Comparing Rating Scales and Preference Judgements in Language Evaluation

Anja Belz **Eric Kow**
Natural Language Technology Group
School of Computing, Mathematical and Information Sciences
University of Brighton
Brighton BN2 4GJ, UK
{asb,eykk10}@bton.ac.uk

Abstract

Rating-scale evaluations are common in NLP, but are problematic for a range of reasons, e.g. they can be unintuitive for evaluators, inter-evaluator agreement and self-consistency tend to be low, and the parametric statistics commonly applied to the results are not generally considered appropriate for ordinal data. In this paper, we compare rating scales with an alternative evaluation paradigm, preference-strength judgement experiments (PJEs), where evaluators have the simpler task of deciding which of two texts is better in terms of a given quality criterion. We present three pairs of evaluation experiments assessing text fluency and clarity for different data sets, where one of each pair of experiments is a rating-scale experiment, and the other is a PJE. We find the PJE versions of the experiments have better evaluator self-consistency and inter-evaluator agreement, and a larger proportion of variation accounted for by system differences, resulting in a larger number of significant differences being found.

1 Introduction

Rating-scale evaluations, where human evaluators assess system outputs by selecting a score on a discrete scale, are the most common form of human-assessed evaluation in NLP. Results are typically presented in rank tables of means for each system accompanied by means-based measures of statistical significance of the differences between system scores.

NLP system evaluation tends to involve sets of systems, rather than single ones (evaluations tend to at least incorporate a baseline or, more rarely, a topline system). The aim of system evaluation is to gain some insight into which systems are better than which others, in other words, the aim is inherently relative. Yet NLP system evaluation experiments have generally preferred rating scale experiments where evaluators assess each system's quality in isolation, in absolute terms.

Such rating scales are not very intuitive to use; deciding whether a text deserves a 5, a 4 or a 3 etc. can be difficult. Furthermore, evaluators may ascribe different meanings to scores and the distances between them. Individual evaluators have different tendencies in using rating scales, e.g. what is known as 'end-aversion' tendency where certain individuals tend to stay away from the extreme ends of scales; other examples are positive skew and acquiescence bias, where individuals make disproportionately many positive or agreeing judgements; see e.g. Choi and Pak, (2005).

It is not surprising then that stable averages of quality judgements, let alone high levels of agreement, are hard to achieve, as has been observed for MT (Turian et al., 2003; Lin and Och, 2004), text summarisation (Trang Dang, 2006), and language generation (Belz and Reiter, 2006). It has even been demonstrated that increasing the number of evaluators and/or data can have no stabilising effect at all on means (DUC literature).

The result of a rating scale experiment is ordinal data (sets of scores selected from the discrete rating scale). The means-based ranks and statistical significance tests that are commonly presented with the results of RSEs are not generally considered appropriate for ordinal data in the statistics literature (Siegel, 1957). At the least, "a test on the means imposes the requirement that the measures must be additive, i.e. numerical" (Siegel, 1957, p. 14). Parametric statistics are more powerful than non-parametric alternatives, because they make a number of strong assumptions (including that the data is numerical). If the assumptions are violated then the risks is that the significance of results is

overestimated.

In this paper we explore an alternative evaluation paradigm, Preference-strength Judgement Experiments (PJEs). Binary preference judgements have been used in NLP system evaluation (Reiter et al., 2005), but to our knowledge this is the first systematic investigation of preference-strength judgements where evaluators express, in addition to their preference (*which system do you prefer?*), also the strength of their preference (*how strongly do you prefer the system you prefer?*). It seems intuitively convincing that it should be easier to decide which of two texts is clearer than to decide whether a text's clarity deserves a 1, 2, 3, 4 or 5. However, it is less clear whether evaluators are also able to express the strength of their preference in a consistent fashion, resulting not only in good self-consistency, but also in good agreement with other evaluators.

