# What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more under-powered than you think

**David M. Howcroft**[*]
School of Computing
Edinburgh Napier University
Edinburgh, Scotland, UK
d.howcroft@napier.ac.uk

**Verena Rieser**
The Interaction Lab, MACS
Heriot-Watt University
Edinburgh, Scotland, UK
v.t.rieser@hw.ac.uk

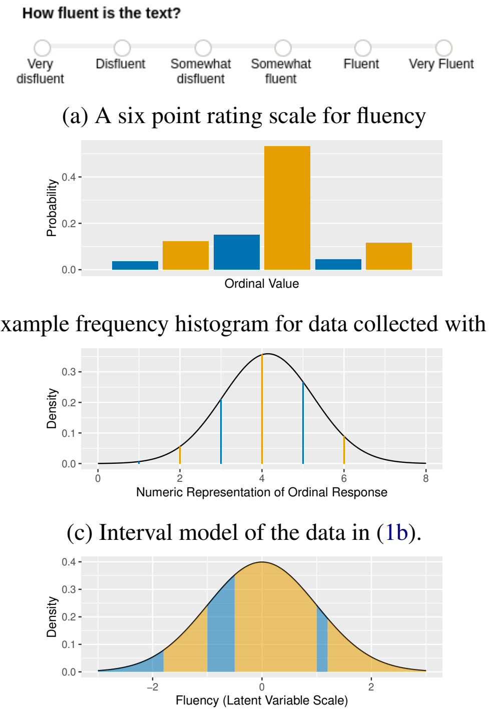
## Abstract

Previous work has shown that human evaluations in NLP are notoriously under-powered. Here, we argue that there are two common factors which make this problem even worse: NLP studies usually (a) treat ordinal data as interval data and (b) operate under high variance settings while the differences they are hoping to detect are often subtle. We demonstrate through simulation that ordinal mixed effects models are better able to detect small differences between models, especially in high variance settings common in evaluations of generated texts. We release tools for researchers to conduct their own power analysis and test their assumptions. We also make recommendations for improving statistical power.

## 1 Introduction

Human evaluation remains the gold standard for many natural language generation tasks, including machine translation, data-to-text, summarisation, and dialogue & interactive systems. One common way to elicit text quality ratings from study participants is to use a rating scale, e.g. a Likert scale which measures agreement with a statement, or other visual or verbal analogue scales, as in Figure 1a. Unfortunately, typically chosen statistical analyses of these scores often rely on the flawed assumption that the rating scales are *interval*, i.e. that the distance between any two adjacent points on the scale is the same across the full range of values, so that, for example, the difference between 'very disfluent' & 'disfluent' is the same as the distance between 'slightly disfluent' & 'slightly fluent' on a 6-point scale (see Figure 1).

The distributions in Figure 1 illustrate the different underlying assumptions of interval and ordinal models of rating scale data: The rating scale in (1a) is used to collect human judgements of text quality (e.g. fluency), which results in a distribution of ordinal data as in (1b). In (1c) we follow the interval assumption, that each point on the six-point rating scale corresponds directly to a real-valued integer and that we can model the relative probability of any pair of ratings based on a Gaussian probability density function. In contrast, (1d) assumes that there is a latent variable for text quality and that the ordinal scores from our surveys correspond to different ranges of values on this latent scale.[1] Different model assumptions influence the choice of statistical significance test: when the data distri-



(a) A six point rating scale for fluency

(b) Example frequency histogram for data collected with (1a).

(c) Interval model of the data in (1b).

(d) Ordered probit model of the data in (1b).

Figure 1: The model in (1d) yields more accurate probabilities than those in (1c).

[*] Work completed while at Heriot-Watt University.

[1] For example, in this figure all latent values below $-1.8$ are associated with the category 'Very disfluent', which has a probability equal to the cumulative probability of the latent Gaussian distribution at $-1.8$.

*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8932–8939
November 7–11, 2021. ©2021 Association for Computational Linguistics

bution is known, it is often possible to choose a *parametric* test to achieve greater statistical power at lower computational cost (Dror et al., 2018).

