

Discrete vs. Continuous Rating Scales for Language Evaluation in NLP

Anja Belz

Eric Kow

School of Computing, Engineering and Mathematics
University of Brighton
Brighton BN2 4GJ, UK
{A.S.Belz, E.Y.Kow}@brighton.ac.uk

Abstract

Studies assessing rating scales are very common in psychology and related fields, but are rare in NLP. In this paper we assess discrete and continuous scales used for measuring quality assessments of computer-generated language. We conducted six separate experiments designed to investigate the validity, reliability, stability, interchangeability and sensitivity of discrete vs. continuous scales. We show that continuous scales are viable for use in language evaluation, and offer distinct advantages over discrete scales.

1 Background and Introduction

Rating scales have been used for measuring human perception of various stimuli for a long time, at least since the early 20th century (Freyd, 1923). First used in psychology and psychophysics, they are now also common in a variety of other disciplines, including NLP. Discrete scales are the only type of scale commonly used for qualitative assessments of computer-generated language in NLP (e.g. in the DUC/TAC evaluation competitions). Continuous scales are commonly used in psychology and related fields, but are virtually unknown in NLP.

While studies assessing the quality of individual scales and comparing different types of rating scales are common in psychology and related fields, such studies hardly exist in NLP, and so at present little is known about whether discrete scales are a suitable rating tool for NLP evaluation tasks, or whether continuous scales might provide a better alternative.

A range of studies from sociology, psychophysiology, biometrics and other fields have compared

discrete and continuous scales. Results tend to differ for different types of data. E.g., results from pain measurement show a continuous scale to outperform a discrete scale (ten Klooster et al., 2006). Other results (Svensson, 2000) from measuring students' ease of following lectures show a discrete scale to outperform a continuous scale. When measuring dyspnea, Lansing et al. (2003) found a hybrid scale to perform on a par with a discrete scale.

Another consideration is the types of data produced by discrete and continuous scales. Parametric methods of statistical analysis, which are far more sensitive than non-parametric ones, are commonly applied to both discrete and continuous data. However, parametric methods make very strong assumptions about data, including that it is numerical and normally distributed (Siegel, 1957). If these assumptions are violated, then the significance of results is overestimated. Clearly, the numerical assumption does not hold for the categorical data produced by discrete scales, and it is unlikely to be normally distributed. Many researchers are happier to apply parametric methods to data from continuous scales, and some simply take it as read that such data is normally distributed (Lansing et al., 2003).

Our aim in the present study was to systematically assess and compare discrete and continuous scales when used for the qualitative assessment of computer-generated language. We start with an overview of assessment scale types (Section 2). We describe the experiments we conducted (Section 4), the data we used in them (Section 3), and the properties we examined in our inter-scale comparisons (Section 5), before presenting our results

Q1: Grammaticality The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.

1. Very Poor
2. Poor
3. Barely Acceptable
4. Good
5. Very Good

Figure 1: Evaluation of Readability in DUC’06, comprising 5 evaluation criteria, including Grammaticality. Evaluation task for each summary text: evaluator selects one of the options (1–5) to represent quality of the summary in terms of the criterion.

(Section 6), and some conclusions (Section 7).

2 Rating Scales

With **Verbal Descriptor Scales** (VDSS), participants give responses on ordered lists of verbally described and/or numerically labelled response categories, typically varying in number from 2 to 11 (Svensson, 2000). An example of a VDS used in NLP is shown in Figure 1. VDSS are used very widely in contexts where computationally generated language is evaluated, including in dialogue, summarisation, MT and data-to-text generation.

Visual analogue scales (VASS) are far less common outside psychology and related areas than VDSS. Responses are given by selecting a point on a typically horizontal line (although vertical lines have also been used (Scott and Huskisson, 2003)), on which the two end points represent the extreme values of the variable to be measured. Such lines can be mono-polar or bi-polar, and the end points are labelled with an image (smiling/frowning face), or a brief verbal descriptor, to indicate which end of the line corresponds to which extreme of the variable. The labels are commonly chosen to represent a point beyond any response actually likely to be chosen by raters. There is only one examples of a VAS in NLP system evaluation that we are aware of (Gatt et al., 2009).

Hybrid scales, known as a graphic rating scales, combine the features of VDSs and VASSs, and are also used in psychology. Here, the verbal descriptors are aligned along the line of a VAS and the endpoints are typically unmarked (Svensson, 2000). We are aware of one example in NLP (Williams and Reiter, 2008);

Q1: Grammaticality The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.


extremely
bad  excellent

Figure 2: Evaluation of Grammaticality with alternative VAS scale (cf. Figure 1). Evaluation task for each summary text: evaluator selects a place on the line to represent quality of the summary in terms of the criterion.

we did not investigate this scale in our study.