We present three pairs of directly comparable RSE and PJE evaluations, and investigate how they compare in terms of (i) the amount of variation accounted for by differences between systems (the more the better), relative to the amount of variation accounted for by other factors such as evaluator and arbitrary text properties (the less the better); (ii) inter-evaluator agreement, (iii) evaluator self-consistency, (iv) the number of significant differences identified, and (v) experimental cost.

2 Overview of Experiments

In the following three sections we present the design and results of three pairs of evaluations. Each pair consists of a rating-scale experiment (RSE) and a preference-strength judgement experiment (PJE) that differ only in the rating method they employ (relative ratings in the PJE and absolute ratings in the RSE).[1] In other words, they involve the same set of system outputs, the same instructions and method of presentating system outputs. Each pair is for a different data domain and system task, the first for generating chains of references to people in Wikipedia articles (Section 3); the second for weather forecast text generation (Section 4); and the third for generating descriptions of images of furniture and faces (Section 5).

All experiments use a Repeated Latin Squares design which ensures that each subject sees the same number of outputs from each system and for each test set item. Following detailed instructions, subjects first do 2 or 3 practice examples, followed by the texts to be evaluated, in an order randomised for each subject. Subjects carry out the evaluation over the internet, at a time and place of their choosing. They are allowed to interrupt and resume (but are discouraged from doing so).

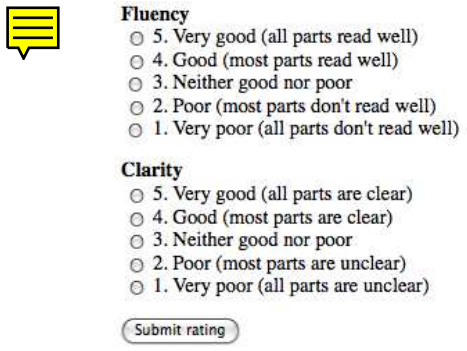


Figure 1: Standardised 1–5 rating scale representation for Fluency and Clarity criteria.

There are subtle differences between the three experiment pairs, and for ease of comparison we provide an overview of the six experiments we investigate in this paper in Table 1. Each of the aspects of experimental design and execution shown in this table is explained and described in more detail in the relevant subsection below, but some of the important differences are highlighted here.

In GREC-NEG PJE, each system is compared with only one other comparisor system (a human-authored topline), whereas in the other two PJE experiments, each system is compared with all other systems for each test data set item.

In the two versions of the METEO evaluation, evaluators were not drawn from the same cohort of people, whereas in the other two evaluation pairs they were drawn from the same cohort. GREC-NEG RSE and METEO RSE used radio buttons (as shown in Figure 1) as the rating-scale evaluation mechanism whereas in TUNA RSE it was an unmarked slider bar. While slightly different names were used for the evaluation criteria in two of the evaluation pairs, Fluency/Readability were explained in very similar terms (*does it read well?*), and Adequacy in TUNA was explained in terms of clarity of reference (*is it clear which entity the description refers to?*), so there are in fact just two evaluation criteria (albeit with different names).

Where we use preference-strength judgements,

[1] We are currently preparing an open-source release of the RSE/PJE toolkit we have developed for implementing the experiments described in this paper which automatically generates an experiment, including webpages, given some user-specified parameters and the data to be evaluated.

Data set	GREC-NEG		METEO		TUNA	
Type	RSE	PJE	RSE	PJE	RSE	PJE
Criteria names	Fluency, Clarity		Readability, Clarity		Fluency, Adequacy	
Evaluator type	linguistics students		uni staff	ling stud	linguistics students	
Num evaluators	10	10	22	22	8	28
Comparisor(s)	–	human topline	–	all systems	–	all systems
Test set size	30		22		112	
N trials	300	300	484	1210	896	3136
Rating tool	radio buttons	slider	radio buttons	slider	slider bar	slider
Range	1–5	$-10.0.. + 10.0$	1–7	$-50.0.. + 50.0$	0–100	$-50.0.. + 50.0$
Numbers visible?	yes	no	yes	no	no	no

Table 1: Overview of experiments with details of design and execution. (Comparisor(s) = the other systems against which each system is evaluated.)

