The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations

Jacopo Amidei and **Paul Piwek** and **Alistair Willis**
School of Computing and Communications
The Open University
Milton Keynes, UK
{jacopo.amidei, paul.piwek, alistair.willis}@open.ac.uk

Abstract

Rating and Likert scales are widely used in evaluation experiments to measure the quality of Natural Language Generation (NLG) systems. We review the use of rating and Likert scales for NLG evaluation tasks published in NLG specialized conferences over the last ten years (135 papers in total). Our analysis brings to light a number of deviations from good practice in their use. We conclude with some recommendations about the use of such scales. Our aim is to encourage the appropriate use of evaluation methodologies in the NLG community.

1 Introduction

Rating and Likert scales are popular tools used in surveys to estimate feeling, opinions or attitudes of responders. Although both instruments are widely used, their nature and their appropriate statistical analysis remain a matter of controversy. In particular, it can be controversial whether rating and Likert scales should be considered as ordinal or interval scales; see for example Knapp (1990), Jamieson (2004), Pell (2005), Carifio et al. (2008), Norman (2010) and Sullivan and Artino (2013). However, this distinction is of capital importance because it determines whether the statistical tool to be used on the collected data is parametric or non-parametric. Guidelines and good practice descriptions for the use and analysis of rating and Likert scales have been developed; see for example Knapp (1990), Kuzon et al. (1996), Pell (2005), Carifio et al. (2008), De Winter and Dodou (2010), Sullivan and Artino (2013), Harpe (2015), Joshi et al. (2015) and Johnson and Morgan (2016).

For this paper we analysed 135 papers published in NLG specialist conferences.[1] Our analysis brings to light common deviations from good practice in the use of rating and Likert scales. The aim of the present paper is to enhance awareness about the use of these scales in the NLG community. Indeed, both rating and Likert scales are widely used in evaluation experiments to measure the quality of NLG systems.

2 Related work

Our paper follows the path started by Robertson (2012), which highights deviations from statistical good practice in the area of Human Computer Interaction (HCI) and computer science education.

Regarding the basic statistics concepts and statistical analyses we refer to Witte and Witte (2017) and Johnson and Morgan (2016). A detailed description of Likert scales and their analysis is given in Joshi et al. (2015). Regarding the recommendations on the use of rating and Likert scales we refer to Knapp (1990), Kuzon et al. (1996), Pell (2005), Carifio et al. (2008), De Winter and Dodou (2010), Sullivan and Artino (2013), Harpe (2015), Joshi et al. (2015), Johnson and Morgan (2016).

A complete list of the papers we examined can be found via the following link https://bit.ly/2lKL516.

3 Rating and Likert scales

In this section we use illustrative examples to underline the differences between rating and Likert scales. We use the term *scale* with the following two meanings:

- Given a statement, the term *scale* is the group of points making up the options offered to respondents. We refer to the combination of the statement and the scale as an *item*.
- In the case of an aggregate scale[2], such as the

[1] Further information about the paper selection can be found in the supplementary material via the following link https://bit.ly/2lKL516.

[2] An aggregate or summated scale is a set of rating scales.

Likert scale, we use the term *scale* to indicate a collection of items.

Rating scales: Rating scales are items used in surveys to estimate feeling, opinions or attitudes of responders. The data collected through a rating scale can be interpreted both as ordinal and interval. A rating scale is composed of an n-point scale. Scales with 3, 5, 7, 10 or 11 points are used most often. Rating scales can be both numerical and verbal.

In a *numerical rating scale*, a number is associated with each point. A variation of a numerical scale uses label words at the extreme values and leaves the intermediate values with a numerical label, as for example shown in Figure 1. A rating scale that uses words as labels for the points is named a *graphic rating scale*[3]. An example of this kind of rating scale is pictured in Figure 2. Sometimes the points of a graphic rating scale can also be labelled with numbers. Another sort of rating scale is the *comparative rating scale*. This kind of scale is used to ask respondents to answer a question in terms of a comparison. An example of a comparative rating scale is given in Figure 3.

