Design and analysis for ranking of human- and machine-rated applications to a veterinary program

Henrik Stryhn [1], Raphael Vanderstichel [2]

[1] Centre for Veterinary Epidemiological Research, University of PEI

[2] Long Island University

ISVEE 17
Sydney
2024

Human-Rated versus Machine-Rated Essays

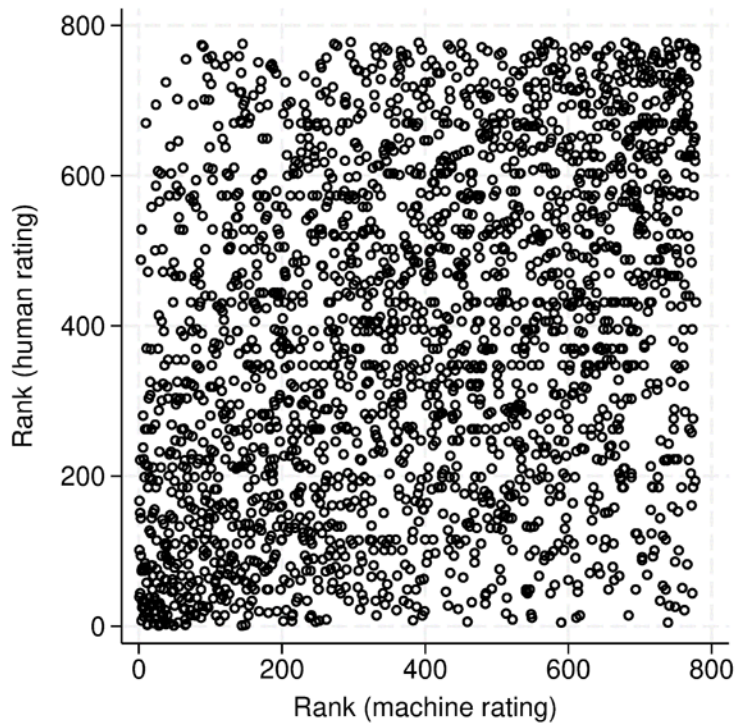
First look at some data (Long Island University, 2021−2022 admission cycle):

- 778 applicants to a veterinary medicine program rated on "writing quality" (1−10 integer scale),
- 13 human raters and Large Language Model (machine) rating (previous presentation).

Raw counts of human versus machine ratings:

human rating	machine rating 3	4	5	6	7	8	9	10	total
3	0	1	1	0	0	0	0	0	2
4	0	0	2	2	3	0	0	0	7
5	0	1	5	21	12	4	0	0	43
6	0	2	6	26	36	43	3	0	116
7	0	1	7	35	50	68	4	0	165
8	0	1	9	42	61	112	12	0	237
9	0	1	4	26	45	87	9	0	172
10	0	0	0	4	11	19	2	0	36
total	0	7	34	156	218	333	30	0	778

Predicted (model-based) rank scatter plot:



Context and Objectives

Role of essays in the admissions process:

- each applicant writes two essays (with detailed specifications), which are graded on several features (items) — for the 2021 – 2022 admission cycle:
  - writing quality (overall, across both essays),
  - content essay I (∼ Veterinary Medicine, as defined by AAVMC[1]),
  - content essay II (∼ Long Island University, with focus on diversity),
- traditionally, essays are rated by a panel of human reviewers (often from the Admissions Committee),
- with machine rating, it is more efficient to grade several essays together (∼ an "adjudication set"), however, (currently) not feasible to include very many essays simultaneously.

Objective of work:
"to assess the reliability and agreement of essay scores, as graded by both machines and humans, for use in the admissions process" (previous presentation[2]).

Objective of this presentation:
"to contrast methods (of design and analysis) for human and machine rating
— as well as their results in our application".

[1] AAVMC = American Association of Veterinary Medical Colleges.
[2] Published as: Vanderstichel & Stryhn (2024), J. Vet. Med. Educ., doi.org/10.3138/jvme-2023-0162.