the evaluation mechanism is implemented using slider bars as shown at the bottom of Figure 2 which map to a scale $-X.. + X$. The evaluator's task is to express their preference in terms of each quality criterion by moving the pointers on the sliders. Moving the pointer to the left means expressing a preference for the text on the left, moving it to the right means preferring the text on the right; the further to the left/right the slider is moved, the stronger the preference. It was not evident to the evaluators that sliders were associated with numerical values. Slider pointers started out in the middle of the scale (the position corresponding to no preference). If they wanted to leave the pointer in the middle (i.e. if they had no preference for either of the two texts), evaluators had to check a box to confirm their rating (to avoid evaluators accidentally not rating a text and leaving the pointer in the default position).

3 GREC-NEG RSE/PJE: Named entity reference chains

3.1 Data and generic design

In our first pair of experiments we used system and human outputs for the GREC-NEG task of selecting referring expressions for people in discourse context. The GREC-NEG data[2] consists of introduction sections from Wikipedia articles about people in which all mentions of people have been annotated by marking up the word strings that function as referential expressions (REs) and annotating them with coreference information as well as syntactic and semantic features. The following is an example of an annotated RE from the corpus:

```
<REF ENTITY="0" MENTION="1" SEMCAT="person" SYNCAT="np"
  SYNFUNC="subj"><REFEX ENTITY="0" REG08-TYPE="name"
  CASE="plain">Sir Alexander Fleming</REFEX> </REF>
(6 August 1881 - 11 March 1955) was a Scottish biol-
ogist and pharmacologist.
```

This data was used in the GREC-NEG'09 shared-task competition (Belz et al., 2009), where the task was to create systems which automatically select suitable REs for all references to all person entities in a text.

The evaluation experiments use Clarity and Fluency as quality criteria which were explained in the introduction as follows (the wording of the first is from DUC):

1. **Referential Clarity**: It should be easy to identify who the referring expressions are referring to. If a person is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if a person is referenced, but their identity or relation to the story remains unclear.

2. **Fluency**: A referring expression should 'read well', i.e. it should be written in good, clear English, and the use of titles and names should seem natural. Note that the Fluency criterion is independent of the Referential Clarity criterion: a reference can be perfectly clear, yet not be fluent.

The evaluations involved outputs for 30 randomly selected items from the test set from 5 of the 6 systems which participated in GREC-NEG'10, the four baseline systems developed by the organisers, and the original corpus texts (10 systems in total).

3.2 Preference judgement experiment

The human-assessed intrinsic evaluation in GREC'09 was designed as a preference-judgement test where subjects expressed their preference, in terms of the two criteria, for either the original Wikipedia text (human-authored 'topline') or the version of it with system-selected referring expressions in it. There were three 10x10 Latin Squares, and a total of 300 trials (with two judgements in each, one for Fluency and one for Clarity) in this evaluation. The subjects were 10

[2] The GREC-NEG data and documentation is available for download from http://www.nltg.brighton.ac.uk/home/Anja.Belz