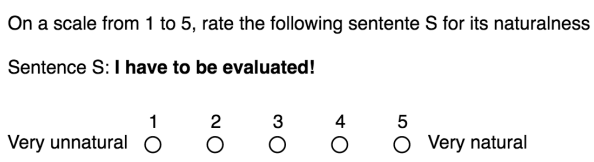

Figure 1: Example of a numerical rating scale.

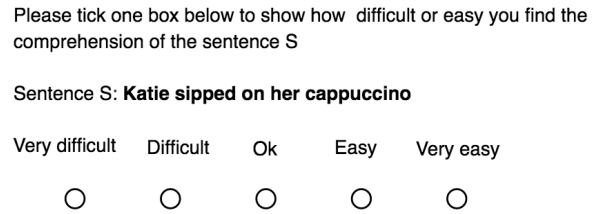

Figure 2: Example of a graphic rating scale.

Likert scale: A Likert scale is an aggregate scale. The items that make a Likert scale are graphic rating scales. In this context, each graphic rating scale is called a *Likert item*. Likert scales are usually expressed in terms of agreement and disagreement. An example of a Likert scale is shown in Figure 4. The items that make a Likert scales are designed to collectively capture the phenomenon under analysis. Accordingly, they shouldn't be considered in isolation and they should be summed or averaged to produce a total score. However, individual items by themselves are often considered as a single scale. Because of this ambivalent use of the Likert scale and its items, the nature of the Likert scale is highly controversial. Researchers are split between who consider it an interval scale and those who consider it an ordinal scale; see for example Jamieson (2004), Pell (2005), Norman (2010).

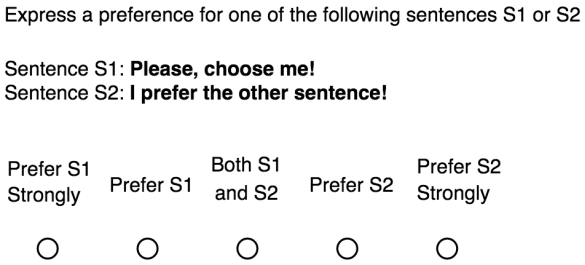

Figure 3: Example of a comparative rating scale.

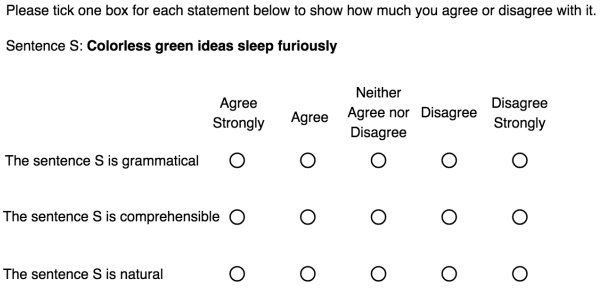

Figure 4: Likert scale example.

The confusion generated by the ambivalent use of the Likert scale and its items is well illustrated and explained in Joshi et al. (2015), where an image similar to Figure 5 is introduced. Likert scales are built in such a way that respondents express their level of agreement or disagreement with the sentences expressed by the Likert items. Because all the items are presented all together and with the same point labels, it is assumed that each respondent gives the same interpretation to the answer points – that is, as suggested by Likert, the

In other words, it is a composite of items which are summed or averaged all together to get an overall positive or negative orientation towards the object under examination in the survey.

[3]Sometimes a graphic rating scale is called *Likert item* or *Likert-style scale*. However, Likert items and Likert-style scale are particular cases of graphic rating scales.

