



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Adaptive Comparative Judgement: A Tool to Support Students' Assessment Literacy

Citation for published version:

Rhind, SM, Hughes, KJ, Yool, D, Shaw, D, Kerr, W & Reed, N 2017, 'Adaptive Comparative Judgement: A Tool to Support Students' Assessment Literacy' *Journal of Veterinary Medical Education*, vol 44, no. 4, pp. 686-691. DOI: 10.3138/jvme.0616-113R

Digital Object Identifier (DOI):

[10.3138/jvme.0616-113R](https://doi.org/10.3138/jvme.0616-113R)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Peer reviewed version

Published In:

Journal of Veterinary Medical Education

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



1 **Adaptive Comparative Judgement: A Tool to Support Students' Assessment Literacy**

2

3 **Susan M. Rhind.** Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter
4 Bush Veterinary Centre, Roslin, Midlothian, EH25 9RG. BVMS, PhD, FRCPath, PFHEA,
5 MRCVS, Director of Veterinary Teaching, Chair of Veterinary Medical Education and
6 Assistant Principal (Assessment and Feedback). Specific interests in assessment and
7 feedback, e-learning, curriculum development and student well-being/ support. E-mail:
8 susan.rhind@ed.ac.uk.

9 **Kirsty J. Hughes.** Royal (Dick) School of Veterinary Studies, University of Edinburgh,
10 Easter Bush Veterinary Centre, Roslin, Midlothian, EH25 9RG. BVM&S, BSc, MSc, PhD,
11 FHEA, MRCVS, Research Assistant in Veterinary Medical Education. Specific interests
12 include assessment and feedback, e-learning, the student experience and staff
13 development.

14 **Donald Yool.** Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter
15 Bush Veterinary Centre, Roslin, Midlothian, EH25 9RG. BVMS, PhD, DipECVS, CertSAS,
16 MRCVS is Head of Companion Animal Soft Tissue Surgery Specific interests in assessment
17 and in use of digital media in supporting undergraduate surgical teaching.

18 **Darren Shaw.** Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter
19 Bush Veterinary Centre, Roslin, Midlothian, EH25 9RG. BSc, PhD, Senior Lecturer in
20 Epidemiology. Specific interests in quantitative epidemiology of zoonotic / purely veterinary
21 diseases as well as the clinical epidemiology associated with animal health that is of
22 veterinary importance.

23 **Wesley Kerr.** Information Services at the University of Edinburgh, 19 Buccleuch Place,
24 Edinburgh EH8 9LN. GRSC, BSc, PGCert. Senior eLearning advisor with the Educational
25 Design and Engagement team. Specific interests include assessment, feedback and student
26 engagement.

27 **Nicki Reed.** Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush
28 Veterinary Centre, Roslin, Midlothian, EH25 9RG. BVM&S, Cert VR, DSAM (Feline),
29 Diploma ECVIM-CA, MRCVS. Senior Lecturer in Internal Medicine. Specific interest in
30 teaching clinical application of theoretical knowledge. *NR's current address is Lumbry Park
31 Veterinary Specialists, Lumbry Park, Selborne Road, Alton, Hampshire GU34 3HF.

32

33

34 **Abstract**

35 Comparative judgement in assessment is a process whereby repeated comparison of two
36 items (e.g. assessment answers) can allow an accurate ranking of all the submissions to be
37 achieved. In Adaptive Comparative Judgement (ACJ), technology is used to automate the
38 process and present pairs of pieces of work over iterative cycles. An on-line ACJ system
39 was used to present students with work prepared by a previous cohort at the same stage of
40 their studies. Objective marks given to the work by experienced faculty were compared to
41 the rankings given to the work by a cohort of veterinary students (n=154). Each student was
42 required to review and judge 20 answers provided by the previous cohort to a free text short
43 answer question. The time that students spent on the judgement tasks was recorded and
44 students were asked to reflect on their experiences after engaging with the task. There was
45 a strong positive correlation between student ranking and faculty marking. A weak positive
46 correlation was found between the time students spent on the judgements and their
47 performance on the part of their own examination which contained questions in the same
48 format. Slightly less than half of the students agreed that the exercise was a good use of
49 their time, but 78% agreed that they had learnt from the process. Qualitative data highlighted
50 different levels of benefit from the simplest aspect of learning more about the topic to an
51 appreciation of the more generic lessons to be learned.