We used the following two specific scale designs in our experiments:

VDS-7: 7 response categories, numbered (7 = best) and verbally described (e.g. 7 = “perfectly fluent” for Fluency, and 7 = “perfectly clear” for Clarity). Response categories were presented in a vertical list, with the best category at the bottom. Each category had a tick-box placed next to it; the rater’s task was to tick the box by their chosen rating.

VAS: a horizontal, bi-polar line, with no ticks on it, mapping to 0–100. In the image description tests, statements identified the left end as negative, the right end as positive; in the weather forecast tests, the positive end had a smiling face and the label “statement couldn’t be clearer/read better”; the negative end had a frowning face and the label “statement couldn’t be more unclear/read worse”. The raters’ task was to move a pointer (initially in the middle of the line) to the place corresponding to their rating.

3 Data

Weather forecast texts: In one half of our evaluation experiments we used human-written and automatically generated weather forecasts for the same weather data. The data in our evaluations was for 22 different forecast dates and included outputs from 10 generator systems and one set of human forecasts. This data has also been used for comparative system evaluation in previous research (Langner, 2010; Angeli et al., 2010; Belz and Kow, 2009). The following are examples of weather forecast texts from the data:

- 1: SSE 28-32 INCREASING 36-40 BY MID AFTERNOON
- 2: S’LY 26-32 BACKING SSE 30-35 BY AFTER-

NOON INCREASING 35-40 GUSTS 50 BY MID
EVENING

Image descriptions: In the other half of our evaluations, we used human-written and automatically generated image descriptions for the same images. The data in our evaluations was for 112 different image sets and included outputs from 6 generator systems and 2 sets of human-authored descriptions. This data was originally created in the TUNA Project (van Deemter et al., 2006). The following is an example of an item from the corpus, consisting of a set of images and a description for the entity in the red frame:



the small blue fan

4 Experimental Set-up

4.1 Evaluation criteria

Fluency/Readability: Both the weather forecast and image description evaluation experiments used a quality criterion intended to capture ‘how well a piece of text reads’, called Fluency in the latter, Readability in the former.

Adequacy/Clarity: In the image description experiments, the second quality criterion was Adequacy, explained as “how clear the description is”, and “how easy it would be to identify the image from the description”. This criterion was called Clarity in the weather forecast experiments, explained as “how easy is it to understand what is being described”.

4.2 Raters

In the image experiments we used 8 raters (native speakers) in each experiment, from cohorts of 3rd-year undergraduate and postgraduate students doing a degree in a linguistics-related subject. They were paid and spent about 1 hour doing the experiment.

In the weather forecast experiments, we used 22 raters in each experiment, from among academic staff at our own university. They were not paid and spent about 15 minutes doing the experiment.

4.3 Summary overview of experiments

Weather VDS-7 (A): VDS-7 scale; weather forecast data; criteria: Readability and Clarity; 22 raters (university staff) each assessing 22 forecasts.

Weather VDS-7 (B): exact repeat of Weather VDS-7 (A), including same raters.

Weather VAS: VAS scale; 22 raters (university staff), no overlap with raters in Weather VDS-7 experiments; other details same as in Weather VDS-7.

Image VDS-7: VDS-7 scale; image description data; 8 raters (linguistics students) each rating 112 descriptions; criteria: Fluency and Adequacy.

Image VAS (A): VAS scale; 8 raters (linguistics students), no overlap with raters in Image VAS-7; other details same as in Image VDS-7 experiment.

Image VAS (B): exact repeat of Image VAS (A), including same raters.

4.4 Design features common to all experiments

In all our experiments we used a Repeated Latin Squares design to ensure that each rater sees the same number of outputs from each system and for each text type (forecast date/image set). Following detailed instructions, raters first did a small number of practice examples, followed by the texts to be rated, in an order randomised for each rater. Evaluations were carried out via a web interface. They were allowed to interrupt the experiment, and in the case of the 1 hour long image description evaluation they were encouraged to take breaks.

5 Comparison and Assessment of Scales

Validity is to the extent to which an assessment method measures what it is intended to measure (Svensson, 2000). Validity is often impossible to assess objectively, as is the case of all our criteria except Adequacy, the validity of which we can directly test by looking at correlations with the accuracy with which participants in a separate experiment identify the intended images given their descriptions.

A standard method for assessing **Reliability** is Kendall’s W, a coefficient of concordance, measuring the degree to which different raters agree in their ratings. We report W for all 6 experiments.