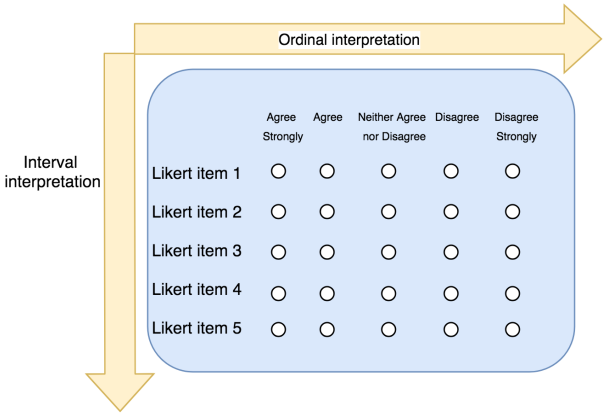

Figure 5: Likert scale interpretations.

distances between the points in the scale can be considered equal.[4] This assumption licenses use of the scale as an interval scale. Consequently, adding or averaging the items annotated by the same respondent is justified. This raises the interval interpretation, depicted by the vertical arrow of Figure 5.

Otherwise, an item-by-item analysis – that is a separate analysis of a single item extracted from an aggregate scale – cannot justify the assumption that the difference between adjacent points is equal. Indeed, we cannot assume that different respondents perceive the difference between adjacent label points as being of equal distance. The difference between "agree" and "strongly agree" can be perceived differently from one respondent to another. Consequently, the addition or the average of the items extracted from an aggregate scale is not justified. In such cases, the median or mode can be used as the measure of central tendency. This follows the ordinal interpretation, depicted by the horizontal arrow of Figure 5.

Unfortunately, in many cases there is not a clear understanding of the difference between the horizontal and the vertical direction of the aggregate scale. It is common to see item-by-item analysis (that is the horizontal direction) that makes use of parametric statistics without a justification of this choice. Indeed, as shown in Section 4, the interpretation of Likert items as interval scales has become a common practice. This particularly applies to the use of the mean for measuring the central tendency for the analysis of Likert items.

[4]Some authors, for example Jamieson (2004), do not accept such an assumption and do not consider the points as equally distant. In this case the Likert scales themselves, and not only the Likert items, are considered ordinal.

4 The use of rating and Likert scales in NLG evaluation tasks

In this section we present our analysis of 135 NLG papers.

First of all, it is important to note that several papers report the evaluation study in a very succinct manner that makes it difficult to understand and interpret the authors' conclusions.

From this observation follow two recommendations. First, researchers should be careful in the way they report the evaluation study. For instance, readers can benefit from examples and graphical and/or tabular presentation of data. Second, for the purpose of reproducibility, it is essential that evaluation guidelines and data are shared.

# Papers	# Rating	# Likert	# Others
135	48	37	50

Table 1: # *Papers*: Number of papers used in the study. # *rating*: Number of papers that use rating scales. # *Likert*: Number of papers that use Likert scales. # *Others*: Number of papers that use different kinds of human evaluation methodologies.

Table 1 shows that 63% of the papers used either a rating scale or a Likert scale. Between these papers, rating scales are used 56% of the time whereas Likert scales 44% of the time.

Because the majority of the papers we analysed report the evaluation study in an approximate manner, it is impossible to provide a statistic for the type of rating or Likert scale used. We found that in 64 papers, either it was not stated whether the rating or Likert scale was used, or the rating or Likert scale name used was imprecise. However, we can go as far as to say that the graphic scales and Likert item are the preferred rating scales used.

We found that the favourite scale dimension both for rating and Likert scales was the 5-point scale. Indeed, 31 papers use 5-point rating scales, and 23 papers use 5-point Likert scales.

Table 2 shows how the rating or Likert scales are interpreted. 16% of rating scales are interpreted as ordinal, whereas the 77% are interpreted as interval.[5] Likewise 16% of Likert scales are interpreted as ordinal, whereas the 84% are interpreted as interval. Table 2 shows the predom-

[5]7% of the papers do not give enough information to determine the interpretation used.