Rating Designs

Human rating:

- assessments by multiple raters:
  - provide replication in the design, corresponding to blocks,
  - rater effects are rarely ignorable and need to be accounted for,
  - the block design is typically incomplete, but subjects (applicants) need to be "connected" across raters in order to eliminate rater biases,
- actual design dimensions:
  - 13 raters, on average each scoring $987/13 = 76$ subjects (range: $47-95$),
  - $209/778$ (27%) subjects scored twice, all other subjects scored once.

Machine rating — no obvious equivalent of raters,

- ratings are still stochastic (repeating does not give totally identical answers),
- how do we generate meaningful replication? — can we utilize the adjudication sets?

3 designs studied, each with 3 rounds of rating:

- identical replicates (same adjudication sets),
- "rolling" adjudication sets (for connectedness):
  - round 1: subjects $1-5$, $6-10$, $11-15$ etc.
  - round 2: subjects $3-7$, $8-12$, $13-17$ etc.
  - round 3: subjects $5-9$, $10-14$, $15-19$ etc.
- random allocation of subjects to adjudication sets in rounds $2-3$.

Rating Models

- items can be analyzed individually or together as a multivariate outcome,
- ratings are ordered categorical, but can maybe be analyzed by linear models, in particular if one chooses to average ratings per subject across items,
- traditionally, subjects are modelled by random effects, with predictions as BLUPs (best linear unbiased predictors), or their GLMM equivalent.

Human rating:

- rater effects:
  - fixed effects typical for a small, heterogeneous rater group,
  - variability may be unequal across raters,
- single item: linear/ordinal probit mixed models (estimable with standard mixed model packages/`gllamm` command for Stata),
- multiple items together: complex random effects models.

Machine rating:

- adjudication sets:
  - invite random effects modeling, with equal variances,
  - but are within-set correlations positive?
- single item: cross-classified random effects with many levels $\Rightarrow$ `lme4` library in R is almost indispensible (or model builder Laplace approximations),
- multiple items together: even more complex random effects models.

Some Statistical Model Equations

For a single item with quantitative ("continuous") outcomes, the data layout invites a linear mixed model (or a two-way ANOVA),

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \text{where} \tag{1}$$

- rater effects (α_i) are either fixed or random,
- the random essay/subject effects (β_j) and errors (ε_{ij}) have suitable $\mathrm{N}(0, \sigma^2)$ distributions,
- one natural extension is to allow unequal rater error variances (σ_i^2), but many other models exist.

An ordinal probit model for an ordered categorical outcome (of a single item) takes the form:

$$\begin{aligned} p_{ijc} &= \mathrm{P}(y_{ij} \leq c), \quad \text{for a set of ordered categories } \{c\}, \\ \Phi^{-1}(p_{ijc}) &= \mu_c + \alpha_i + \beta_j, \quad \text{with similar specifications for } (\alpha_i) \text{ and } (\beta_j), \end{aligned} \tag{2}$$

and where Φ is the standard normal cumulative distribution function. [3]

Multivariate linear mixed models for multiple items (k) have more terms (compared to (1)), e.g.:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \varepsilon_{ijk}, \quad \text{where} \tag{3}$$

- overall item effects (γ_k) are typically fixed, the effects involving subjects ($\alpha\beta_{ij}$ and $\beta\gamma_{jk}$) are typically random, and the item-specific rater effects ($\alpha\gamma_{ik}$) can be either fixed or random.

[3] A latent variable formulation of the model can incorporate unequal rater variances (at the latent scale).

Sample Mixed Model Results

- the overall writing quality item analysed separately (for simplicity),
- linear mixed model variance components unless indicated otherwise.

Human rating:

	outcome scale	
parameter	linear	ordinal
σ^2(subject)	0.39 (.09)	0.45 (.17)
σ^2(error)	0.77 (.08)	1 (.)
$(\sigma^2(\text{rater}))^a$	0.53 (.21)	0.84 (.)

