experimental design, and experimental evaluation (Dean et al., 1999; Field and Hole, 2002; Field, 2013; Montgomery, 2017). However, they can become overwhelming for a practically-oriented researcher due to their breadth of topics. On the other hand, there are task-specific NLP resources. For example, van der Lee et al. (2019) provide guidelines on human evalu-

[1] https://removed-for-anonymity.edu

ation of natural language generation (NLG) and Sedoc et al. (2019) present an evaluation methodology specifically for chat bots. These contain valuable details for the particular evaluation setting, but lack discussions of broader human aspects, such as ethical considerations and cross-task NLP topics, such as crowdsourcing. Similarly, Dror et al. (2018) focus on statistical significance testing in NLP for automatic evaluation, but do not touch upon the needs of human evaluation analyses.

In this paper, we provide an overview that focuses on commonalities of human evaluation across NLP without restriction to a single task and thereby aim to provide a good balance between generality and relevance and foster an overall understanding of important aspects in human evaluation, how they are connected, and where to find more information.

In particular, we address NLP researchers who are new to human evaluation and walk them through how to formulate hypotheses (§2), determine which are (in)dependent and confounding variables (§3), choose appropriate metrics and questionnaires and know their level of measurement (§4), select a suitable experimental design (§5), set up a crowdsourced study (§6), calculate appropriate statistics (§7), and be aware of ethical considerations (§8). We complement our discussions with concrete examples from various NLP tasks. In the paper, we mainly include example snippets due to space reasons. In addition, we provide comprehensive examples for automatic summarization, explanation generation and dialog modeling in the supplementary material. Finally, we publish toy data with corresponding statistical evaluation code.

2 Research Questions and Hypotheses

In essence, the purpose of a user study is to answer one or more research questions. These broadly fall into two categories: (i) *Confirmatory*, where the research question aims to test a specific assumption, e.g., “Does the persona of dialog system B change the users’ enjoyment of the interaction compared to that of system A?” and (ii) *exploratory*, where the research question’s purpose is to generate assumptions, which can then be tested in a subsequent confirmatory research question, e.g., “Which factors influence the users’ enjoyment of system B?”. This distinction has a direct influence on all later stages of the study. In NLP research, the approach of sequentially improving models is predominant. For this, confirmatory, comparative research questions fit the best. Thus, we will focus on those in the remainder of the paper.

Once one or more research questions have been chosen, they need to be transformed into hypotheses, which propose a relationship between multiple variables. Staying with our example, the hypothesis “The new system B changes users’ enjoyment compared to the old system A” is called the *alternative hypothesis* in contrast to the *null hypothesis* that postulates there will be no change. In the former, our hypothesis assumes an effect of the variable “system type” on the variable “user enjoyment”.

3 Variables

Before we discuss experimental designs and evaluation methods, we first have to distinguish, which variables we are changing, which variables we are measuring and which we cannot control.

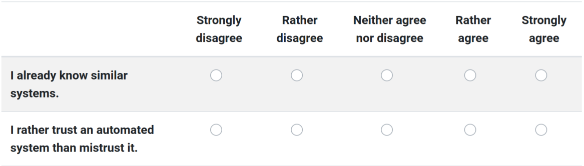
Independent. The independent variable(s) are those which we control in our study, also called *factors*. Experimental designs involving a single (or multiple) independent variable(s) are referred to as *unifactorial* (or *multifactorial*). The values a variable can take are called *levels*. For example, if the variable is “translation system”, levels might be “old system” and “new system”.

Dependent. The dependent or *response* variable(s) are those which we observe when changing the independent variables. For this, it is important to consider not just the general concept (*construct*), but also what concrete measurement to take. This process is known as *operationalization*. For example, we could measure the dependent variable “intelligibility” when varying “translation system” between “old system” and “new system”.

Confounding. A confounding variable or *confounder* is a variable that affects the dependent variable, but cannot be controlled for, e.g., age or gender of the participant. Education, for example, might affect “intelligibility” but one cannot deliberately change the education level of participants. Potential confounding variables should either be accounted for in the experiment design or in the statistical evaluation of the collected responses.

4 Metrics

Depending on the dependent variable, there are different means of quantifying user responses.



	Strongly disagree	Rather disagree	Neither agree nor disagree	Rather agree	Strongly agree
I already know similar systems.	○	○	○	○	○
I rather trust an automated system than mistrust it.	○	○	○	○	○

Figure 2: A subset of Likert items from the trust in automation scale by Körber (2018).