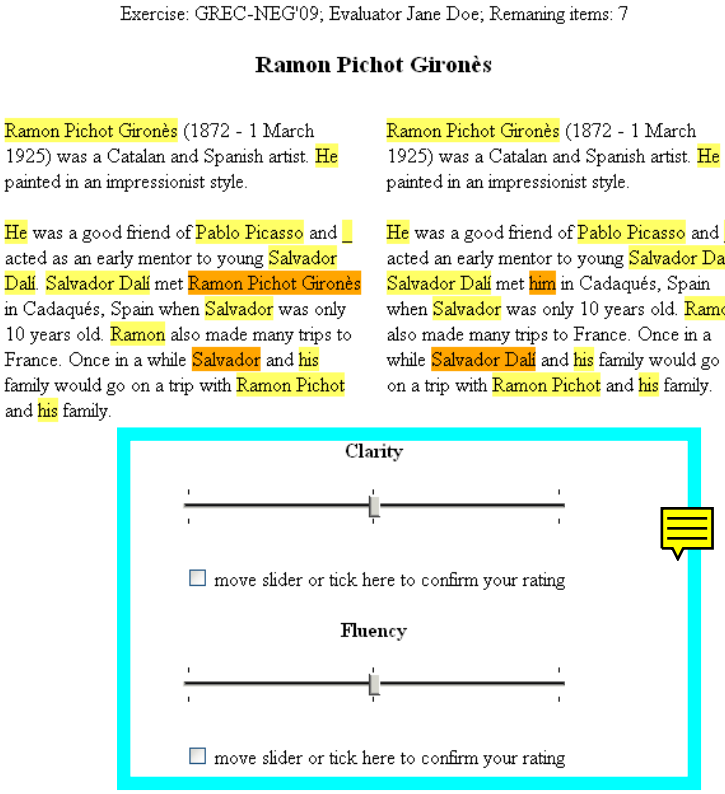


Figure 2: Example of text pair presented in human intrinsic evaluation of GREC-NEG systems.

native speakers of English recruited from cohorts of students currently completing a linguistics-related degree at Kings College London and University College London.

Figure 2 shows what subjects saw during the evaluation of an individual text pair. The place (left/right) of the original Wikipedia article was randomly determined for each individual evaluation of a text pair. People references are highlighted in yellow/orange, those that are identical in both texts are yellow, those that are different are orange.[3] The sliders are the standardised design described in the preceding section.

3.3 Rating scale experiment

Our new experiment used our standardised radio button design for a 1–5 rating scale as shown in Figure 1. We used the same Latin Squares design as for the PJE version, and recruited 10 different evaluators from the same student cohorts at Kings College London and University College London. Evaluators saw just one text in each trial, with the people references highlighted in yellow.

3.4 Results and comparative analysis

Measures comparing the results from the two versions of the GREC-NEG evaluation are shown in Table 2. The first row for each experiment type shows the F ratio as determined by a one-way ANOVA with the evaluation criterion in question as the dependent variable and System as the factor. F is the ratio of between-groups variability over within-group (or residual) variability, i.e. the larger the value of F, the more of the variability observed in the data is accounted for by the grouping factor, here System, relative to what variability remains within the groups.

The second row shows the number of significant differences out of the possible total, as determined by a Tukey's HSD analysis. Kendall's W (interpretable as a coefficient of concordance) is

Type	Measure	Clarity	Fluency
RSE	$F_{(9,290)}$	10.975**	35.998**
	N sig diffs	19/45	27/45
	K's W (inter)	.543**	.760**
	avg W (intra)	.5275	.7192
(Text	$F_{(29,270)}$	2.512**	1.825**)
(Evaluator	$F_{(9,290)}$	3.998**	.630)
PJE	$F_{(9,290)}$	29.539**	26.596**
	N sig diffs	26/45	24/45
	K's W (inter)	.717**	.725**
	avg W (intra)	.6909	.7125
(Text	$F_{(29,270)}$	.910	1.237)
(Evaluator	$F_{(9,290)}$	1.237	4.145**)

Table 2: GREC-NEG RSE/PJE: Results of analyses looking at effect of System.

[3] When viewed in black and white, the orange highlights are the slighly darker ones.

a commonly used measure of the agreement between judges and is based on mean rank. It ranges from 0 to 1, and the closer to 1 it is the greater the agreement. The fourth row (K's W, inter) shows the standard W measure, estimating the degree to which the evaluators agreed. The 5th row (K's W, intra) shows the average W for repeated ratings *by the same judge*, i.e. it is a measure of the average self-consistency achieved by the evaluators. Finally, in the last two rows we give F-ratios for Text (test data set item) and Evaluator, estimating the effect these two have independently of System.

The F ratios and numbers of significant differences are very similar in the PJE version, but very dissimilar in the RSE version of this experiment. For Fluency, F is greater in the RSE version than in the PJE version where there appear to be bigger differences between scores assigned by evaluators. However, Kendall's W shows that in terms of mean score ranks, the evaluators agreed to a similar extent in both experiment versions.