While the debate about whether and when ordinal scales can be treated as interval has been fought for several decades (Glass et al., 1972; Knapp, 1990; Jamieson, 2004; Carifio and Perla, 2007; Wu and Leung, 2017; Liddell and Kruschke, 2018), we argue that ordinal data needs to be analysed as ordinal in NLP: in this paper we demonstrate that this misinterpretation of rating scales does in fact limit the statistical soundness of our studies by simulating the effects. Previous research has shown that human evaluations are notoriously under-powered (Card et al., 2020). We show that these effects will be exaggerated if we treat ordinal data as interval. We compare the linear mixed effects models proposed by Card et al. (2020), which treats rating scale data as interval, and compare it to a corrected version, which uses ordered probit models and appropriately treats the data as ordinal. We show that ordinal models are more likely to detect a real effect, especially when the effect size is small, the variance is high, or the sample size is small, all of which are common in human evaluations.

We release all of our code so that other researchers can adjust the assumptions of our models to match the reality of their evaluation settings and easily estimate appropriate sample sizes using the same simulation methods: `https://www.github.com/dmhowcroft/ordinal-models`

## 2 Current reporting practices

Significance testing provides an assessment of how extreme the observed values are according to a random noise model. For example, if an observed difference in performance between two systems is not distinguishable from noise centered at zero, then we would not want to rank one system above the other, with implications for leaderboards and the replicability of results (van der Lee et al., 2019; Dror et al., 2018; Card et al., 2020).[2] However, not many studies include significance tests: regardless of whether using automated metrics or human evaluations, only about a third of studies reported significance tests according to recent surveys (Dror et al., 2018; van der Lee et al., 2019). And even when researchers do include significance tests, they often apply the tests incorrectly (Dror et al., 2018; Amidei et al., 2019), with Amidei et al. (2019) reporting that the majority of recent papers incorrectly interpret rating and Likert scales as interval data (up to 84% for Likert scales; Figure 1 illustrates why this is a problem).

## 3 Models of Ordinal Data

Card et al. (2020) suggest that NLP researchers follow psycholinguists in adopting linear mixed effects (LME) models[3] for statistical modelling and significance testing. Mixed effects models control for random noise due to the individual items and participants in an experiment and allow for richer statistical comparisons than $t$-tests or ANOVAs, though they have the same drawbacks in assuming that the data is metric. The general form is given in Equation 1:

$$Y = X\boldsymbol{\beta} + Z\mathbf{u} + \epsilon, \tag{1}$$

where $X$ and $Z$ are design matrices for fixed and random effects, respectively, $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{u}$ is a vector of random effects, and $\epsilon$ is the residual noise in the model, assumed to be Gaussian. Given some observed data ($Y$) we estimate the fixed ($\boldsymbol{\beta}$) and random ($\mathbf{u}$) effects.

In the common `lme4` (Bates et al., 2015) notation, a model comparing ratings for several systems with random effects for participants and items is:

```
rating ~ system +
           (system|participant) +
           (system|item)
```

This specifies a model for `ratings` with a fixed effect of `system` and random effects `(system|participant)` and `(system|item)`. As a 'maximal model' (Barr et al., 2013), it includes random *intercepts* to represent the general bias of individual participants and items (e.g. some participants give higher or lower ratings on average) and random *slopes* to represent the interactions between random and fixed effects (e.g. some users may systematically prefer system A or system B). These random effects are designed to control for the fact that our participants and items are samples from larger populations: we are not interested in the behaviors of these individuals, but rather in more general assessments of text quality that should generalise to a larger population

[2] Note that some argue against statistical significance tests for NLP because the independence assumptions do not necessarily hold (Koplenig, 2019; McShane et al., 2019; Carver, 1978; Leek et al., 2017).

[3] Also known as hierarchical or multi-level models.