Stability refers to the extent to which the results of an experiment run on one occasion agree with the results of the same experiment (with the same

raters) run on a different occasion. In the present study, we assess stability in an intra-rater, test-retest design, assessing the agreement between the same participant's responses in the first and second runs of the test with Pearson's product-moment correlation coefficient. We report these measures between ratings given in Image VAS (A) vs. those given in Image VAS (B), and between ratings given in Weather VDS-7 (A) vs. those given in Weather VDS-7 (B).

We assess **Interchangeability**, that is, the extent to which our VDS and VAS scales agree, by computing Pearson's and Spearman's coefficients between results. We report these measures for all pairs of weather forecast/image description evaluations.

We assess the **Sensitivity** of our scales by determining the number of significant differences between different systems and human authors detected by each scale.

We also look at the relative effect of the different experimental factors by computing the F-Ratio for System (the main factor under investigation, so its relative effect should be high), Rater and Text Type (their effect should be low). F-ratios were determined by a one-way ANOVA with the evaluation criterion in question as the dependent variable and System, Rater or Text Type as grouping factors.

6 Results

6.1 Interchangeability and Reliability for system/human authored image descriptions

Interchangeability: Pearson's r between the means per system/human in the three image description evaluation experiments were as follows (Spearman's ρ shown in brackets):

		VAS (A)	VAS (B)
		VDS-7	.957**(.958**)
Flue.	VAS (A)	—	.874** (.810*)
	VDS-7	.948**(.922**)	.864** (.850**)
	VAS (A)	—	.937** (.929**)

For both Adequacy and Fluency, correlations between Image VDS-7 and Image VAS (A) (the main VAS experiment) are extremely high, meaning that they could substitute for each other here.

Reliability: Inter-rater agreement in terms of Kendall's W in each of the experiments:

	VDS-7	VAS (A)	VAS (B)
K's W Adequacy	.598**	.471**	.595*
K's W Fluency	.640**	.676**	.729**

W was higher in the VAS data in the case of Fluency, whereas for Adequacy, W was the same for the VDS data and VAS (B), and higher in the VDS data than in the VAS (A) data.

6.2 Interchangeability and Reliability for system/human authored weather forecasts

Interchangeability: The correlation coefficients (Pearson's r with Spearman's ρ in brackets) between the means per system/human in the image description experiments were as follows:

		VDS-7 (B)	VAS
		VDS-7 (A)	.995** (.989**)
Clar.	VDS-7 (B)	—	.939** (.836**)
	VDS-7 (A)	.981** (.870**)	.947** (.709*)
	VDS-7 (B)	—	.951** (.656*)

For both Adequacy and Fluency, correlations between Weather VDS-7 (A) (the main VDS-7 experiment) and Weather VAS (A) are again very high, although rank-correlation is somewhat lower.

Reliability: Inter-rater agreement in terms of Kendall's W was as follows:

	VDS-7 (A)	VDS-7 (B)	VAS
W Clarity	.497**	.453**	.485**
W Read.	.533**	.488**	.480**

This time the highest agreement for both Clarity and Readability was in the VDS-7 data.

6.3 Stability tests for image and weather data

Pearson's r between ratings given by the same raters first in Image VAS (A) and then in Image VAS (B) was .666 for Adequacy, .593 for Fluency. Between ratings given by the same raters first in Weather VDS-7 (A) and then in Weather VDS-7 (B), Pearson's r was .656 for Clarity, .704 for Readability. (All significant at $p < .01$.) Note that these are computed on individual scores (rather than means as in the correlation figures given in previous sections).

6.4 F-ratios and post-hoc analysis for image data

The table below shows F-ratios determined by a one-way ANOVA with the evaluation criterion in question (Adequacy/Fluency) as the dependent variable and System/Rater/Text Type as the grouping factor. Note

that for System a high F-ratio is desirable, but a low F-ratio is desirable for other factors.

Image descriptions			
		VDS-7	VAS (A)
Adequacy	System	8.822**	6.371**
	Rater	12.623**	13.136**
	Text Type	1.193	1.519**
Fluency	System	13.312**	17.207**
	Rater	27.401**	17.479**
	Text Type	.894	1.091

Out of a possible 28 significant differences for System, the main factor under investigation, VDS-7 found 8 for Adequacy and 14 for Fluency; VAS (A) found 7 for Adequacy and 15 for Fluency.

6.5 F-ratios and post-hoc analysis for weather data

The table below shows F-ratios analogous to the previous section (for Clarity/Readability).