4.1 Likert Scales

While it is clear how to collect objective measures, e.g., the length of a dialog, it is less straightforward how to collect scores of trust or cognitive load. For such subjective metrics one usually obtains scores via a scale (Hart and Staveland, 1988; Langer and König, 2018; Körber, 2018), e.g., a questionnaire. A scale is designed to quantify a construct, e.g., "system usability", that may comprise of multiple dimensions, e.g, efficiency, effectiveness, and satisfaction (Brooke, 1996; Finstad, 2010). The most common type of scale is the Likert scale, containing (multiple) items, rated by the user on a discrete range. Figure 2 shows an example for a scale containing five-point Likert items. The overall score is calculated by combining the numbers related to the answer from each item (Körber, 2018). It is important to stress that the single questions are not scales themselves but items and the group of items makes the scale.

Using multiple items instead of a single rating allows one to assess the scale's internal consistency, e.g., via Cronbach's alpha (DeVellis, 2016).[2] Designing a valid and reliable scale requires a precise development process, summarized by Boateng et al. (2018) and explained in detail by DeVellis (2016). For NLP, the fields of psychology, human computer interaction, and robotics already offer a valuable range of scales. Validated questionnaires exist, for example, for evaluating trust (Körber, 2018), usability (Brooke, 1996; Finstad, 2010), cognitive load (Hart and Staveland, 1988), social attribution (Carpinella et al., 2017), or user interface language quality (Bargas-Avila and Brühlmann, 2016).

4.2 Other Useful Metrics for NLP

In tasks like generating text or speech, direct comparisons or ranked order comparisons (ranked output from multiple systems best to worst) can be a good option (Vilar et al., 2007; Bojar et al., 2016). Another option for tasks involving text generation is error classification, which involves users annotating text output from a set of predefined error labels (Secară, 2005; Howcroft et al., 2020). Santhanam and Shaikh (2019) showed that continuous rating scales can yield more consistent results than Likert scales for dialog system evaluation. Other measurements of interest include completion time, and bio-signals, such as gaze, EEG, ECG, and electrodermal activity. Bio-signals may provide insight into, e.g., emotional state (Kim and André, 2008), engagement (Renshaw et al., 2009), stress (McDuff et al., 2016), and gestures (Kim et al., 2008).

4.3 Level of Measurement

It is important to consider the scale on which a variable is measured in order to choose a correct statistical test (§7) and measures of central tendency (i.e., mode, median and mean).

On a **nominal** (categorical) scale, items are simply named, with no concept of order or distance between them. An example is emotions perceived in a generated voice ("happiness", "sadness","fear", etc.). The only measure of central tendency applicable to such data is the mode. An **ordinal** scale adds order to the elements. An example is measuring intelligibility using the values "very low", "low", "medium", "high", and "very high". In addition to the mode, ordinal data also allows to derive the median. On an **interval** scale, the elements are ordered with an equal distance between them, allowing one to additionally take the mean. Scores obtained from Likert scales are commonly regarded as interval data.[3] An example is measuring user trust in an explainable AI system using the scale from Figure 2. A **ratio** measurement adds the property of a true zero point making ratios of interval measurements sensible. An example is interaction time with a system.

5 Experimental Designs

Next one has to choose how participants are assigned to *conditions*, i.e., to levels of the independent variable(s). This design determines applicable

[2] Although, we cannot directly assess how well an item is related to the latent variable of interest (e.g., trust) because this is what we want to capture via the items, we still can quantify these relationships indirectly via item-item correlations. If the items have a high correlation with the latent variable, they will have a high correlation among each other (DeVellis, 2016).

[3] There has been a long debate between *ordinalists* who claim that Likert scales should be treated as ordinal data and non-parametric statistics have to be used, and *intervalists* who argue for an interval interpretation and thus support parametric approaches (Jamieson, 2004; Carifio and Perla, 2008; De Winter and Dodou, 2010) For a deeper discussion as well as practical recommendations, we refer to Harpe (2015).

statistical tests and can mitigate confounding effects. To illustrate the design choices, we will use the example of investigating the correctness of a summarization system with the independent variable "system", the levels "old" and "new", and the confounding variable that some participants are native speakers while others are not.

Within-Subject. In this study design, also called a *repeated-measures* design, participants are exposed to all study conditions. This may cause participant responses of later conditions to be affected by their responses to earlier conditions due to *carry-over effects*. One way to account for this is to control the order of conditions the participants are exposed to. Typical approaches are *randomization*, *blocking*, and *Latin square* designs. For details, we refer to Dean et al. (1999). For a within-subject design, a paired statistical test must be used.