Instead of using LME models, we argue that researchers should use *ordinal* mixed-effects models to analyse ordinal data. Unlike LME models, ordinal regression models do not assume that the data is metric. Here we focus on *ordered probit* models as implemented in `ordinal` (Christensen, 2019) for ease of explication, but researchers are free to use alternative linking functions (e.g. logit) or tools as needed.[4] The key difference from the LME model is that we no longer assume that our observed ratings $Y$ are on a continuous scale we can model directly. Instead, we assume that there is an underlying latent variable $\mathbf{Y_l}$ which is continuous. We represent this latent variable with a standard Gaussian distribution with a mean of 0 and a standard deviation of 1. The link between the observed variables $\mathbf{Y}$ with $k$ possible categories and $\mathbf{Y_l}$ is then based on fitting a series of $k-1$ thresholds $\tau$ such that:

$$\begin{aligned} P(Y=i) = \quad & \Phi(\tau_i - X\boldsymbol{\beta} - Z\mathbf{u}) - \\ & \Phi(\tau_{i-1} - X\boldsymbol{\beta} - Z\mathbf{u}), \end{aligned} \tag{2}$$

where $\Phi$ is the cumulative density function for the Gaussian distribution and other terms are as defined above. This corresponds to a model where, when participants are asked to rate an item, they are implicitly accessing this continuous variable and determining how best to bin it based on the categories available to them. Figure 1 exemplifies this for a single system (i.e. fitting $\tau_i$ but no fixed or random effects).

When comparing systems, then, the goal of the model is to fit these thresholds *as well as* a fixed effect representing the differences between the systems while controlling for noise. These differences can be thought of as shifting the thresholds along the latent variable axis or, equivalently, as shifting the mean of the underlying latent variable.

## 4 Simulation Experiments

Our experiments take the form of a *power simulation*, i.e. an analysis of the statistical power of a given test in typical or expected experimental conditions. In a power simulation we generate a set of data with known parameters (e.g. a known 'effect size' difference between conditions) and measure how often a statistical test correctly identifies that effect in the simulated data. In order to limit the complexity of a power simulation, the researcher estimates 'typical' values of as many model parameters as possible and then systematically explores possible values for the other parameters.

In our case, we begin by fitting ordinal models on several NLG evaluation datasets. The resulting models then allow us to simulate NLG datasets with different numbers of raters, different amounts of variance, and different effect sizes in order to understand how many participants are needed to detect effects of different sizes.

### 4.1 Datasets

We use 6 datasets to estimate parameters for our simulations: 4 datasets used by (Card et al., 2020, HUSE$_{1\text{-}3}$ & PPLM), the dataset used by (Novikova et al., 2017, NEM$_{1\text{-}2}$) and a reproduced version of the NEM dataset with new ratings gathered using different instructions (reNEM).

The HUSE$_{1\text{-}3}$ datasets include 'typicality' judgements from crowdworkers on a 6-point scale ranging from 'invalid' to 'very typical' for 3 different tasks: sampling from LMs, summarisation, and chit-chat conversational turn generation (Hashimoto et al., 2019). PPLM includes 'fluency' judgements from expert annotators on a 5-point scale ranging from 1 = "not fluent at all" to 5 = "very fluent" for texts generated in a lightly-conditioned style- or topic-transfer task (Dathathri et al., 2020). NEM$_{1\text{-}2}$ and reNEM include 'quality', 'naturalness', and 'informativeness' judgements from crowdworkers on a 6-point scale for data-to-text generation in the restaurant domain. We included NEM$_{1\text{-}2}$ and reNEM to compensate for the fact that two of the datasets used by Card et al. (2020) are not publicly available and to include reproduced ratings for the same outputs. The more detailed instructions provided to reNEM raters result in a higher degree of interannotator agreement.

Note that the datasets we had access to for this study are mostly use 6-points, while (van der Lee et al., 2019) found that the most frequently used rating scale in NLG research is the 5-point scale, which is also confirmed by (Howcroft et al., 2020). Other frequently used scales are 3,4,6,7-point. However, the methodology we propose should generalise well to other scale sizes.

Among the datasets we use, there is wide variation in the number of ratings included (ranging from 4k to 41k ratings, median 7.4k, mean 13k). Further details are provided in Appendix A.