Clarity in the RSE version has lower values across the board than the rest of Table 2: it accounts for less of the variation, has fewer significant differences and lower levels of inter-evaluator agreement and self-consistency. If the results from the PJE version were not also available one might be inclined to conclude that there was not as much difference between systems in terms of Clarity as there was in terms of Fluency. However, because Fluency and Clarity have a similarly strong effect in GREC-NEG PJE, it looks instead as though the evaluators found it harder to apply the Clarity criterion in GREC-NEG RSE than Fluency in GREC-NEG RSE, and than Clarity in GREC-NEG PJE.

One way of interpreting this is that it is possible to achieve the same good levels of inter-evaluator and intra-evaluator variation for the Clarity criterion as for Fluency (both as defined and applied within the context of this specific experiment), and that it is therefore worrying that the RSE version does not achieve it.

4 METEO RSE/PJE: Weather forecasts

4.1 Data

Our second pair of evaluations used the Prodigy-METEO[4] version (Belz, 2009) of the SUMTIME-METEO corpus (Sripada et al., 2002) which contains system outputs and the pairs of wind forecast texts and wind data the systems were trained on, e.g.:

Data: `1 SSW 16 20 - - 0600 2 SSE - - - - NOTIME 3 VAR 04 08 - - 2400`

Text: `SSW 16-20 GRADUALLY BACKING SSE THEN FALLING VARIABLE 4-8 BY LATE EVENING`

The input vector is a sequence of 7-tuples $\langle i, d, s_{min}, s_{max}, g_{min}, g_{max}, t\rangle$ where i is the tuple's ID, d is the wind direction, s_{min} and s_{max} are the minimum and maximum wind speeds, g_{min} and g_{max} are the minimum and maximum gust speeds, and t is a time stamp (indicating for what time of the day the data is valid). The wind forecast texts were taken from comprehensive maritime weather forecasts produced by the professional meteorologists employed by a commercial weather forecasting company for clients who run offshore oilrigs.

There were two evaluation criteria; Clarity was explained as indicating how understandable a forecast was, and Readability as indicating how fluent and readable it was. The experiment involved 22 forecast dates and outputs from the 10 systems described in (Belz and Kow, 2009) (also included in the corpus release) for those dates (as well as the corresponding forecasts in the corpus) in the evaluation, i.e. a total of 242 forecast texts.

4.2 Rating scale experiment

We used the results of a previous experiment (Belz and Kow, 2009) in which participants were asked to rate forecast texts for Clarity and Readability, each on a scale of 1–7.

The 22 participants were all University of Brighton staff whose first language was English and who had no experience of NLP. While earlier experiments used master mariners as well as lay-people in a similar evaluation (Belz and Reiter, 2006), these experiments also demonstrated that the correlation between the ratings by expert evaluators and lay-people is very strong in the METEO domain (Pearson's $r = 0.845$).

We used a single 22 (evaluators) by 22 (test data items) Latin Square; there were 484 trials in this experiment.

4.3 Preference judgement experiment

Our new experiment used our standardised preference strength sliders (bottom of Figure 2). We recruited 22 different evaluators from among students currently completing or recently having

[4]The Prodigy-METEO corpus is freely available here: http://www.nltg.brighton.ac.uk/home/Anja.Belz

Type	Measure	Clarity	Readability
RSE	$F_{(10,473)}$	23.507**	24.351**
	N sig diffs	24/55	23/55
	K's W	.497**	.533**
(Text	$F_{(21,462)}$	1.467	1.961**)
(Evaluator	$F_{(21,462)}$	4.832**	4.824**)
PJE	$F_{(10,1865)}$	45.081**	41.318**
	N sig diffs	34/55	32/55
	K's W	.626**	.542**
(Text	$F_{(21,916)}$	1.436	1.573)
(Evaluator	$F_{(21,921)}$	.794	1.057)

Table 3: METEO RSE/PJE: Results of analyses looking at effect of System.

completed a linguistics-related degree at Oxford, KCL, UCL, Sussex and Brighton.