In our example, we could use a within-subject approach and mitigate carry-over effects by sampling all possible four combinations[4] equally often. We could account for the possibly confounding effect of being a native speaker by balancing the number of native/non-native speakers per condition.

Between-Subjects. In this design, each participant is only exposed to one condition. Participant responses collected with a between-subjects design must use unpaired tests.

In our example, it could be preferable to use a between-subjects approach if the interaction of the users with the system takes long and, thus, users could become fatigued when being exposed to both conditions (i.e., old and new system).

Comparison. With the same number of participants, a within-subject design allows one to collect more samples than a between-subjects design. In contrast, a between-subjects design can easily be scaled to an arbitrarily high number of conditions while a within-subjects design is limited by participants' fatigue and willingness to participate in very long studies.

6 Crowdsourcing for NLP

Crowdsourcing provides an attractive way to quickly collect responses from a population that has been shown to be more diverse than samples from, e.g., college students and internet samples (Buhrmester et al., 2016). In NLP, Schnoebelen and Kuperman (2010) find crowdsourcing to be a reliable source for linguistic data. However there are differences between designing a crowdsourcing study and a traditional lab experiment, which we will overview in the following paragraphs.

Fair Compensation. In a traditional study, participants are often volunteers interested in aiding research. On crowdsourcing platforms, participants might not have another full time job and rely on the money they earn by completing tasks (Williamson, 2016). Therefore it is important to ensure your pay structure is non-exploitative. If a user is unable to complete a task due to an error in the task, their time should still be respected.

Platform rules. Different platforms, e.g., Amazon Mechanical Turk, CrowdFlower, MicroWorkers, Prolific Academy or Qualtrics, have different rules and capabilities. For example, some require participants to be paid on completion of task, while others allow the results to be reviewed first. Some only support users filling out surveys, while others allow for building more complex interactions.

Task description. The task description should explicitly contain all necessary steps that a worker needs to fulfill in order to be paid, as well as the estimated time a task will take. It should give workers an accurate idea of expectations so they can make an informed choice about accepting the task.

Incentives. Crowdsourcing workers often want to get through an experiment quickly to maximize their pay, so this should be kept in mind when designing an experience. Including attention checking or free-response questions can help ensure workers are not just clicking through tasks to finish quickly (Meade and Craig, 2012). We also recommend that experiments are designed such that workers cannot submit a task unless they have completed all subtasks. For example, if evaluating a speech generation system, the user must actually play samples before they can be evaluated. Finally, keep interactions as short as possible as participants may suffer from survey fatigue (i.e., giving less thoughtful answers over time) (Ben-Nun, 2008).

Pilot study. Pilot studies, i.e., small scale trials before a larger study, allow for testing the experimental design and technical set-up. Performing pilot studies allows researchers to discover errors

[4] (i) native speaker: "old" first → "new" second, (ii) native speaker: "new" → "old", (iii) not native speaker: "old" → "new", (iv) not native speaker: "new" → "old".

early on, saving resources and time. For more details on designing pilot studies, c.f., Van Teijlingen and Hundley (2001) and Hassan et al. (2006).

Note that pilot studies conducted in a lab setting may not generalize to the data collected on crowdsourcing websites, due to the difference in populations. Thus, it is a good idea to also conduct a small pilot study on the crowdsourcing platform.

Data Logging. If an experiment involves anything more than a survey, the interaction of the user with the system will probably generate interesting data itself. Even if it does not seem immediately relevant to the research goal, logging is relatively cheap and can provide insights when analysing the experimental data. Additionally, if the focus of the experiment shifts, rather than re-running the study, the "extra" data logged might already contain the needed information. For example if we want to measure translation quality, it might also be a good idea to log mouse movement and time taken to rate each translation as these might later provide insights into how comprehensible translation were. It is important to note, however, that users should be informed of any data collected and personally identifying data should be avoided.

Further reading. We refer to Pavlick et al. (2014) for a discussion of Mechanical Turk's language demography and to Paolacci (2010), Schnoebelen and Kuperman (2010) for further advice on conducting a crowdsourcing study as well as Jacques and Kristensson (2019) for information on crowdsourcing economics.

7 Statistical Evaluation for NLP

In their review of INLG and ACL papers that conduct a human evaluation, van der Lee et al. (2019) note that only 33% of the papers report statistical analyses. This section aims to offer a guideline to choose an appropriate sample size, select an applicable statistical test and decide whether a post-hoc test and a multiplicity adjustment need to be used.