[4] This is a *cumulative link model*. Alternatives include *sequential* and *adjacent category* models. Bürkner and Vuorre (2019) provide an overview. Appendix C lists other tools.

### 4.2 Experiment Settings

We estimate parameter settings for our simulations by fitting an ordered probit model for each dataset above. The low (high) variance setting uses the smallest (largest) observed by-participants and by-items variances, while the 'General' condition is based on the mean observed variances.[5] We also base the distance between thresholds in the latent variable space on the estimates from these fitted models.

We then simulate 100 experiments comparing two systems for each combination of experimental design factors considered: (1) 3 or 10 participants per item; (2) 50, 100, or 500 items per system; and (3) an effect size of 0.25, 0.5, 0.75, or 1 times the distance between adjacent thresholds. We base the effect sizes on the settings used by Card et al. (2020): in their experiment, the average distance between adjacent values was 0.2 on a 0-1 scale and they used effect sizes of 0.05, 0.1, 0.15, and 0.2.

Unlike Card et al. (2020), we construct design matrices to create our item lists such that each participant sees only 25 items and never sees the same item in multiple conditions (rather than seeing all items in every condition). This represents a more realistic experimental design, since designs requiring every participant to rate every item are rare.

The interval assumption in (Card et al., 2020) also influences the quality of the simulations used for their analyses: since they do not model variance in an ordinal regression model, their simulated data will, in fact, be interval data, unlike the data they seek to model. We also correct for the calculation of $p$-values for LMEs by using the `lmerTest` library (Kuznetsova et al., 2017), which is designed to produce accurate $p$-values by approximating the number of degrees of freedom.[6]

For all of our tests we used the conventional $p < 0.05$ significance threshold. For the ordinal models, the `ordinal` package itself provides $p$-values. Our plots show the proportion of simulations for a particular condition where the statistical test identified the underlying effect as significant.

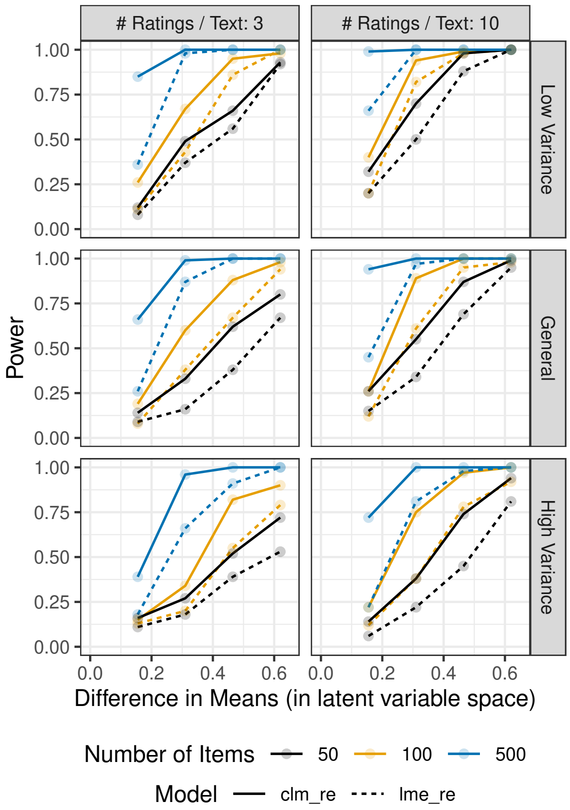


Figure 2: Power curves for the same data fit with both linear mixed-effects models (lme_re) and ordinal mixed-effects models (clm_re). Additional variance settings are included in Appendix B.

### 4.3 Results

Figure 2 shows the results of these simulations. Each point on the curve represents the proportion of times the given statistical model (either an ordinal – represented with solid lines – or a linear mixed-effects model – dashed lines) is able to detect an effect of the size given on the x-axis (i.e. the model's power at that effect size). The message is clear: the ordinal model is always more likely to detect a true effect of any size than the corresponding linear model is (all of the solid lines of a given color are always above their dashed counterpart). However, this is especially true for settings with high variance and for smaller effect sizes. Moreover, the ordinal model using only 50 items is approximately as powerful as the linear model using 100 items! As such, we can conclude that using an ordinal model for rating and Likert scales will always lead to more reliable results. However, for settings with high variance and small data samples, as typically the case for human NLP evaluations, using ordinal models is even more crucial.