[a] for comparison only, modelled by fixed effects

- improved fit with unequal rater variances: $\Delta \log L \approx 66$ (df$=$11),
- close to perfect agreement between subject estimates in linear and ordinal models: $r = 0.98-0.99$ (BLUPs/rankings). [4]

Machine rating:

	adjudication sets		
parameter	replic	rolling	random
σ^2(subject)	0.87 (.05)	0.54 (.03)	0.57 (.04)
σ^2(error)	0.08 (.003)	0.31 (.01)	0.38 (.02)
σ^2(adj.set)	0 (.)	0.02 (.01)	0.01 (.01)

- small variance component for adjudication sets $\sim$ small (positive) within-set correlation: e.g., $\hat{\rho}=0.026$ (rolling).
- fair agreement only between subject estimates (rankings) across designs: $r = 0.70-0.86$. [4]

Overall, the agreement between human and machine ratings is low (slide 1): $r \approx 0.31-0.42$ (BLUPs/rankings). [4]

[4] r = Pearson/Spearman/concordance correlation coefficient.

How to Assess Performance of (Human or Machine) Grading Designs

Potential uses:

- for choosing between human versus machine grading (as direct comparison of estimates was not helpful),
- for optimizing machine grading designs.

Some options:

- look at sizes of the variance components,
- explore associations with other applicant characteristics, e.g. GPA,
- compute repeatability and/or reliability,
- quantify uncertainty in subject estimates/rankings, or more specifically the robustness of subject rankings to individual ratings:
  - ideally, the subjects' ranking should not depend too strongly on individual ratings (a particular concern with human rating).

How to Assess Performance of (Human or Machine) Grading Designs

Potential uses:

- for choosing between human versus machine grading (as direct comparison of estimates is not helpful),
- for optimizing machine grading designs.

Some options:

- ✓ look at sizes of variance components: not too informative, scale-dependent and lacking direct interpretation,
- explore associations with other applicant characteristics, e.g. GPA,
- compute repeatability and/or reliability,
- quantify uncertainty in subject estimates/rankings, or more specifically the robustness of subject rankings to individual ratings:
  - ideally, the subjects' ranking should not depend too strongly on individual ratings (a particular concern with human rating).

How to Assess Performance of (Human or Machine) Grading Designs

Potential uses:

- for choosing between human versus machine grading (as direct comparison of estimates is not helpful),
- for optimizing machine grading designs.

Some options:

- ✓ look at sizes of variance components: not too informative, scale-dependent and lacking direct interpretation,
- ✓ explore associations with other applicant characteristics, e.g. GPA: weak, slightly better with machine than human ratings,
- compute repeatability and/or reliability,
- quantify uncertainty in subject estimates/rankings, or more specifically the robustness of subject rankings to individual ratings:
  - ideally, the subjects' ranking should not depend too strongly on individual ratings (a particular concern with human rating).

How to Assess Performance of (Human or Machine) Grading Designs

Potential uses:

- for choosing between human versus machine grading (as direct comparison of estimates is not helpful),
- for optimizing machine grading designs.

Some options:

- ✓ look at sizes of variance components: not too informative, scale-dependent and lacking direct interpretation,
- ✓ explore associations with other applicant characteristics, e.g. GPA:
  weak, slightly better with machine than human ratings,
- ✓ compute repeatability and/or reliability (e.g. Hecker & Violato, 2010[5]):
  highest with replic design, still good (ICC $\approx$ 0.6) for rolling/random, not comparable to human,
- quantify uncertainty in subject estimates/rankings, or more specifically the robustness of subject rankings to individual ratings:
  - ideally, the subjects' ranking should not depend too strongly on individual ratings
    (a particular concern with human rating).

[5] Hecker K, Violato C (2010), Using standardized essays in the veterinary admissions process: are the ratings reliable and valid?, *J. Vet. Med. Educ.*, 37, 254−257.