7.1 Estimating the Required Sample Size

Before starting a user study, an important step is to consider what sample size will be necessary to make meaningful claims about the results. This number will greatly depend on what *power* the study should have. Statistical power expresses the probability of recognizing a statistically significant difference in the data if one occurs, in short the likelihood of not reporting a false negative. A power level of 0.80 or higher is generally recommended (Bausell and Li, 2002).

The power of a study is dependent on the expected *effect size*, the number of participants (N), and the statistical test that will be used. The effect size refers to how large the expected difference is between experimental groups. The smaller the effect size is, the greater the required number of participants will be, in order to show that differences between experimental groups are not just due to chance. While estimating effect size can be difficult, some useful starting points could come from previous research or from looking at the results from a pilot study — which is always a good idea to conduct before launching a large-scale study. Additionally, a meta study of 302 social and behavioral meta-analsyses, Lipsey and Wilson (1993), found the average effect size to be exactly 0.5.

For more information, including tables with the relationship between power, N, and hypothesized effect size as well details on calculating power, Dean et al. (1999), Bausell and Li (2002) and Montgomery (2017) provide a solid introduction to the topic and VanVoorhis et al. (2007) discuss common rules of thumbs of sample size.

7.2 Choosing the Correct Statistical Test

The (set of) applicable statistical test(s) is determined by the experimental setup including the choice of measurement scale (§4.3) and the experimental design (§5). To choose a test, one has to determine the number of levels (groups), whether the samples were collected in a paired or unpaired design, the measurement scale of the dependent variable and whether parametric assumptions are met. In the following, we discuss these aspects and present common tests. Table 1 lists the discussed tests along the conditions in which they are applicable.

Paired and Unpaired Tests. Whether a paired or an unpaired test is the correct choice directly depends on the choice of experimental design, which we discussed in §5. The paired test is applicable if the samples were collected in a within-subject design (repeated measures), i.e., from one group. In the in-between design, the two samples were collected from different groups and an unpaired test has to be applied.

Parametric and Non-Parametric Tests. Parametric tests make assumptions on the underlying population distribution (such as normality), non-parametric tests do not make assumptions on the distributions, but still can make other assumptions (Colquhoun, 1971). Therefore, the measurement scale of the dependent variable can directly determine whether a parametric test is applicable. For example, we cannot run a parametric t-test (which is parametric) on ordinal responses from {“often”, “sometimes”, “never”}. It is often claimed that parametric tests offer higher statistical power. This statement has to be restricted to very specific conditions and Colquhoun (1971) argues to prefer non-parametric tests as long as there is no experimental evidence of the error distribution. We refer to Colquhoun (1971) for a discussion of the differences between parametric and non-parametric methods and to Sprent (2012) and Corder and Foreman (2014) for details on non-paramteric statistics.

Paired	Param.	Scale	Test
✗	✓	interval	unpaired t-test
✗	✗	ordinal	Mann-Whitney U test
✗	✗	nominal	Chi-square (χ^2) test
✗	✗	dichotomous	Fisher’s exact test
✓	✓	interval	paired t-test
✓	✗	ordinal*	Wilcoxon signed-rank test
✓	✗	ordinal	sign test
✓	✗	nominal	McNemar test
✗	✓	interval	one-way ANOVA
✗	✗	ordinal	Kruskal-Wallis test
✓	✓	interval	repeated-measures ANOVA
✓	✗	ordinal	Friedmann test

Table 1: Frequently used parametric and corresponding non-parametric tests. The upper part of the table contains tests that compare two groups, the lower part lists tests that compare more groups. *The pairwise differences have to be on an ordinal scale, see Colquhoun (1971) for more details.

7.2.1 Frequently-Used Tests for NLP

In the following, we present a selection of the common statistical tests, highlight important assumptions they make and provide examples of NLP applications they are relevant to. We do not exhaustively discuss all assumptions of each test here, but instead, want to offer first guidance in choosing the right test. We first discuss tests that are applicable to experiment designs with one factor that has two levels (e.g., the factor chatbot system with the levels “system A” and “system B”).

Thereafter, we consider tests involving one factor with more than two levels (e.g., the factor chatbot system with an additional third “system C”). These tests are called *omnibus tests*, which means that they only can detect that “there is a difference” but make no statement about pairwise differences. Therefore, pairwise post-hoc tests are usually used after detecting a significant difference with an omnibus test.