52

53 **Key Words: assessment literacy, peer assessment, feedback**

54

55 **Background**

56 The concept of comparative judgement in assessment is not new, having first been
57 described by Thurstone in 1927 ¹ as a process whereby repeated comparison of two items, if
58 carried out a sufficient number of times across all the items, can allow an accurate ranking of
59 these items to be achieved. In 2012, Pollit, extended this concept to one of 'Adaptive
60 Comparative Judgement' (ACJ) by describing a system where technology can be used to
61 automate the operation of the underlying algorithm and present judges with pairs of pieces of
62 work for comparison over iterative cycles ². Through these iterative cycles of judgement
63 rather than assigning of marks to pieces of work, all the items are ultimately sorted into a
64 rank order ³. It is then possible for assessors to agree an aligned marking scheme if required
65 or simply provide students with the rank ordering as feedback on their performance. Whilst

66 ACJ has mainly been described in the context of faculty judging students work, there is also
67 the potential to explore the principle in the context of peer assessment.

68 It has been shown that peer assessment, if well supported can help students develop their
69 own ability to self-assess⁴. By giving students the opportunity to act as judges, there are
70 potential additional benefits to them in terms of increasing their skills in understanding the
71 range of quality in other students work. Smith et al (2013) presented a study with business
72 students showing that the development of students' ability to judge standards of
73 performance on student work correlated with enhanced marks⁵. This type of activity is being
74 used in other disciplines in Higher Education to build students 'assessment literacy' skills, a
75 term encompassing the range of knowledge, skills and attributes necessary to understand
76 the purpose and process of assessment^{6,7}.

77 Developing skills in self-assessment requires practice, time and the opportunity to consider
78 the strengths and weaknesses of a range of pieces of assessed work. This aspect has been
79 emphasized by Sadler who in particular has championed the notion of provision of
80 'substantial evaluative experience' as a core part of the design of a curriculum⁸. Boud has
81 also published frequently in this area, emphasizing the importance of students becoming
82 assessors in order to fully understand the nature of good quality work⁹⁻¹¹.

83 In the study reported here, we sought to compare the objective marks (and implicitly the
84 ranking) given to pieces of work by experienced faculty to the rankings given to the work by
85 a larger group of veterinary students who had recently learned the material relevant to the
86 question. The students were engaging in the ranking process as part of a curriculum
87 intervention to build skills in assessment literacy and allow them an explicit opportunity to
88 experience a range of pieces of work of different quality from a previous cohort at the same
89 stage in the curriculum.

90 In carrying out this study we sought to explore the following hypotheses:

- 91 1. Students' abilities as markers will positively correlate with actual faculty marks given
92 to pieces of work.
- 93 2. Students' will find engaging with a range of pieces of work of differing standards
94 helpful beyond the context of the specific course in the study.

95

96 **Methods**

97 **Context**

98 The context of this study was a small animal medicine course in the second or third year of a
99 veterinary degree programme. Students in the cohort under study included students with
100 previous degrees on an accelerated 4 year programme and students on a five year
101 programme entering straight from high school. The cohort size was 154. The course is
102 examined by in course assessment) and an end of course degree examination comprising
103 MCQs and short answer questions.

104

105 ***Ranking activity***

106 The answers to an examination question from the previous cohort (n=162) were transcribed
107 and entered into a commercially available system (Digital Assess^a) which allows pairs of
108 pieces of work to be presented to assessors who then compare the work and select which of
109 the 2 is better using the principle of Adaptive Comparative Judgement. As the students work
110 through their comparisons, the system fine tunes the comparisons to focus on those that are
111 most similar and ultimately presents a complete rank order of all pieces of work with an
112 associated reliability statistic. The system also collects data on the amount of time spent on
113 both each individual judgement and also overall time spent per student on the judging
114 process.

115 Students were given an introduction to the system and an explanation of the question they
116 were assessing and an outline expected answer for the question in a plenary session at the
117 start of a lecture. The question had previously been marked by a single faculty member who
118 taught the material and given a mark out of 10. The question selected for the exercise was
119 chosen due to the spread of marks given (consistent with a range of answer quality). The
120 overall mark profile is shown in Figure 1 (Mean mark 4.7, median 5). The question used is
121 shown in Box 1.