How to Assess Performance of (Human or Machine) Grading Designs

Potential uses:

- for choosing between human versus machine grading (as direct comparison of estimates is not helpful),
- for optimizing machine grading designs.

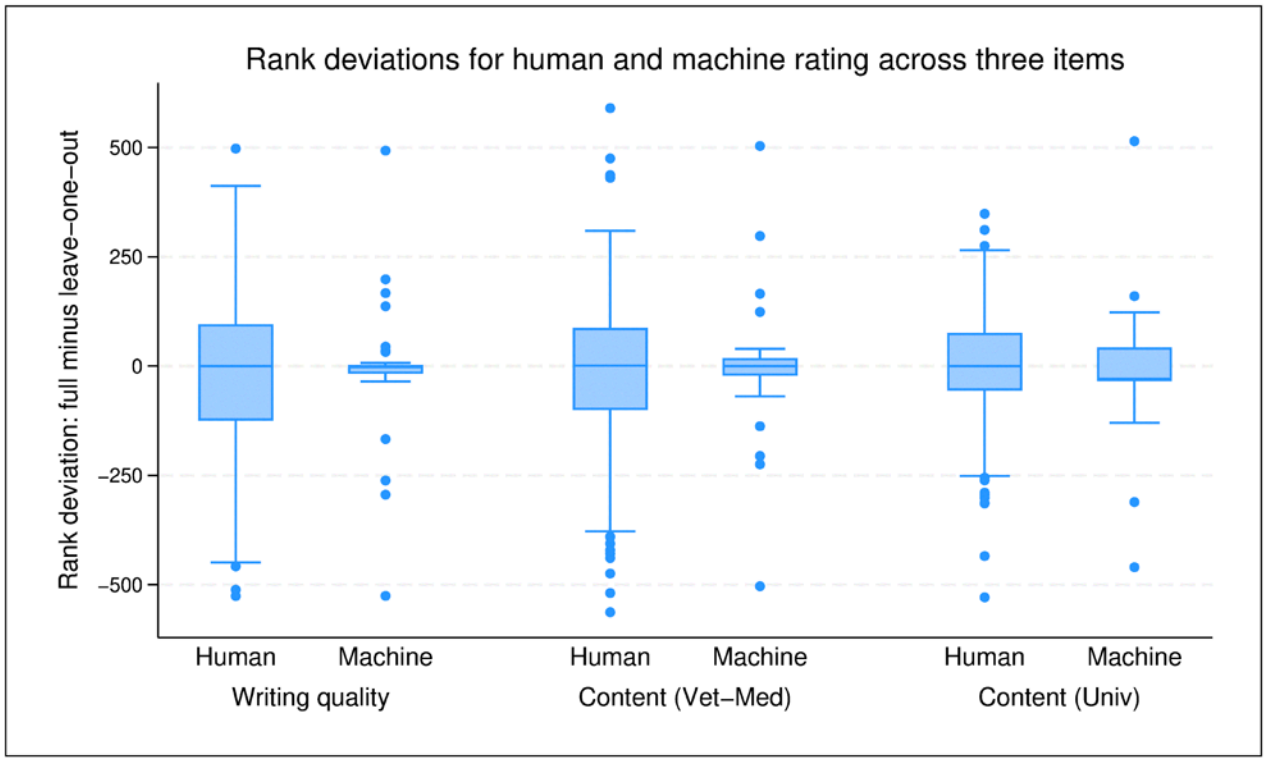
Some options:

- ✓ look at sizes of variance components: not too informative, scale-dependent and lacking direct interpretation,
- ✓ explore associations with other applicant characteristics, e.g. GPA:
  weak, slightly better with machine than human ratings,
- ✓ compute repeatability and/or reliability (e.g. Hecker & Violato, 2010[5]):
  highest with replic design, still good (ICC ≈ 0.6) for rolling/random, not comparable to human,
- ✓ quantify uncertainty in subject estimates/rankings, or more specifically the robustness of subject rankings to individual ratings:
  - leave-one-out analysis across subjects with replication,
  - target: the change (deviation) in subject ranking from omitting one rating,[6]
  - designs with more replication should be more robust.