We had at our disposal 11 METEO systems, so there were $\binom{11}{2} = 55$ system combinations to evaluate on the 22 test data items. We decided on a design of ten 11 × 11 Latin Squares to accommodate the 55 system pairings, so there was a total of 1210 trials in this experiment.

4.4 Results and comparative analysis

Table 3 shows the same types of comparative measures as in the previous section. Note that the self-consistency measure is missing, because for METEO-PJE we do not have multiple scores for the same pair of systems by the same evaluator.

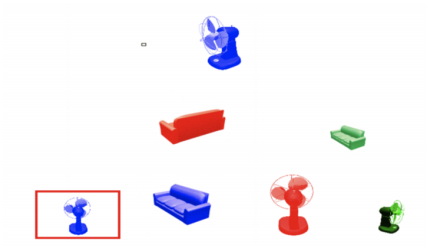
For the METEO task, the relative amount variation in Clarity and Radability accounted for by System is similar in the RSE, and again similar in the PJE. However, F ratios and numbers of significant differences found are higher in the latter than in the RSE. The inter-evaluator agreement measure also has higher values for both Clarity and Readability in the PJE, although the difference is much more pronounced in the case of Clarity.

In the RSE version, Evaluator has a small but significant effect on both Clarity and Readability, which disappears in the PJE version. Similarly, a small effect of Text (date of weather forecast in this data set) on Fluency in the RSE version disappears in the PJE version.

5 RSE/PJE Pair 2: Descriptions of furniture items and faces

5.1 Data and generic design

In our third pair of evaluations, we used the system outputs from the TUNA'09 shared-task competition (Gatt et al., 2009).[5] The TUNA data is a collection of images of domain entities paired with descriptions of entities. Each pair consists of seven entity images where one is highlighted (by a red box surrounding it), paired with a description of the highlighted entity, e.g.:



`the small blue fan`

The descriptions were collected in an online experiment with anonymous participants, and then annotated for semantic content. In TUNA'09, the task for participating systems was to generate descriptions of the highlighted entities given semantic representations of all seven entities. In the evaluation experiments, evaluators were asked to give two ratings in answer to the following questions (the first for Adequacy, the second for Fluency):

1. How clear is this description? Try to imagine someone who could see the same grid with the same pictures, but didn't know which of the pictures was the target. How easily would they be able to find it, based on the phrase given?

2. How fluent is this description? Here your task is to judge how well the phrase reads. Is it good, clear English?

Participants were shown a system output, together with its corresponding domain, displayed as the set of corresponding images on the screen. The intended (target) referent was highlighted by a red frame surrounding it on the screen.

Following detailed instructions, subjects did two practice examples, followed by the 112 test items in random order.

There were 8 'systems' in the TUNA evaluations: the descriptions produced by the 6 systems and two sets of humans-authored descriptions.

5.2 Rating scale experiment

The rating scale experiment that was part of the TUNA'09 evaluations had a design of fourteen 8 × 8 squares, and a total of 896 trials.

[5]The TUNA'09 data and documentation is available for download from http://www.nltg.brighton.ac.uk/home/Anja.Belz

Type	Measure	Adequacy	Fluency
RSE	$F_{(7,888)}$	6.371**	17.207**
	N sig diffs	7/28	15/28
	K's W	.471**	.676**
(Text1	$F_{(111,784)}$	1.519**	1.091)
(Text2	$F_{(14,881)}$	8.992**	4.694**)
(Evaluator	$F_{(7,888)}$	13.136**	17.479**)
PJE	$F_{(7,6264)}$	46.503**	89.236**
	N sig diffs	19/28	22/28
	K's W	.573**	.654**
(Text1	$F_{(111,3024)}$	.746	.921)
(Text2	$F_{(14,3121)}$	.856	.853)
(Evaluator	$F_{(27,3108)}$	1.3	1.638*)

Table 4: TUNA RSE/PJE: Results of analyses looking at effect of System.

Subjects were asked to give their judgments for Clarity and Fluency for each item by manipulating a slider. The slider pointer was placed in the center at the beginning of each trial. The position of the slider selected by the subject mapped to an integer value between 1 and 100. However, the scale was not visible to participants who knew only that one end of the scale corresponded to the worst possible score and the opposite end to the best.