122 [Insert Figure 1 here]

123

You have made a clinical diagnosis of allergic skin disease in a three year old, male West Highland white terrier called Angus. You have decided to start Angus on a food trial to determine whether a protein in his diet is contributing to his pruritus and inflammation. Angus is 6/10 pruritic but your cytology samples did not reveal any evidence of bacterial or Malassezia skin infections.

- A) **LIST** the different options for an appropriate diet in this case. (3 marks)
- B) Briefly discuss how to select an appropriate food for Angus. (3 marks)
- C) Briefly discuss how the diet trial should be conducted to maximise compliance and a diagnostic outcome. (4 marks)

124

125 Box 1

126

127 Students were given a 2 week period to perform 10 parallel judgements using the on-line
128 system i.e. each student made judgements on a total of 20 pieces of work. Throughout the
129 rounds of comparisons the system calculates a reliability statistic (between 0 and 1) which
130 was recorded.

131

132 ***Correlation of Faculty and Student Judgements***

133 The final output from the ACJ system ranks all pieces of work in order based on the
134 judgements made. To compare faculty and student opinions on the work, a correlation of the
135 actual marks given by faculty was carried out against the rank order using Spearman Rank
136 correlation.

137

138 ***Performance on the Degree Examination***

139 The final degree examination marks were correlated against the time students spent
140 comparing questions and completing their judgement.

141

142 **Survey**

143 Following the ranking exercise, students were asked to complete a short on-line
144 questionnaire on their experiences comprising a mixture of Likert scale and open ended
145 questions. Responses to open ended questions were grouped into themes.

146

147 Ethical approval for this study was sought and received from the College of Medicine and
148 Veterinary Medicine's Committee for the use of student volunteers at the University of
149 Edinburgh (Ref: 2014/24)

150

151 **Results**

152 ***Reliability***

153 The reliability statistic at the end of the 20 rounds of judgement was 0.98 indicating a high
154 level of reliability.

155

156 ***Time spent on judgement***

157 The average time taken per judgement was 139 Seconds; standard deviation 83. Median
158 126 seconds (Figure 2).

159

160 [Insert Figure 2 here]

161

162 ***Student judging compared to faculty judging***

163 There was a positive correlation between faculty mark and ranking by students (0.690,
164 $p < 0.001$, Figure 3)

165

166 [Insert Figure 3 here]

167

168 ***Performance on Degree Examination with Time spent on Ranking Task***

169 The degree examination was divided into multiple choice elements and short answer and
170 then correlated to the total and average time spent making judgements (Table 1).

171

172 [Insert table 1 here]

173 **Student Evaluation**

174 The survey was completed by 67 of the total 154 students (44% response rate). 68% of
175 students were extremely positive or positive about their experience with the software, 24%
176 quite positive and 8% not at all positive.

177 Responses to a series of Likert scale questions on the intervention are shown in Table 2.

178 [Insert Table 2 here]

179

180 Although the majority (>75%) agreed or strongly agreed that they had learned from what
181 their peers had submitted, slightly less than 50% agreed that it was a good use of their time.
182 A similar percentage (49%) indicated that they would recommend the tool for use in other
183 courses.

184 Students were asked whether reviewing other people's answers had made them think
185 differently about how they answered questions. 69% answered yes, 31% answered no.

186 Students were then asked to explain their answer in more detail. Themes in this free text
187 could be broadly grouped into those relating to the benefit being around revision of the
188 particular topic and more generic benefits.

189

190 Learning more about the specific topic:

191

192 'Learned about the topic being addressed and also the type of question that can be
193 asked and the answer expected from us.' S4

194 'I learned about the topic while doing the exercise and it was really interesting to see
195 other peoples answers.' S10

196

197 Many more comments related to the broader learning around assessment and approach in
198 general; particularly where it highlighted good practice

199 'It was interesting to see different ways that students laid out their work. It was also
200 interesting to see different angles from which students approached the questions.' S3

201 'Some people approach things in a very different way to me - that was interesting'
202 S25

203

204 And this point was extended by several students to indicate that reflection on this had helped
205 them with their future approach to similar style of questions:

206 'I am typically someone who waffles during exams. While reading the various
207 answers, I was able to pick up useful tips on keep the answers short, simple and to
208 the point.' S7

209 'seeing some of the answers was definitely a bit of "What not to do!" S2

210 'It reminded me of the need to use appropriate terms/language during writing
211 answers and that concise, well-informed answers are better than long rambles. S18

212

213 Linked to this were several comments on the benefit of being given the examiner perspective

214 'This has certainly given me a level of sympathy for graders' S16

215 'Now I know which methods are easier for the professors to grade.' S20

216

217 The main negative issue raised by students in the data was the focus on one question only
218 with students expressing a desire to have more examples of different subject areas to
219 review.

220 'It would have been much better if every question was a different question instead of
221 the same one 10 times as this just became boring. And we would have learnt more if
222 different questions were used.' S2

223

224 **Limitations**

225 The exercise was carried out in the context of one course in one school and only utilised one
226 question. However the short answer question format is commonly used in many contexts

227 and schools so results are likely to be of relevance to other educators interested in
228 assessment of free text format. The system of evaluation used was different for faculty and
229 students however it was impractical to have the answers ranked for a second time by the
230 same individual using the ACJ system. The response rate to the survey of 44% is a further
231 limitation.

232

233 **Discussion**

234 The aim of this study was to facilitate student exposure to work of different quality and
235 engage them in the process of assessing peers work as part of a series of interventions to
236 help support development of assessment literacy skills and understanding of different
237 standards in assessment. We aimed to explore the hypotheses that students' abilities as
238 markers would positively correlate with actual faculty marks given to pieces of work on a
239 subject they had recently covered in the curriculum and that students would find engaging
240 with a range of pieces of work of differing standards helpful beyond the context of the
241 specific course in the study.

242

243 We have previously shown that students abilities as judges when assigning marks to work is
244 variable ⁷. Although the correlation between the students' ranking and faculty marks was
245 strong in this study, there were clearly outliers where students had ranked an answer
246 markedly different from faculty. Although not possible to explore in the current study design
247 (because the ranking data generated is a cumulative estimate and is not associated with an
248 individual student), it seems a reasonable hypothesis that the students who were more
249 accurate in their ranking, were more academically able.

250 Student feedback on using the system was mixed with less than half agreeing it was a good
251 use of their time and recommending the exercise for future classes. Despite this,
252 approximately two thirds of students did agree that they had learnt from the process. It would
253 be interesting to explore this dichotomy in future studies through more detailed qualitative
254 approaches such as focus groups. Qualitative data highlighted different levels of benefit from
255 the simplest aspect of learning more about the topic to those who exhibited more
256 metacognitive reflection on the process appreciating the more generic lessons to be learned
257 from such a process.

258 Furthermore there was a positive correlation between those who spent longer on the
259 judgements and their performance on the part the examination that the exercise was
260 designed to support i.e. free text short answer questions. Whilst this may be a simple

261 reflection of the more conscientious, able and engaged students spending longer on the
262 task, if we consider the performance on the MCQ part of the examination a 'control' (in terms
263 of a measure of knowledge not requiring skills in description of diagnostic/therapeutic
264 approach), then this suggests that the benefit was more of a direct relation to time spent
265 improving their abilities in the short answer section of the examination. This is consistent
266 with a study reported by Li and Gao,¹² who showed that students who conducted peer
267 assessment performed significantly better than students who did not.

268 The main negative issue described by the students in the qualitative data was around the
269 focus on one question only with students expressing a desire to have more examples of
270 different subject areas to review. Whilst this would have allowed them to review different
271 topics, the main aim of this study was to give the students insights into differing approaches
272 to answering questions and differing quality work rather than revision of content per se.
273 However given the student feedback discussed earlier, having a system which allowed
274 review of a larger number of topics may well have added to the perceived value of the
275 exercise from the student perspective.

276 The benefits of peer assessment have been demonstrated elsewhere but often in the
277 context of students marking or giving feedback on one or two other pieces of work¹³. The
278 advantage of the system reported in this study is the ability to electronically facilitate access
279 to a larger number of answers and therefore a wider range of work of differing quality. This
280 approach is consistent with the model of self-regulated learning described by Nicol and
281 Macralane-Dick¹⁴ which highlights the importance of assessment strategies which facilitate
282 self and peer assessment and help clarify what good performance is.