	Rating	Likert
Ordinal	8	6
Interval	37	31
(?)	3	0

Table 2: Number of rating and Likert scales which are considered Ordinal or Interval scales. The symbol (?) means we cannot determine the scale interpretation from the information given in the paper. We classified scales as ordinal or interval based on the statistic that was reported in the paper, i.e., whether the statistic used was parametric or nonparametric.

inant use of parametric statistics over nonparametric statistics in the papers we analysed.

Between the 68 papers (37 rating and 31 Likert) that interpret the data as interval, only 3 papers justify such an interpretation (2 rating and 1 Likert).

Regarding the use of Likert scales, we note that only one paper uses the Likert scale suitably, that is as an aggregate scale. All the other papers used Likert scales in order to perform item-by-item analysis.

For statistical significance testing, we found that ANOVA and the t-test are the preferred parametric statistics. Among nonparametric statistics the most commonly used are χ^2 and the Mann-Whitney U test.

5 Conclusion

From our analysis the following two main deviations from good practice in the use of rating and Likert scales in NLG evaluation tasks emerge:

1. Many studies confuse Likert scales and Likert items. Often Likert scales are used for an item-by-item analysis.

2. Scales are often analysed with parametric statistics without a justification.[6]

Regarding 1: Aggregate scales such as Likert scales are created to estimate the overall opinion of a responder about some phenomenon by the use of aggregate items. Indeed, the design of a Likert scale is aimed at reaching an overall opinion by analysing together the answers given by the responder about the single items. Accordingly, items extracted from an aggregate scale reveal one aspect of the phenomenon and can lose meaning if analysed in isolation from the other items. Also, the use of parametric statistics for Likert scales can be better justified in the case of item aggregation. It is difficult to justify the assumption of equal distance between the scale points across different responders when doing an item-by-item analysis. If researchers are interested in performing parametric statistics using Likert items, or better graphic rating scales, we refer to Harpe (2015) for some recommendations. It is important to decide the scale as part of the experimental design and not at the time of analysis[7]. In case one Likert scale is used, because the items are considered as pieces of a bigger picture, it is important to check their internal consistency. To this end Cronbach's α[8], Revelle's β, McDonald's ω_h, ω_{Total} or Kuder-Richardson 20 can be used. A review of different measures of internal consistency can be found in Revelle and Zinbarg (2009) and McNeish (2018).

Finally, it is important to use appropriate language to avoid confusion and allow the readers to form a better understanding of the results. For example, one should avoid using the term Likert scale to refer to a graphic rating scale or Likert items, especially if Likert items are analysed in isolation.

Regarding 2: Although there is no right way to interpret rating and aggregate scales, such as Likert scales, it is good practice to justify the scale interpretation and the choice of the statistic used in their analysis. As proved by previous studies, see for example Norman (2010), the use of parametric statistics is quite robust with ordinal data. Generally, is not clear whether authors are aware of the controversy about scale interpretations, and many do not provide an argument for using one interpretation rather than another.

Due to the fact that using parametric statistics for ordinal data can lead to unwanted conse-

[6] We note that the use of parametric statistics without a justification is also present for evaluation methodologies other than rating and Likert scales (for instance ranking experiments). This is in general true also for nonparametric statistics. Although nonparametric methods do not require assumptions about the distribution of the population probability, they do require assumptions such as randomness and independence of the samples. This suggests that in general researchers have to pay more attention to the statistics use in their evaluation studies.

[7] Decisions about the levels of measurement and the choice of analysis method should be made at the design stage (for example, whether to use ordinal or interval scale, or by using descriptive or inferential statistics). This way, the researchers can create a survey compatible with the chosen methodology.