[5] Since these variances are in part dependent on the size of the scale, we use only the 6-point scales in these analyses, with 'extra low' and 'extra high' variance settings based on the PPLM dataset included in the appendices.

[6] Card et al. instead directly used $z$-values output by the `lme4` package as though they corresponded to $z$-values in other kinds of statistical tests, with a clean mapping to $p$-values. The authors of `lme4` advise against doing this.

In the meta-analysis comparing different datasets mentioned above, we found that the difference between models ranged from -0.6 to 1.0, with 9 out of 12 systems for which an effect was estimated having a difference less than 0.46 (i.e. 0.75 times the average distance between adjacent thresholds). The above analysis indicates that a study with 100 items and only 3 ratings per text would require an ordinal model to detect an effect of this size with 80% power, except in the low variance setting. While van der Lee et al. (2019) found that the median/average study did use 100 items and 4 annotators, they also found that "only 55% of papers specified the number of participants" and they did not report on how many items each participant rated. Since most studies are not using ordinal analyses of their data (Amidei et al., 2019), our simulation results suggest that **most human evaluations are underpowered to detect typical system differences**, exaggerating the effects reported in Card et al. (2020).

## 5 Discussion

In contrast to the (common) assumption followed by Card et al. (2020) that ordinal data can be analysed as interval data, we show that treating ordinal data as interval makes human ratings even more under-powered. This is a problem because, in practice, NLP evaluations often aim to detect small differences (i.e. effect sizes) in high variance settings while operating under a limited budget or with limited access to human raters.

Since our proposed framework is independent from the concrete instantiation of the scale and generalises well, our hope is that other researchers can adapt our code to gain a better understanding of what kind of scale and statistical model to use for their next experiment. We also recommend setting simulation parameters based on e.g. their own past experiments if similar.

One open question is how to best choose the best scale and model. In general, each researcher needs to choose appropriate tools based on their knowledge of the data. On the one hand, they may prefer to start with 5+ points on their scale, use ordinal regression to measure variance, and only later conclude that the differences seem large enough for their task & survey instruments that they can switch to simpler scales and/or models. On the other hand, they may reason that 'yes-no questions are easy/cheap to ask, so let's see if those are informative enough for our needs'. If the differences between systems are large enough, they may even be able to use an even simpler model than an LME model (for example, a simple Chi-squared test on 'the proportion of positive responses'). However, if the differences are not in fact large enough for such a simple scale & analysis to capture, then they have wasted time and resources to collect data they cannot use. Both approaches are reasonable, but researchers should be aware of the power problems highlighted in our paper when they start planning and choosing an approach.

## 6 Conclusion

We see three core ways to improve the power of human evaluations: First, reduce noise in human ratings. The reNEM dataset's clear definitions, guidance, and training reduced noise in the resulting human ratings, which reduces between-participants variance and increases the ability of a statistical model to distinguish between similar systems. Similar studies have been conducted for machine translation (Freitag et al., 2021).

In addition to providing clear instructions, we can also design experiments to include more items and more participants, using power analyses like the ones presented in this paper to estimate how large a sample we need before collecting any data.

Most importantly, however, we recommend researchers use ordinal models to analyse ordinal data to have the greatest statistical power when testing hypotheses. This is especially important for setups with high variance and small data samples, as often the case for human evaluations in NLP.

## Acknowledgements

This research was was supported by the EPSRC project MaDrIgAL (EP/N017536/1).

## References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. The use of rating and Likert scales in natural language generation human evaluation tasks: A review and some recommendations. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 397–402, Tokyo, Japan. Association for Computational Linguistics.

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

Douglas Bates, Martin Mächler, Benjamin M. Bolker, and Steven C. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1).