283 In conclusion, there is evidence that those students who spent longer on carrying out the
284 judgements benefitted when it came to the short answer section of the assessment. Whilst it
285 is not possible to state definitely that this is causative, the lack of correlation with the other
286 aspect of the assessment (the MCQ) suggests it gave these students an advantage. As the
287 survey was anonymous it is not possible to say whether those who enjoyed and engaged
288 with the process more were the ones who benefitted more.

289 This study provides further support for the utility of assessment literacy interventions in
290 helping students understand more about the assessment process, quality, standards and the
291 challenges assessors face. In particular, the ability of such interventions to encourage
292 students to reflect on their own assessment practice and learn from others was highlighted
293 as a major benefit.

294

295

296 **References**

297 ^a <http://digitalassess.com/>

298

299

- 300 1 Thurstone LL. A law of comparative judgment. *Psychological Review* 34:273-86 1927.
- 301 2 Pollitt A. Comparative judgement for assessment. *International Journal of Technology and*
- 302 *Design Education* 22(2):157-70, 2012.
- 303 3 Whitehouse C, Pollitt A. Using Adaptive Comparative Judgement to Obtain a Highly
- 304 Reliable Rank Order in Summative Assessment. <
- 305 https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP_RP_CW_20062012_2.pdf >. The
- 306 Assessment and Qualifications Alliance (AQA), 2012.
- 307 4 Reinholz D. The assessment cycle: a model for learning through peer assessment.
- 308 *Assessment & Evaluation in Higher Education* 41(2):301-15, 2016.
- 309 5 Smith CD, Worsfold K, Davies L, Fisher R, McPhail R. Assessment literacy and student
- 310 learning: the case for explicitly developing students 'assessment literacy'. *Assessment &*
- 311 *Evaluation in Higher Education* 38(1):44-60, 2013.
- 312 6 Price M, Rust C, O'Donovan B, Handley K, Bryant R. *Assessment Literacy: The*
- 313 *Foundation for Improving Student Learning*. Oxford: The Oxford Centre for Staff and
- 314 Learning Development; 2012.
- 315 7 Rhind SM, Patterson J. Assessment Literacy: Definition, Implementation, and Implications.
- 316 *Journal of Veterinary Medical Education* 42(1):28-35, 2015.
- 317 8 Sadler DR. Beyond feedback: developing student capability in complex appraisal.
- 318 *Assessment & Evaluation in Higher Education* 35(5):535-50, 2010.
- 319 9 Boud D, Falchikov N. Quantitative studies of student self-assessment in higher-education -
- 320 a critical analysis of findings. *Higher Education* 18(5):529-49, 1989.
- 321 10 Boud D, Falchikov N. Aligning assessment with long term learning. *Assessment &*
- 322 *Evaluation in Higher Education* 31(4):399-413, 2006.
- 323 11 Boud D, Molloy E. Rethinking models of feedback for learning: the challenge of design.
- 324 *Assessment & Evaluation in Higher Education* 38(6):698-712, 2013.
- 325 12 Li L, Gai F. The effect of peer assessment on project performance of students at different
- 326 learning levels, . *Assessment & Evaluation in Higher Education*, 2015.
- 327 13 Mostert M, Snowball JD. Where angels fear to tread: online peer-assessment in a large
- 328 first-year class. *Assessment & Evaluation in Higher Education* 38(6):674-86, 2013.
- 329 14 Nicol DJ, Macfarlane-Dick D. Formative assessment and self-regulated learning: a model
- 330 and seven principles of good feedback practice. *Studies in Higher Education* 31(2):199-218,
- 331 2006.

332

333 **Figure legends**

334

335 Figure 1

336 Range of marks given by faculty to the question selected for the adaptive comparative

337 judgement exercise.