Unpaired and Paired Two-Sample t-Test. In the context of user studies, the t-test is usually used to test if the means of two samples differ significantly, i.e., a *two-sample* t-test.[5] In NLG evaluation, the time a participants take to read a sentence generated by one versus another system could be compared using a t-test. For the two-sample test one further distinguishes an *unpaired* or *independent* test and a *paired* or *dependent* test. The t-test assumes that the errors follow a normal distribution which is usually decided subjectively by inspecting the quantile-quantile (Q-Q) plot of the data (Hull, 1993). When analyzing Likert scale responses, the choice of test depends on whether one regards the scale scores to be measures to be ordinal or interval measures (§4.3). However, De Winter and Dodou (2010) compare error rates between the non-parametric Mann-Whitney U test with the parametric t-test for five-point Likert items and find that both tests yield similar power.

Mann-Whitney U and Wilcoxon Signed-Rank. Although the t-test can be robust to violations of normality (Hull, 1993), non-parametric alternatives, such as the Mann-Whitney U for unpaired samples and the Wilcoxon signed-rank test for paired samples are preferable. The Mann-Whitney U test is the non-parametric counterpart to the unpaired t-test. In contrast to the t-test, which is restricted to interval data, it is additionally applicable to ordinal data as well as interval data that does not fulfill the parametric assumptions. For example, testing user acceptance of a voice assistant could involve asking participants how often they would use the system: "daily", "weekly", "monthly" or "never". The paired counterpart to the Mann-Whitney U test is the Wilcoxon signed-rank test which compares median differences between the two groups and can be applied as long as the pairwise differences between samples can be ranked. If this is not possible, a sign test should be used instead (Colquhoun, 1971).

[5] A *one-sample* t-test compares a sample’s mean with a predefined reference mean.

Fisher's Exact, χ^2 and McNemar Test. If the measurement scale nominal, the Mann-Whitney U and the Wilcoxon signed rank test are not applicable. Instead, Fisher's exact test test should be used for unpaired groups if the dependent variable is *dichotomous*, i.e., can only take two values like "yes" and "no", e.g. for rating the correctness of answers generated by a question answering system. If it can take more values, e.g. additionally "I do not know", a chi-square (χ^2) test can be used. When the samples are paired, the test of choice should be the McNemar test.

One-Way and Repeated-Measures ANOVA. So far, we only addressed tests that compare two groups, such as samples from "dialog system A" to samples from "dialog system B". When we add a third or more conditions, the discussed tests are no longer applicable. Instead, if the samples are parametric, a one-way ANOVA can be applied to unpaired samples and a repeated-measures ANOVA can be applied to paired samples.

Kruskal-Wallis and Friedmann Test. Like the Mann-Whitney U and the Wilcoxon-signed rank test are the non-parametric counterparts to the paired and unpaired t-test, one can use the non-parametric Kruskal-Wallis test instead of a one-way ANOVA and the non-parametric Friedmann test instead of a repeated-measures ANOVA.

More Complex Models and Tests. Besides the discussed tests, there also are more general models and accompanying tests. If the experimental setup requires to account for, e.g., subject-specific influences (e.g., mother tongue or literacy) or repeated measures of one factor within a mixed design (e.g., a design in which each participant uses one dialog system, i.e. a between-subjects factor, but all participants perform the same tasks, i.e., a within-subject factor), *(generalized) linear mixed models (GLMMs)* can be the appropriate evaluation tool. The difference between a linear model and a generalized linear mixed model is that the linear model is (i) *generalized* with respect to the distribution of the response variable (e.g., to predict a binary response) and (ii) extended to include *random effects* such as individual participant characteristics on top of *fixed effects* such as "system type" resulting in a *mixed* model. Intuitively, the purpose of including random effects is to get a clearer picture of the fixed effects and not to falsely attribute, e.g., the effect of participant age to be a difference between two chat bots. These models can be particularly relevant for crowdsourcing studies where crowdworkers participate in some, but neither all nor only one condition. An introduction to linear mixed models and their usage in R is provided by Winter (2013). More details can be found in McCulloch and Neuhaus (2005) and Jiang (2007).

7.3 Post Hoc Tests

The presented omnibus tests do not allow to make statements about pairwise differences between conditions. For example, an ANOVA might detect a significant difference within the groups {"system A", "system B", "system C"} but makes no statement if there is for example a significant difference between "system A" and "system B". In such cases one needs to use a post-hoc test. The respective post-hoc test is typically only applied if the omnibus test found a significant effect and — depending on the method — requires a multiple testing adjustment. Commonly used tests are Tukey HSD, Scheffé, Games-Howell, Nemenyi and Conover.