[8] Although the use of Cronbach's α was recently criticized in McNeish (2018).

quences, sometimes substantial and sometimes inconsequential, the use of parametric statistics on data which are not interval should be clearly justified. Liddell and Kruschke (2018) present several problematic cases where parametric statistic were used for ordinal data. Liddell and Kruschke, for instance, discuss examples with low correct detection rates, risk of inflated Type I and Type II error rates and distorted effect size estimates. The first problem reduces statistical power. The second can result in a false positive conclusion or false negative conclusion.[9] Finally, the last problem can lead to either overestimating or underestimating the size of the difference between two groups.

Without a preliminary verification of the parametric statistic assumptions, the use of such a statistic is controversial. Although parametric statistics allow more powerful and nuanced data analyses than nonparametric statistics, sometimes the use of nonparametric statistics is enough. If the data collected fails to satisfy the conditions required from the parametric statistic, it does not mean that the data lost statistical importance. Indeed, the use of percentages or central tendency measures such as mode or median as well as statistical significance such as Mann-Whitney U Test, Kruskal-Wallis, χ^2 etc., can give a good picture of the generative abilities of a NLG system. Furthermore, recent advances in statistics have introduced new options for ordinal data which are worth to be taken into account, for example ordinal regression models (Bürkner and Vuorre, 2019) or generalized mixed effect models (Faraway, 2016) which are able to work with several different data distributions.

To our knowledge, there is currently a lack of robustness studies in the NLG area. Such studies would be greatly valuable for the discussion of the use of parametric and nonparametric statistics as well as the use of ordinal regression models and generalized mixed effect models.

Acknowledgments

We warmly thanks the anonymous reviewers for their helpful suggestions.

[9] Type I error is the rejection of the null hypothesis when it is true. Type II error is the failure to reject the null hypothesis when it is false.

References

P. C. Bürkner and M. Vuorre. 2019. Ordinal regression models in psychology: a tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101.

L. Carifio, R. Perla, and P. Giraux. 2008. Resolving the 50-year debate around using and misusing likert scales. *Med Educ.*, 42(12):1150–1152.

J. CF De Winter and D. Dodou. 2010. Five-point likert items: t test versus mann-whitney-wilcoxon. *Practical Assessment, Research & Evaluation*, 15(11):1–12.

J. J. Faraway. 2016. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC, Boca Raton, FL.

S. E. Harpe. 2015. How to analyze likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, 7(6):836–850.

S. Jamieson. 2004. Likert scales: how to (ab)use them. *Med Educ.*, 38(12):1217–1218.

R. L. Johnson and G. B. Morgan. 2016. *Survey Scales. A Guide to Development, Analysis, and Reporting*. The Guilford Press, New York, NY.

A. Joshi, S. Kale, S. Chandel, and D. K. Pal. 2015. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4):396–403.

T. R. Knapp. 1990. Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nurs Res.*, 39:121–123.

W. Kuzon, M. Urbanchek, and S. McCabe. 1996. The seven deadly sins of statistical analysis. *Annals of plastic surgery*, 37:265–272.

T. M. Liddell and J. K. Kruschke. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79:328–348.

R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22:1–55.

D. McNeish. 2018. Thanks coefficient alpha, well take it from here. *Psychological Methods*, 23(3):412.

G. Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Adv Health Sci Educ Theory Pract.*, 15(10):625–632.

G. Pell. 2005. Uses and misuses of likert scales. *Med Educ.*, 39(9):970.

W. Revelle and R. E. Zinbarg. 2009. Coefficients alpha, beta, omega, and the glb: Comments on sijtsma. *Psychometrika*, 74(1):145.

J. Robertson. 2012. Likert-type scales, statistical methods, and effect sizes. *Commun. ACM*, 55(5):6–7.

G. M. Sullivan and A. R. Artino. 2013. Analyzing and interpreting data from likert-type scales. *Journal of Graduate Medical Education*, pages 541–542.

R. S. Witte and J. S. Witte. 2017. *Statistics, Eleventh Edition*. John Wiley & Sons, Inc., LaVergne, Tennessee, USA.

A complete list of the papers we examined can be found via the following link https://bit.ly/2lKL516.