338

339 Figure 2

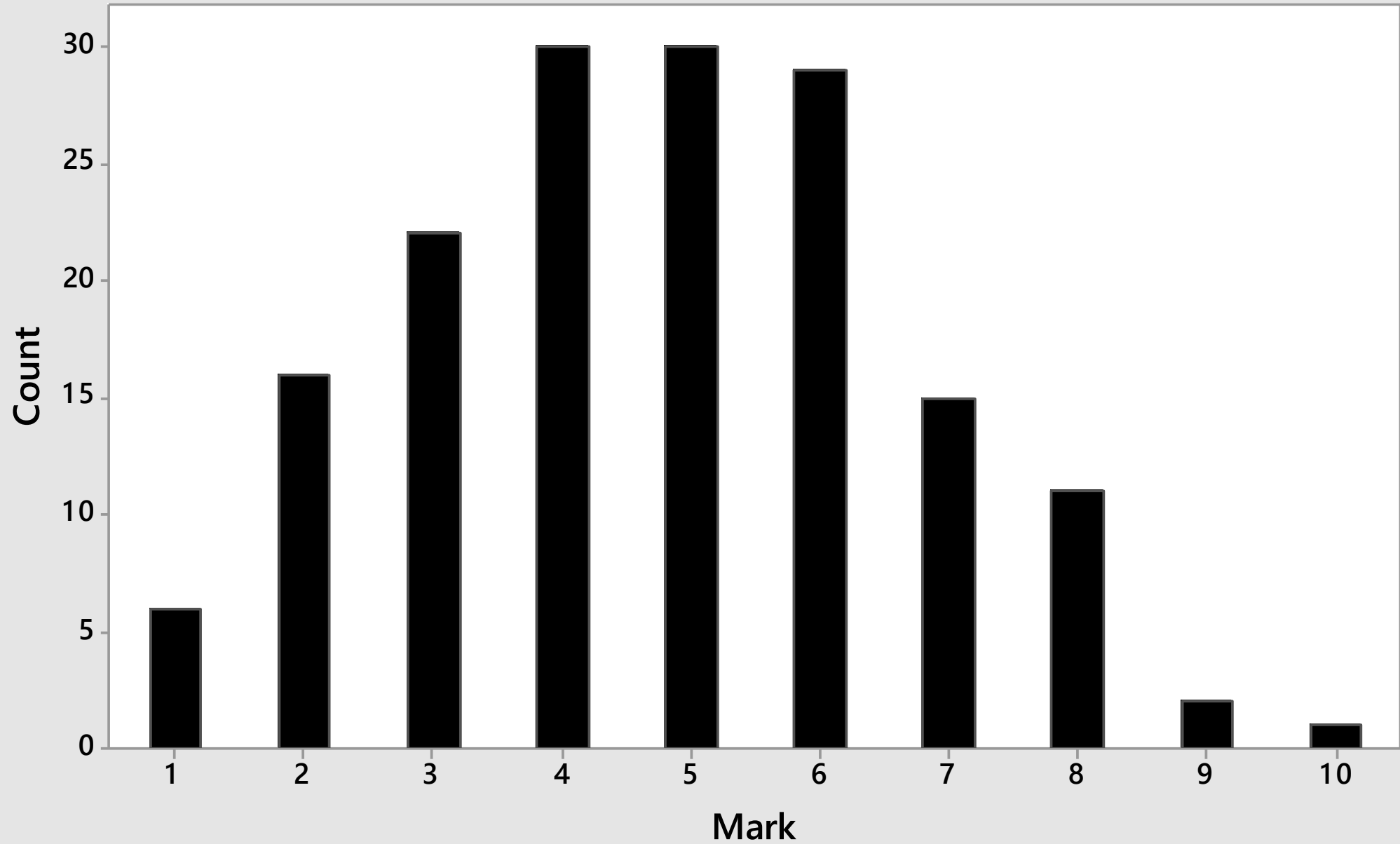
340 Average time in seconds, spent per student on the judging process captured by the on-line
341 system