Paul Christian Bürkner and Matti Vuorre. 2019. Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

James Carifio and Rocco J. Perla. 2007. Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. *Journal of Social Sciences*, 3(3):106–116.

R. P. Carver. 1978. The Case Against Statistical Significance Testing. *Harvard Educational Review*, 48:378–399.

Rune Haubo B Christensen. 2019. Cumulative Link Models for Ordinal Regression with the R Package ordinal.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A simple approach to controlled text generation. In *Proc. of the International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume Long Papers, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation.

Gene V. Glass, Percy D. Peckham, and James R. Sanders. 1972. Consequences of failure to meet assumptions underlying the analyses of variance and covariance. *Review of Educational Research*, 42:237–288.

F. E. Harrell, Jr. 2021. Rms: Regression Modelling Strategies.

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.

David M Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, and Simon Mille. 2020. Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. In *Proc. of the 13th International Conference on Natural Language Generation (INLG)*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

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

Thomas R Knapp. 1990. Treating Ordinal Scales as Interval Scales: An Attempt to Resolve the Controversy. *Nursing Research*, 39:121–123.

Alexander Koplenig. 2019. Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory*, 15(2):321–346.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13).

Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-Attribute Text Rewriting. In *Proc. of the International Conference on Learning Representations (ICLR)*, New Orleans, Louisiana, USA.

Jeff Leek, Blakeley B. McShane, Andrew Gelman, David Colquhoun, Michèle B. Nuijten, and Steven N. Goodman. 2017. Five ways to fix statistics. *Nature*, 551(7682):557–559.

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

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

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why We Need New Evaluation Metrics for NLG. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

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 *Proc. of the 12th International Conference on Natural Language Generation*

*(INLG)*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

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.

W. N. Venables and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Statistics and Computing. Springer, New York, New York, USA.

Huiping Wu and Shing-On Leung. 2017. Can Likert Scales be Treated as Interval Scales?—A Simulation Study. *Journal of Social Service Research*, 43(4):527–532.

Thomas W. Yee. 2010. The VGAM package for categorical data analysis. *Journal of Statistical Software, Articles*, 32(10):1–34.

## A Datasets

**HUSE$_{1\text{-}3}$** The first three datasets come from Hashimoto et al. (2019), and all measure the ‘typicality’ of a text on a 6-point scale ranging from ‘invalid’ to ‘very typical’. These three datasets represent 3 different tasks: sampling from LMs, summarisation, and chit-chat conversational turn generation. The authors report collecting judgements on 100 human and 100 model texts from 20 human participants who each provided 25 ratings. However, these numbers do not match what is found after downloading the data, which we report in Table 1.

**PPLM** Collected by Dathathri et al. (2020), this data includes 5-point rating scale judgements assessing fluency ranging from 1 = “not fluent at all” to 5 = “very fluent” as in (Lample et al., 2019). The task in this case is style- or topic-transfer, and the authors report using 9 professional annotators to rate texts for 4 different models. See further detail in Table 1.

**NEM$_{1\text{-}2}$** Collected by Novikova et al. (2017) in order to assess correlations between automated and human evaluation metrics for data-to-text generation. Participants saw the input slot-value pairs along with two candidate utterances which they then rated on 6-point scales for ‘informativeness’, ‘naturalness’, and ‘quality’. Each crowdworker evaluated a maximum of 20 utterances; each text was scored by 3 different crowdworkers. See further detail in Table 1.

**reNEM** A local re-collection of ratings for the NEM$_{1\text{-}2}$ dataset, this study provided annotators with training in how to use the rating scales and assessed each dimension of quality (‘informativeness’, ‘naturalness’) separately. There are 3 ratings for each text. See further detail in Table 1.

## B Simulating extra low and extra high variance

The PPLM dataset exhibited more extreme values for the random effects structure of the fitted ordered probit model. Since this is 1 of 6 datasets and the other values were closer together, we omit this analysis from the main text. Note, however, that the results support the primary findings: ordered probit models always have more power to detect a true effect than a linear model, though these differences nearly disappear when variance is extremely low and are more pronounced in extremely high variance settings, as seen in the top and bottom plots in Figure 3.