[5] Hecker K, Violato C (2010), *J. Vet. Med. Educ.*, 37, 254−257.
[6] Adjustment needed to avoid change in BLUPs due to the shrinkage depending on within-subject sample size.

Leave-one-out Analysis Results

Comparisons between human and machine rating across the 3 items:[7]
— (one rank difference per subject: full data subject rank minus leave one out data subject rank)



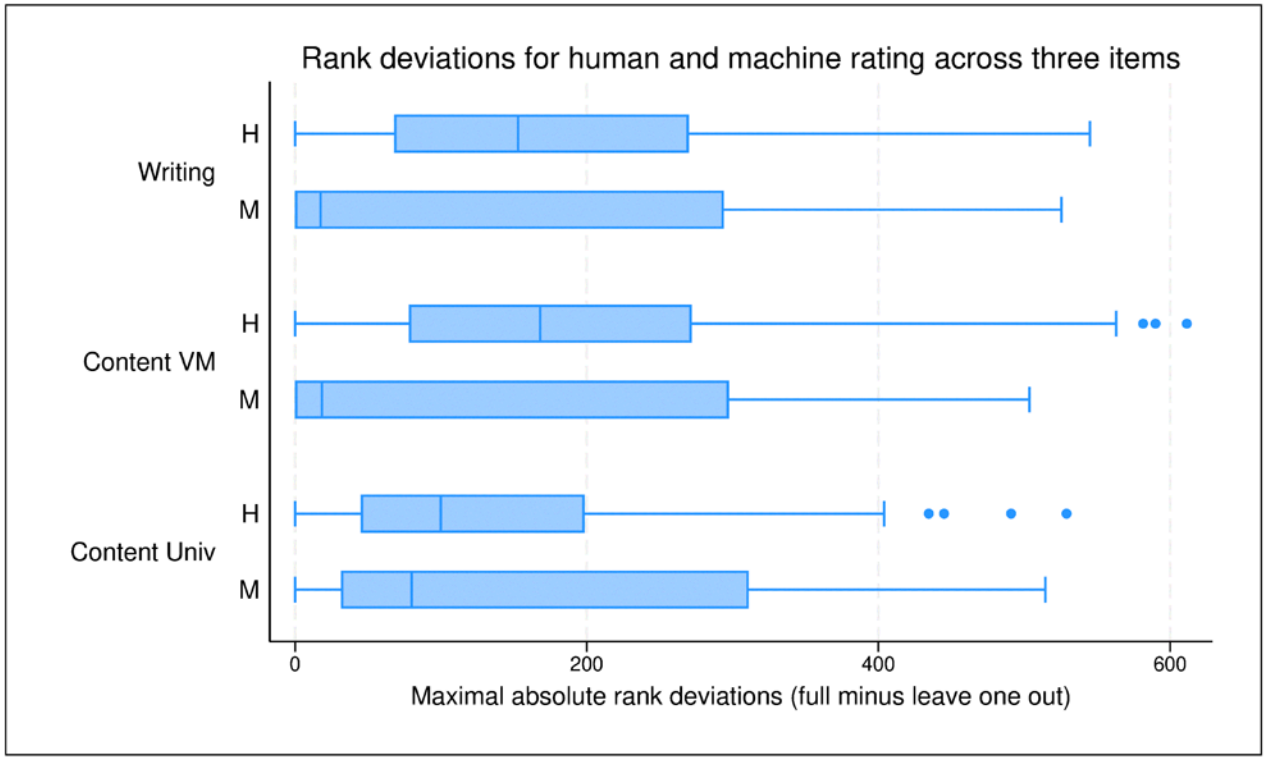
- signs are immaterial (depend on the order of records within subjects),
- larger spread in human rating distributions, especially for the first two items,
- large fluctuations in ranking occur in all designs, when subject scores differ in the center of the distribution (due to the discrete outcome).