7.4 The Multiple Comparisons Problem

The intuition behind the multiple comparisons problem or *multiplicity* is that when many tests are conducted, each test bears the risk of a Type I error and that with many tests, one is, overall, much more likely to mistakenly report a statistically significant difference. In such cases the individual α levels need to be adjusted. A simple and well-known adjustment method is the Bonferroni correction, that divides the α level by the number of tests. However, this method is typically considered to be too conservative. Therefore, improved methods, such as the Holm procedure or the Hochberg technique are recommended (Bender and Lange, 2001; Streiner and Norman, 2011). When and when not to apply α adjustments was discussed vividly (Rothman, 1990; Ottenbacher, 1998; Moyé, 1998; Bender and Lange, 2001; Streiner and Norman, 2011).[6]

7.5 Further Analysis Methods for NLP

As NLP systems are frequently evaluated in side-by-side comparisons, the collected variables can be ranks or preferences (Callison-Burch et al., 2007; Grundkiewicz et al., 2015). For example, participants can be asked to rank pairs of translations or

[6] We refer to Bender and Lange (2001) and Streiner and Norman (2011) for brief but comprehensive discussions of the multiple comparisons problem and different recommendations when and when not to correct for multiple testing.

generated speech snippets. TrueSkill™ (Herbrich and Graepel, 2006; Sakaguchi et al., 2014) can be used to construct ranks from pairwise preferences. Pairwise preferences can be analyzed statistically using (log-linear) Bradly-Terry models (Bradley and Terry, 1952; Dras, 2015) or approaches based on item response theory (Sedoc et al., 2019). Further, hybrid approaches that combine ranking with scale ratings (Novikova et al., 2018) or human judgements with automatic evaluation (Hashimoto et al., 2019) have been proposed for NLG.

7.6 Example

To showcase a complete statistical analysis, we consider a scenario in which we want to compare three chat bot systems with respect to the levels of trust they evoke in users. More formally, we investigate the effect of three levels of the independent variable "personalization" on the variable user trust. We suppose that we operationalize user trust using the trust scale by Körber (2018) and consider the scale scores to lie on an interval scale. We assume that we conducted a pilot study and collected the full study data using a within-subject design balancing for native speakers. The next step is to determine an appropriate statistical test. For this example, we suppose that a Q-Q plot indicated that the collected responses are not parametric. Since we chose a within-subject design, the ratings are paired. Therefore we need to use a paired non-parametric test and choose the Friedmann test. Supposing the Friedmann test detects a significant difference, we subsequently run a Nemenyi test to determine which pairs of groups significantly differ. In our example, we find that trust ratings of two levels of personalization are significantly higher than the third level without a significant difference between the two.[7]

8 Ethical and Legal Considerations

When designing an experiment involving human participation, it is also critical to consider ethical and legal implications.

Privacy. In the EU, legal considerations include respecting participant's data privacy in compliance with the GDPR. In particular, the clauses on participant's right to erasure, right to information, and right to restriction of processing should be considered. It is therefore necessary to have a data agreement for participants before they decide to take part in an experiment, informing participants what data will be collected, how it will be used, and how long any personally identifying data (e.g., video or speech recordings) will be stored (Commission, 2018). Legal requirements may vary by country/ locality, therefore it is important to check with your research institution before planning an experiment. However, ethically, data protection should be considered regardless of legal obligations.

Consent. Additionally, it is important to make sure participants have true informed consent before beginning an experiment (Association et al., 2002; Commission, 2018; Association; Code, 1949). This means that participants should know the purpose of the research, that they have the right to end participation at any time, the potential risks an experiment poses/factors why someone might not want to participate, prospective benefits of the experiment, any limits to confidentiality, such as how the data collected will be used or published, incentives for participation, and contacts in case of questions.

Respect for Participants. In addition to consent and privacy issues, researchers should also prioritize the dignity of participants. Studies should be conducted in order to provide a benefit to society rather than randomly. That said, participant welfare must take a priority over the interests of science and society. And studies should be conducted so as to avoid all unnecessary physical and mental suffering (e.g., intentionally inducing negative emotions) and injury (Association et al., 2002; Code, 1949; Association). For further reading we refer to Shaw (2003) and Leidner and Plachouras (2017).

9 Conclusion

In this paper, we provided an overview of the most important aspects for human evaluation in natural language processing. We guided the reader along the way from research questions to statistical analysis, reviewed general experimental design approaches, discussed ethical and legal considerations and gave NLP-specific advice on metrics, crowdsourcing and evaluation techniques. We complemented our discussions with numerous example scenarios from NLP and a code example for a statistical analysis with R. Thereby, we offered a quick start guide for NLP researchers who are new to the field of human evaluation and provided pointers to in-depth resources.

[7] We provide toy data and code for the described statistical analysis at `www.removed-for-anonymity.edu` (see supplementary material).