## C Resources for Ordinal Regression Models

There are several packages available for ordinal regression models in `R`, including `MASS` (Venables and Ripley, 2002), `VGAM` (Yee, 2010), `rms` (Harrell, 2021), `ordinal` (Christensen, 2019), and `brms` (Bürkner and Vuorre, 2019). Resources also exist for ordinal models in SPSS, SAS & S-Plus, and STATA; (Christensen, 2019) and (Bürkner and Vuorre, 2019) include pointers to resources for these tools and briefly describe the other `R` packages mentioned here.

| Dataset | Scale Size | Num. Systems | Num. Items | Num. Texts | Num. Part.s | Num. Ratings | $\frac{\text{Ratings}}{\text{Text}}$ | $\frac{\text{Ratings}}{\text{Participant}}$ |
|---|---|---|---|---|---|---|---|---|
| HUSE$_1$ | 6 | 2 | 50 | 100 | 124 | 4000 | 40 | 25 |
| HUSE$_2$ | 6 | 2 | 99 | 197 | 96 | 4000 | 20 | 25 |
| HUSE$_3$ | 6 | 4 | 99 | 382 | 123 | 12000 | 20 | 25 |
| PPLM | 5 | 4 | 1361 | 1361 | 14 | 19486 | 9 | 1356* |
| NEM$_1$ | 6 | 2 | 202 | 296 | – | 2967 | 3 | – |
| NEM$_2$ | 6 | 2 | 972 | 1954 | – | 40965 | 9* | – |
| reNEM | 6 | 3 | 1174 | 2250 | – | 7380 | 3 | – |

Table 1: *Scale size* is the size of the ordinal rating scale. *Num. Systems* is the number of systems being evaluated, Num. Items is the number of unique inputs to the systems, *Num. Texts* is the number of unique outputs being evaluated, *Num. Raters* is the number of unique participants. Num. Ratings is the total number of judgements recorded. Ratings/Text and Ratings/Participant report how many ratings were associated with each text or participant in the most frequent case (*except in two cases where the median is more representative of the distribution). For the NEM$_{1\text{-}2}$ and reNEM datasets, the number of unique raters is not known.

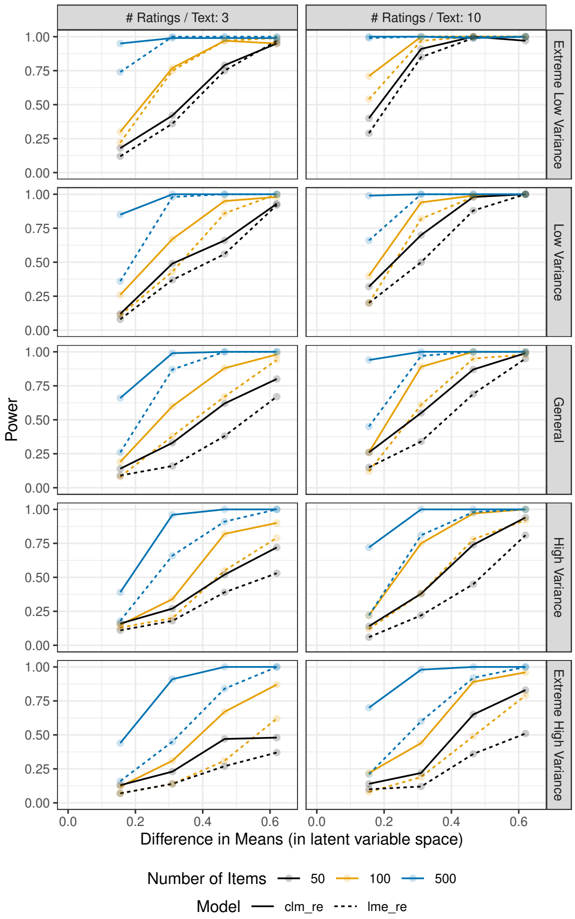


Figure 3: Power curves for the same data fit with both linear mixed-effects models (lme_re) and ordinal mixed-effects models (clm_re).