Eight native speakers of English were recruited for this experiment from among post-graduate students currently doing a Masters degree in a linguistics-related subject at UCL, Sussex and Brighton universities.

5.3 Preference judgement experiment

Our new experiment used our standardised preference strength sliders (bottom of Figure 2). To accommodate all pairwise comparisons as well as all test set items, we used a design of four 28 $\times$ 28 Latin Squares, and recruited 28 evaluators from among students currently completing, or recently having completed, a degree in a linguistics-related subject at Oxford, KCL, UCL, Sussex and Brighton universities. There were 3,136 trials in this version of the experiment.

5.4 Results and comparative analysis

Table 4 shows the same measures as we reported for the other two experiment pairs above. The picture is somewhat similar in that the measures have better values for PJE version except for the inter-evaluator agreement (Kendall's W) for Fluency which is slightly higher for the RSE version.

For the TUNA dataset, we look at two Text factors. Text2 refers to different sets of entities used in trials; there are 15 different ones. Text1 refers to sets of entities and their specific distribution over the visual display grid in trials (see the figure in Section 5.1); there are 112 different combinations of entity set and grid locations.

The most striking aspect of the results in Table 4 is the effect of Evaluator in the RSE version which appears to account for more variability in the data even than System (relative to other factors). In fact, in the case of Adequacy, even Text2 causes more variation than System. In contrast, in the PJE version, by far the biggest cause of variability is System (for both criteria), and the F ratios for Text and Evaluators are not significant except for Evaluator on Fluency (weakly significant at .05).

On the face of it, the variation between evaluators in the RSE version as evidenced by the F ratio is worrying. However, Kendall's W shows that in terms of mean rank, evaluators actually agreed similarly well on Fluency in both RSE and PJE. The F measure is based on mean scores whereas W is based on mean score ranks, so there was more variation in the absolute scores than in the ranks.

The reason is likely to be connected to the way ratings were expressed by evaluators in the TUNA-RSE experiment: recall that evaluators had the task of moving the pointer to the place on the slider bar that they felt corresponded to the quality of text being evaluated. As no numbers were visible, the only information evaluators had to go on was which was the 'worse' end and which was the 'better' end of the slider. It seems that different evaluators used this evaluation tool in very different ways (accounting for the variation in absolute scores), but were able to apply their way of using the tool reasonably consistently to different texts (so that they were able to achieve reasonably good agreement with the other evaluators in terms of relative scores).

6 Discussion

We have looked at a range of aspects of evaluation experiments: the effect of the factors System, Text and Evaluator on evaluation scores; the number of significant differences between systems found; self-consistency; and inter-evaluator agreement (as described by F ratios obtained in one-way ANOVAs for Evaluator, as well as by Kendall's W measuring inter-evaluator agreement).

The results are unambiguous as far as the Clarity criterion (called Adequacy in TUNA) is concerned: in all three experiment pairs, the preference-strength judgement (PSE) version had a greater effect of System, a smaller effect of Text and Evaluator, more significant pairwise differences, better inter-evaluator agreement, and (where we were able to measure it) better self-consistency.

The same is true for Readability in METEO and Fluency in TUNA, in the latter case except for W which is slightly lower in TUNA-PJE than TUNA-RSE. However, Readability in GREC-NEG bucks the trend: here, all measures are worse in the PJE version than in the RSE version (although for the W measures, the differences are small). Part of the reason for this may be that in GREC-NEG PJE each system was only compared to one single other 'system', the (human-authored) original Wikipedia texts.

If we see less effect of Clarity than of Fluency in an experiment (as in GREC-NEG RSE and TUNA RSE), then we might want to conclude that systems differed less in terms of Clarity than in terms of Fluency. However, the real explanation may be that evaluators simply found it harder to apply the Clarity criterion than the Fluency criterion in a given evaluation set-up. The fact that the difference in effect between Fluency and Clarity virtually disappears in GREC-NEG PJE makes this the more likely explanation at least for the GREC-NEG evaluations.