To better quantify the robustness of each subject's ranking, we should combine the two rank differences per subject → (e.g.) the largest numerical value (maximal absolute deviation).

[7] 209 (out of 778) applicants with two human and two machine ratings (random allocation to adjudication sets).

Leave-one-out Analysis Results II

Comparisons between human and machine rating across the 3 items:[8]



- statistical comparison of mean (or median) absolute rank deviations for human and machine rating:
  - first two items: strong significance ($P < .001$),
  - third item: no significant difference,
- results reflect a sparse design (only 209 repeated scores for 778 subjects).

Room for all the methods to try to reduce the sensitivity to individual observations — most obviously by increasing the replication.

[8] 209 (out of 778) applicants with two human and two machine ratings (random allocation to adjudication sets), maximal absolute deviations.

Concluding Remarks

Main message: Machine rating designs are less dependent on adjustment for rater (or other design) effects and can be expanded to generate/increase replication, without adding substantial workload or cost.

Second main message: The discrete nature of the ratings make the resulting rankings sensitive, and it can be suggested to try to increase the resolution of the grading scale.

Machine rating of the writing quality and AAVMC content items showed lowest sensitivity to individual scores (and substantially less than by human rating).

Machine rating designs do matter: the "rolling" and "random" designs performed best (results not shown).

In our view, the low agreement between human and machine ratings does not necessarily mean that human grading is preferable, now or in the future.

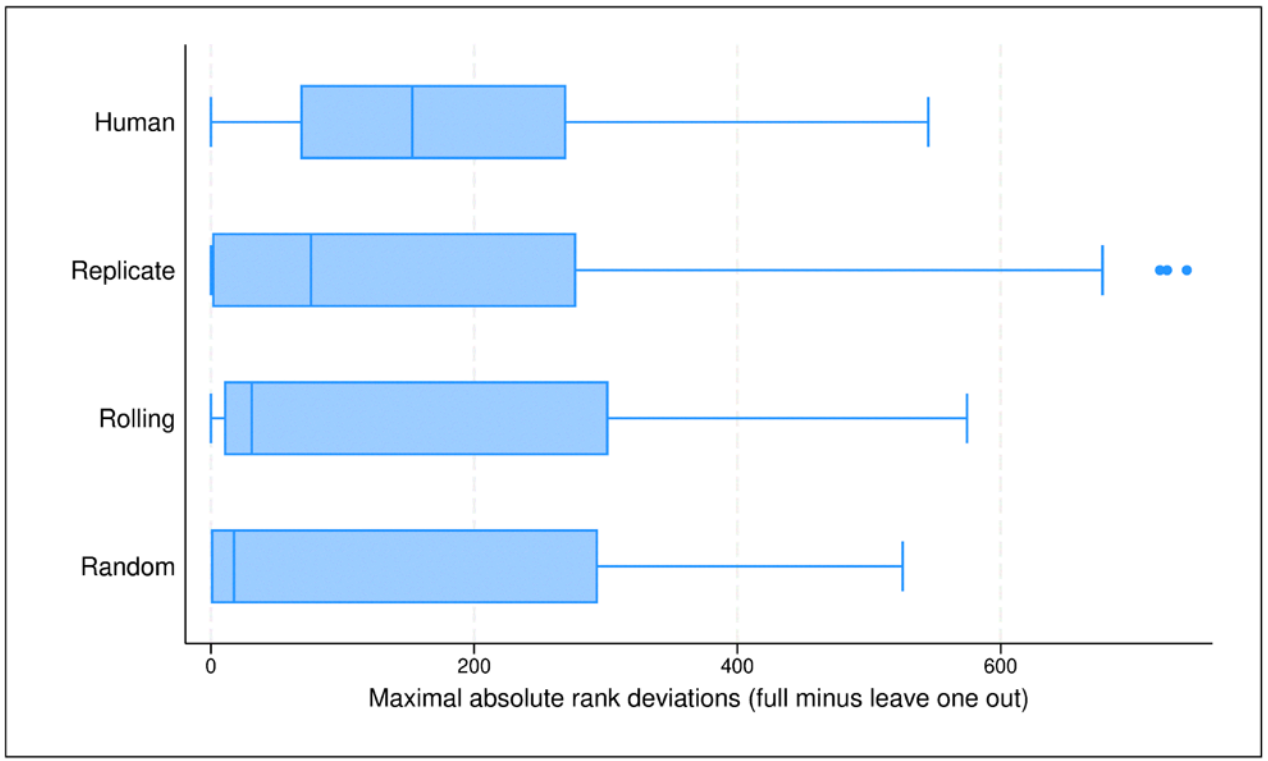
You Have Just Seen ...