342

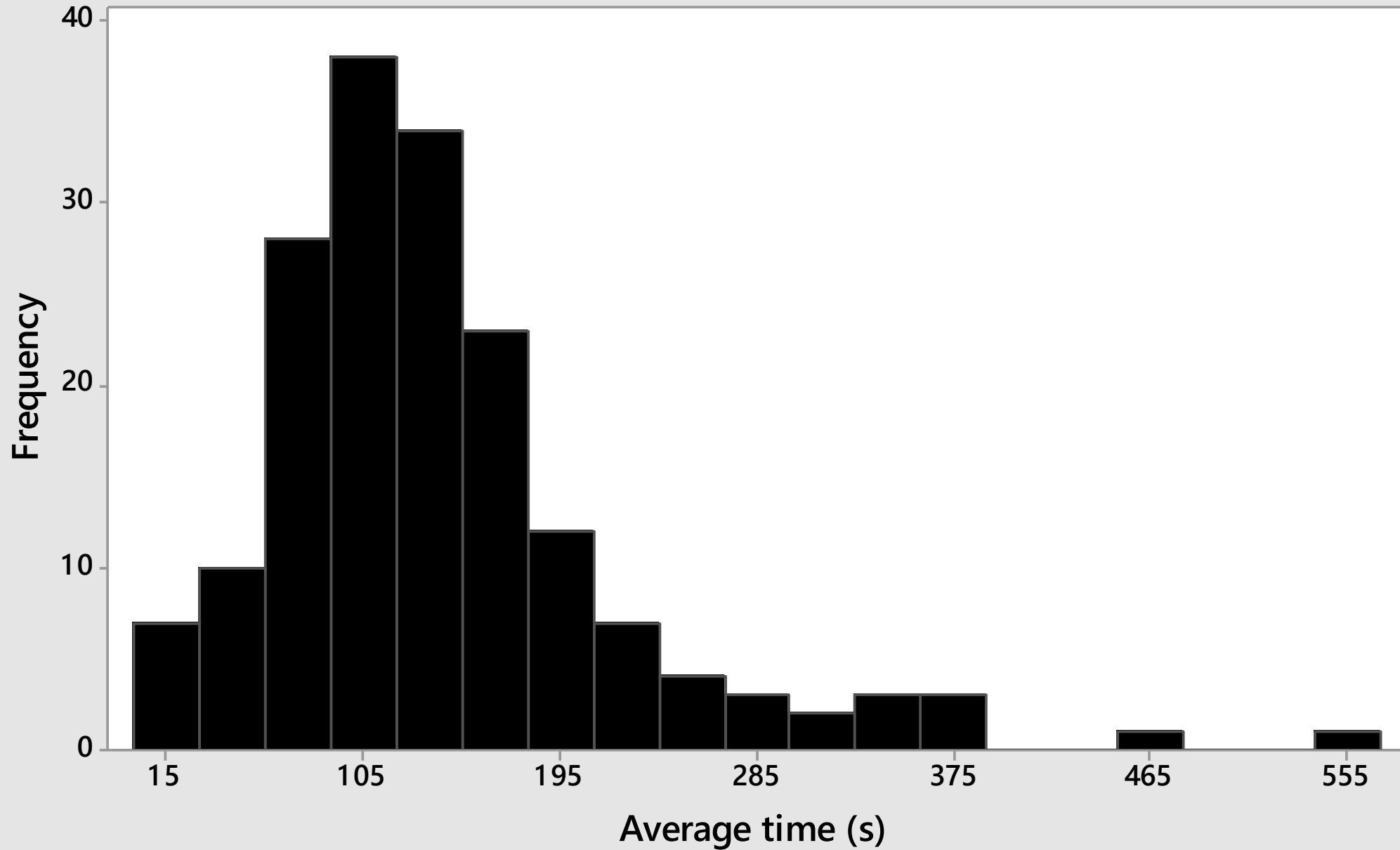
343 Figure 3

344 Correlation of the ranking produced from the cumulative student judgements with the actual
345 faculty marks given to the pieces of work. Pearson correlation = 0.696, P-Value = 0.000

Faculty Mark out of 10



Histogram of Average time (s)