Parametric statistics are more powerful than non-parametric ones because of the strong assumptions they make about the nature of the data. Roughly speaking, they are more likely to uncover significant differences. Where the assumptions are violated, the risk is that significance is overestimated (the likelihood that null hypotheses are incorrectly rejected increases). One might consider using a slider mapping to a continuous scale instead of a multiple-choice rating form in order to overcome this problem, but the evidence from the TUNA RSE evaluation appears to be that this can result in unacceptably large variation in how individual evaluators apply the scale to assign absolute scores.

What seems to make the difference in terms of ease of application of evaluation criteria and reduction of undesirable effects is not the use of continuous scales (as e.g. implemented in slider bars), but the comparative element, where pairs of systems are compared and one is selected as better in terms of a given criterion than the other.

It makes sense intuitively that deciding which of two texts is clearer should be an easier task than deciding whether a system is a 5, 4, 3 or 1 in terms of its clarity. PJEs enabled evaluators to apply the Clarity criterion to determine ranks more consistently in all three experiment pairs.

However, it was an open question whether evaluators would also be able to express the *strength* of their preference consistently. From the results we report here it seems clear that this is indeed the case: the System F ratios which look at absolute scores (in the PJEs quantifying the strength of a preference) are higher, and the Evaluator F ratios lower, in all but one of the experiments.

While there were the same number of trials in the two GREC-NEG evaluations, there were 2.5 times as many trials in METEO-PJE than in METEO-RSE, and 3.5 times as many trials in TUNA-PJE than in TUNA-RSE. The increase in trials is counter-balanced to some extent by the fact that evaluators tend to give relative judgements far more quickly than absolute judgements, but clearly there is an increase in cost associated with including all system pairings in a PJE. If this cost grows unacceptably large, a subset of systems has to be selected as reference systems.

7 Concluding Remarks

Our aim in the research presented in this paper was to investigate how rating-scale experiments compare to preference-strength judgement experiments in the evaluation of automatically generated language. We find that preference-strength judgement evaluations generally have a greater relative effect of System (the factor actually under investigation), a smaller relative effect of Text and Evaluator (whose effect should be small), a larger number of significant pairwise differences between systems, better inter-evaluator agreement, and (where we were able to measure it) better evaluator self-consistency.

References

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.

A. Belz and E. Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of EACL'06*, pages 313–320.

Anja Belz, Eric Kow, and Jette Viethen. 2009. The GREC named entity generation challenge 2009: Overview and evaluation results. In *Proceedings of the ACL-IJCNLP'09 Workshop on Language Generation and Summarisation (UCNLG+Sum)*, pages 88–98.

A. Belz. 2009. Prodigy-METEO: Pre-alpha release notes (Nov 2009). Technical Report NLTG-09-01, Natural Language Technology Group, CMIS, University of Brighton.

Bernard Choi and Anita Pak. 2005. A catalog of biases in questionnaires. *Preventing Chronic Disease*, 2(1):A13.

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.

C.-Y. Lin and F. J. Och. 2004. ORANGE: A method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, pages 501–507, Geneva.

E. Reiter, S. Sripada, J. Hunter, and J. Yu. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167:137–169.

Sidney Siegel. 1957. Non-parametric statistics. *The American Statistician*, 11(3):13–19.

S. Sripada, E. Reiter, J. Hunter, and J. Yu. 2002. SUMTIME-METEO: A parallel corpus of naturally occurring forecast texts and weather data. Technical Report AUCS/TR0201, Computing Science Department, University of Aberdeen.

H. Trang Dang. 2006. DUC 2005: Evaluation of question-focused summarization systems. In *Proceedings of the COLING-ACL'06 Workshop on Task-Focused Summarization and Question Answering*, pages 48–55.

J. Turian, L. Shen, and I. D. Melamed. 2003. Evaluation of machine translation and its evaluation. In *Proceedings of MT Summmit IX*, pages 386–393, New Orleans.