References

American Psychological Association et al. 2002. Ethical principles of psychologists and code of conduct. *American psychologist*, 57(12):1060–1073.

World Medical Association. Wma declaration of helsinki – ethical principles for medical research involving human subjects.

Javier A Bargas-Avila and Florian Brühlmann. 2016. Measuring user rated language quality: development and validation of the user interface language quality survey (lqs). *International Journal of Human-Computer Studies*, 86:1–10.

R Barker Bausell and Yu-Fang Li. 2002. *Power analysis for experimental research: a practical guide for the biological, medical and social sciences*. Cambridge University Press.

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*.

Pazit Ben-Nun. 2008. Respondent fatigue. *Encyclopedia of survey research methods*, 2:742–743.

Ralf Bender and Stefan Lange. 2001. Adjusting for multiple testing—when and how? *Journal of clinical epidemiology*, 54(4):343–349.

Godfred O. Boateng, Torsten B. Neilands, Edward A. Frongillo, Hugo R. Melgar-Quiñonez, and Sera L. Young. 2018. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, 6:149.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016. 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.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs. *Biometrika*.

John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry*, page 189.

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2016. Amazon's mechanical turk: A new source of inexpensive, yet high-quality data?

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Walcir Cardoso, George Smith, and C Garcia Fuentes. 2015. Evaluating text-to-speech synthesizers. In *Critical CALL–Proceedings of the 2015 EUROCALL Conference, Padova, Italy*, pages 108–113. Research-publishing. net.

James Carifio and Rocco Perla. 2008. Resolving the 50-year debate around using and misusing likert scales. *Medical education*, 42(12):1150–1152.

Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. 2017. The Robotic Social Attributes Scale (RoSAS): Development and Validation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 254–262, Vienna Austria. ACM.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124.

Rob Clark, Hanna Silen, Tom Kenter, and Ralph Leith. 2019. Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs. *arXiv preprint arXiv:1909.03965*.

Miruna-Adriana Clinciu, Arash Eshghi, and Helen Hastie. 2021. A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2376–2387, Online. Association for Computational Linguistics.

Nuremberg Code. 1949. The nuremberg code. *Trials of war criminals before the Nuremberg military tribunals under control council law*, 10:181–182.

David Colquhoun. 1971. *Lectures on biostatistics: an introduction to statistics with applications in biology and medicine*. David Colquhoun.

European Commission. 2018. 2018 reform of eu data protection rules.

Gregory W Corder and Dale I Foreman. 2014. *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons.

JFC De Winter and Dimitra Dodou. 2010. Five-point likert items: t test versus mann-whitney-wilcoxon (addendum added october 2012). *Practical Assessment, Research, and Evaluation*, 15(1):11.

Angela Dean, Daniel Voss, Danel Draguljić, et al. 1999. *Design and analysis of experiments*, volume 1. Springer.

Robert F DeVellis. 2016. *Scale development: Theory and applications*, volume 26. Sage publications.

Mark Dras. 2015. Squibs: Evaluating human pairwise preference judgments. *Computational Linguistics*, 41(2):309–317.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 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)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Andy Field. 2013. *Discovering statistics using IBM SPSS statistics*. sage.

Andy Field and Graham Hole. 2002. *How to design and report experiments*. Sage.

Kraig Finstad. 2010. The usability metric for user experience. *Interacting with Computers*, 22(5):323–327.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal. Association for Computational Linguistics.

Spencer E Harpe. 2015. How to analyze likert and other rating scale data. *Currents in pharmacy teaching and learning*, 7(6):836–850.

Sandra G Hart and Lowell E Staveland. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Zailinawati Abu Hassan, Peter Schattner, and Danielle Mazza. 2006. Doing a pilot study: why is it essential? *Malaysian family physician: the official journal of the Academy of Family Physicians of Malaysia*, 1(2-3):70.

Ralf Herbrich and Thore Graepel. 2006. Trueskill(tm): A bayesian skill rating system. Technical Report MSR-TR-2006-80.

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

David Hull. 1993. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–338.

Jason T. Jacques and Per Ola Kristensson. 2019. Crowdworker economics in the gig economy. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–10, New York, NY, USA. Association for Computing Machinery.

Susan Jamieson. 2004. Likert scales: How to (ab) use them? *Medical education*, 38(12):1217–1218.

Jiming Jiang. 2007. *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media.

Jonghwa Kim and Elisabeth André. 2008. Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence*, 30(12):2067–2083.

Jonghwa Kim, Stephan Mastnik, and Elisabeth André. 2008. Emg-based hand gesture recognition for realtime biosignal interfacing. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 30–39.

Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*, pages 13–30. Springer.

Markus Langer and Cornelius J. König. 2018. Introducing and Testing the Creepiness of Situation Scale (CRoSS). *Frontiers in Psychology*, 9:2220.

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

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Mark W Lipsey and David B Wilson. 1993. The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American psychologist*, 48(12):1181.

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

Charles E McCulloch and John M Neuhaus. 2005. Generalized linear mixed models. *Encyclopedia of biostatistics*, 4.

Daniel J. McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W. Picard. 2016. *COGCAM: Contact-Free Measurement of Cognitive Stress During Computer Tasks with a Digital Camera*, page 4000–4004. Association for Computing Machinery, New York, NY, USA.

Adam W Meade and S Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological methods*, 17(3):437.

Douglas C Montgomery. 2017. *Design and analysis of experiments*. John wiley & sons.

Lemuel A Moyé. 1998. P-value interpretation and alpha allocation in clinical trials. *Annals of Epidemiology*, 8(6):351–357.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. Rankme: Reliable human ratings for natural language generation. *arXiv preprint arXiv:1803.05928*.

Kenneth J Ottenbacher. 1998. Quantitative evaluation of multiplicity in epidemiology and public health research. *American Journal of Epidemiology*, 147(7):615–619.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montréal, Canada. Association for Computational Linguistics.

Gabriele Paolacci. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):9.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The Language Demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Tony Renshaw, Richard Stevens, and Paul D Denton. 2009. Towards understanding engagement in games: an eye-tracking study. *On the Horizon*.

Kenneth J Rothman. 1990. No adjustments are needed for multiple comparisons. *Epidemiology*, pages 43–46.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA. Association for Computational Linguistics.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.

Tyler Schnoebelen and Victor Kuperman. 2010. Using amazon mechanical turk for linguistic research. *Psihologija*, 43(4):441–464.

Hendrik Schuff, Heike Adel, and Ngoc Thang Vu. 2020. F1 is Not Enough! Models and Evaluation Towards User-Centered Explainable Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7076–7095, Online. Association for Computational Linguistics.

Alina Secară. 2005. Translation evaluation: A state of the art survey. In *Proceedings of the eCoLoRe/MeLLANGE workshop, Leeds*, volume 39, page 44. Citeseer.

João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. ChatEval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.

Ian F Shaw. 2003. Ethics in qualitative research and evaluation. *Journal of social work*, 3(1):9–29.

Peter Sprent. 2012. *Applied nonparametric statistical methods*. Springer Science & Business Media.

David L Streiner and Geoffrey R Norman. 2011. Correction for multiple testing: is there a resolution? *Chest*, 140(1):16–18.

Edwin R Van Teijlingen and Vanora Hundley. 2001. The importance of pilot studies.

CR Wilson VanVoorhis, Betsy L Morgan, et al. 2007. Understanding power and rules of thumb for determining sample sizes. *Tutorials in quantitative methods for psychology*, 3(2):43–50.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E Banchs. 2007. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 96–103.

Vanessa Williamson. 2016. On the ethics of crowdsourced research. *PS: Political Science & Politics*, 49(1):77–81.

Bodo Winter. 2013. Linear models and linear mixed effects models in r with linguistic applications.

Task	Summarization	Explanation Generation	Dialog Modeling
Object of Investigation	Summarization and Reading Times	Explanations for Usability	Voice Assistant and Cognitive Load
Research Question	Do two summarization methods differ w.r.t. the time users need to read the summaries?	How do three textual explanation methods differ w.r.t. perceived usability?	Does the usage of a voice assistant reduce users' cognitive load during task solving?
Independent Variable(s)	summarization system (two levels)	explainability method (three levels)	assistant support (two levels) and task to solve (5 levels)
Dependent Variable(s)	reading time (ratio level)	scores on a usability scale such as SUS or UMUX (interval/ordinal)	scores from a cognitive load scale such as NASA-TLX (interval/ordinal)
Experimental Design	within-subject	between-subject	mixed design (between-subject for assistant support and within-subject for task)
Statistical Evaluation	paired t-test or Wilcoxon signed-rank test	one-way ANOVA or Kruskal-Wallis test + post hoc test (e.g. Tukey HSD)	Linear mixed model with fixed effects for assistance support and task and random effect for age + Likelihood-ratio test + post hoc test (e.g. Tukey HSD)

Table 1: We provide three fictional scenarios to demonstrate a range of user studies and showcase appropriate experimental designs and statistical evaluation approaches.