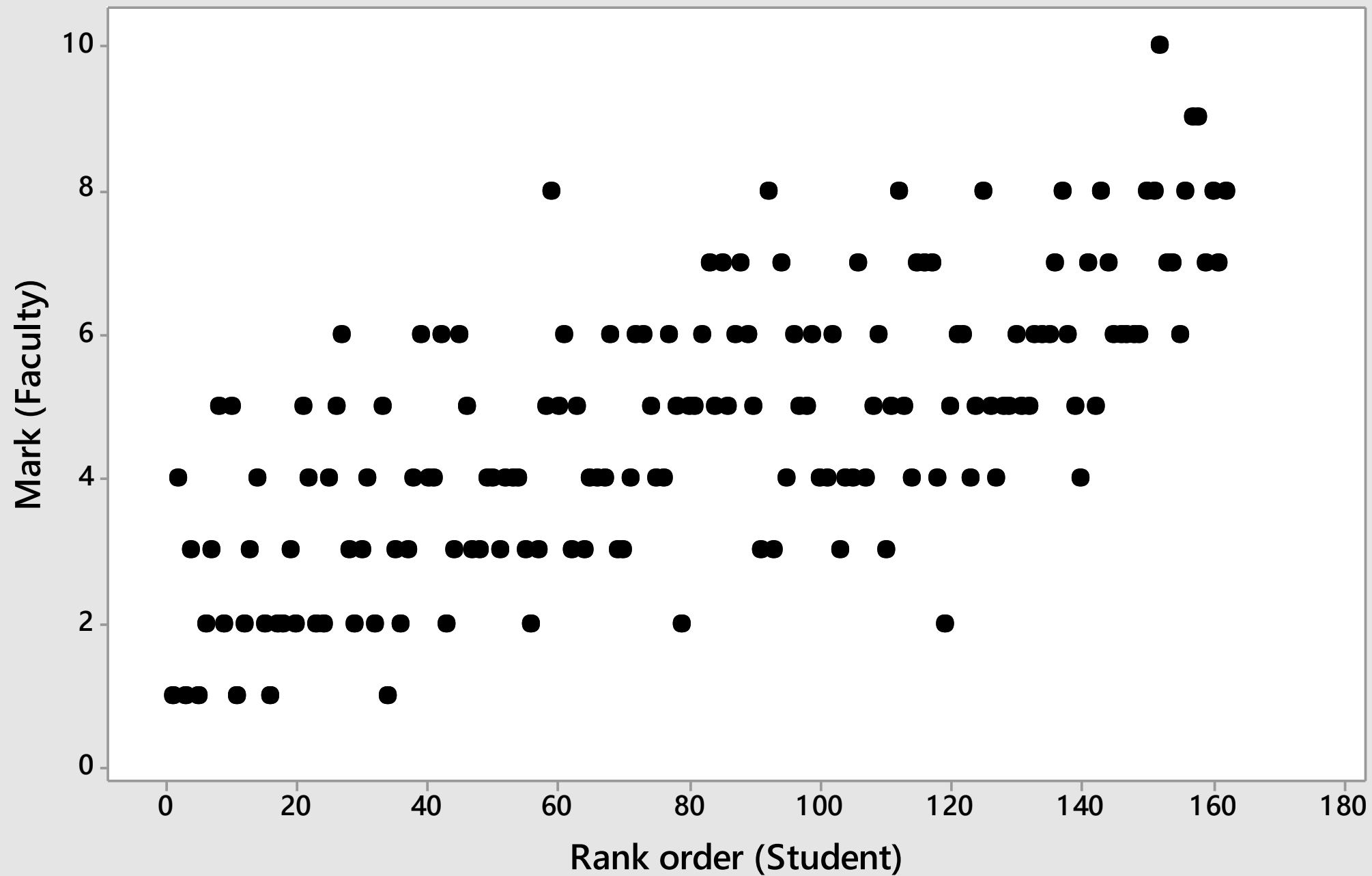


Table 1

Correlation			p-value
Total time spent	MCQ result	0.018	0.153
Average time spent	MCQ result	0.120	0.145
Total time spent	SAQ exam result	0.193*	0.018
Average time spent	SAQ exam result	0.195*	0.017

Spearman rank correlations and associated P values between total and average time students spent making judgements and their performance on the multiple choice (MCQ) and short answer (SAQ) sections of the examination. Significant correlations are asterisked ($p < 0.05$)

Table 2

Survey Statement	Strongly Disagree	Disagree	No Strong Feelings	Agree	Strongly Agree
The judging process was enjoyable	4.5	7.5	46.3	37.3	4.5
I have learned from reading what my peers submitted	3	10.4	9	62.7	14.9
This has helped me to understand what markers are looking for	4.5	10.4	16.4	53.7	14.9
This was a good use of my time	4.5	19.4	26.9	49.3	0
I would recommend this for other courses	6	14.9	29.9	43.3	6

Percentage responses to a series of Likert scale questions from the post intervention questionnaire (n=67).