Design and analysis for ranking of human- and machine-rated applications to a veterinary program
presented by Henrik Stryhn (`http://stryhnstatistics.ca`)

Thank you for your attention!

Leave-one-out Analysis Results III

Comparisons between human rating and machine rating designs for writing quality item:[9]



- the replicate ("same") design is less robust than the rolling and random machine rating designs.

[9] 209 (out of 778) applicants with two human and two machine ratings in three machine-grading designs, maximal absolute deviations.

Sample Mixed Model Results II

- the Veterinary Medicine content item analysed separately,
- linear mixed model variance components unless indicated otherwise;

Human rating:

	outcome scale	
parameter	linear	ordinal
σ^2(subject)	0.26 (.10)	0.23 (.11)
σ^2(error)	1.17 (.11)	1 (.)
$(\sigma^2(\text{rater}))^a$	1.04 (.42)	1.07 (.)

[a] for comparison only, modelled by fixed effects

- improved fit with unequal rater variances: $\Delta \log L \approx 71$ (df$=$11),
- close to perfect agreement between subject estimates in linear and ordinal models: $r = 0.99-1.00$ (BLUPs/rankings).

Machine rating:

	adjudication sets		
parameter	replic	rolling	random
σ^2(subject)	0.85 (.04)	0.47 (.03)	0.50 (.03)
σ^2(error)	0.05 (.002)	0.35 (.01)	0.35 (.02)
σ^2(adj.set)	0 (.)	0.04 (.014)	0.03 (.01)

- small variance component for adjudication sets $\sim$ small (positive) within-set correlation: $\hat{\rho} = 0.046$ (rolling).
- fair agreement only between subject estimates and rankings across designs: $r = 0.64-0.82$.

Overall, the agreement between human and machine ratings is very low: $r \approx 0.20-0.27$ (BLUPs/rankings).

Sample Mixed Model Results III

- the University content ($\sim$ diversity) item analysed separately,
- linear mixed model variance components unless indicated otherwise;

Human rating:

	outcome scale	
parameter	linear	ordinal
σ^2(subject)	1.09 (.15)	0.92 (.22)
σ^2(error)	1.24 (.12)	1 (.)
$(\sigma^2(\text{rater}))^a$	0.99 (.40)	1.02 (.)

[a] for comparison only, modelled by fixed effects

- improved fit with unequal rater variances: $\Delta \log L \approx 24$ (df$=11$),
- close to perfect agreement between subject estimates in linear and ordinal models: $r = 0.98-0.99$ (BLUPs/rankings).

Machine rating:

	adjudication sets		
parameter	replic	rolling	random
σ^2(subject)	1.48 (.08)	0.83 (.05)	0.84 (.05)
σ^2(error)	0.09 (.003)	0.62 (.03)	0.57 (.01)
σ^2(adj.set)	0 (.)	0.03 (.01)	0.02 (.01)

- small variance component for adjudication sets $\sim$ small (positive) within-set correlation: $\hat{\rho} = 0.026$ (.006).
- fair agreement only between subject estimates and rankings across designs: $r = 0.68-0.86$.

Overall, the agreement between human and machine ratings is low:
$r \approx 0.36-0.49$ (BLUPs/rankings).