Weather forecasts			
		VDS-7 (A)	VAS
Clarity	System	23.507**	23.468**
	Rater	4.832**	6.857**
	Text Type	1.467	1.632*
Read.	System	24.351**	22.538**
	Rater	4.824**	5.560**
	Text Type	1.961**	2.906**

Out of a possible 55 significant differences for System, VDS-7 (A) found 24 for Clarity, 23 for Readability; VAS found 25 for Adequacy, 26 for Fluency.

6.6 Scale validity test for image data

Our final table of results shows Pearson's correlation coefficients (calculated on means per system) between the Adequacy data from the three image description evaluation experiments on the one hand, and the data from an extrinsic experiment in which we measured the accuracy with which participants identified the intended image described by a description:

	ID Acc.
Image VAS (A) Adequacy	.870**
Image VAS (B) Adequacy	.927**
Image VDS-7 Adequacy	.906**

The correlation between Adequacy and ID Accuracy was strong and highly significant in all three image description evaluation experiments, but strongest in VAS (B), and weakest in VAS (A). For comparison,

Pearson's between Fluency and ID Accuracy ranged between .3 and .5, whereas Pearson's between Adequacy and ID Speed (also measured in the same image identification experiment) ranged between -.35 and -.29.

7 Discussion and Conclusions

Our interchangeability results (Sections 6.1 and 6.2) indicate that the VAS and VDS-7 scales we have tested can substitute for each other in our present evaluation tasks in terms of the mean system scores they produce. Where we were able to measure validity (Section 6.6), both scales were shown to be similarly valid, predicting image identification accuracy figures from a separate experiment equally well. Stability (Section 6.3) was marginally better for VDS-7 data, and Reliability (Sections 6.1 and 6.2) was better for VAS data in the image description evaluations, but (mostly) better for VDS-7 data in the weather forecast evaluations. Finally, the VAS experiments found greater numbers of statistically significant differences between systems in 3 out of 4 cases (Section 6.5).

Our own raters strongly prefer working with VAS scales over VDSs. This has also long been clear from the psychology literature (Svensson, 2000), where raters are typically found to prefer VAS scales over VDSs which can be a "constant source of vexation to the conscientious rater when he finds his judgments falling between the defined points" (Champney, 1941). Moreover, if a rater's judgment falls between two points on a VDS then they must make the false choice between the two points just above and just below their actual judgment. In this case we know that the point they end up selecting is not an accurate measure of their judgment but rather just one of two equally accurate ones (one of which goes unrecorded).

Our results establish (for our evaluation tasks) that VAS scales, so far unproven for use in NLP, are at least as good as VDSs, currently virtually the only scale in use in NLP. Combined with the fact that raters strongly prefer VASs and that they are regarded as more amenable to parametric means of statistical analysis, this indicates that VAS scales should be used more widely for NLP evaluation tasks.

References

- Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *Proceedings of the 15th Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*.
- Anja Belz and Eric Kow. 2009. System building cost vs. output quality in data-to-text generation. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 16–24.
- H. Champney. 1941. The measurement of parent behavior. *Child Development*, 12(2):131.
- M. Freyd. 1923. The graphic rating scale. *Biometrical Journal*, 42:83–102.
- A. Gatt, A. Belz, and E. Kow. 2009. The TUNA Challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG'09)*, pages 198–206.
- Brian Langner. 2010. *Data-driven Natural Language Generation: Making Machines Talk Like Humans Using Natural Corpora*. Ph.D. thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University.
- Robert W. Lansing, Shakeeb H. Moosavi, and Robert B. Banzett. 2003. Measurement of dyspnea: word labeled visual analog scale vs. verbal ordinal scale. *Respiratory Physiology & Neurobiology*, 134(2):77–83.
- J. Scott and E. C. Huskisson. 2003. Vertical or horizontal visual analogue scales. *Annals of the rheumatic diseases*, (38):560.
- Sidney Siegel. 1957. Non-parametric statistics. *The American Statistician*, 11(3):13–19.
- Elisabeth Svensson. 2000. Comparison of the quality of assessments using continuous and discrete ordinal rating scales. *Biometrical Journal*, 42(4):417–434.
- P. M. ten Klooster, A. P. Klaar, E. Taal, R. E. Gheith, J. J. Rasker, A. K. El-Garf, and M. A. van de Laar. 2006. The validity and reliability of the graphic rating scale and verbal rating scale for measuring pain across cultures: A study in egyptian and dutch women with rheumatoid arthritis. *The Clinical Journal of Pain*, 22(9):827–30.
- Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Conference on Natural Language Generation*, pages 130–132, Sydney, Australia, July.
- S. Williams and E. Reiter. 2008. Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, 14(4):495–525.