The influence of rater training on inter-rater reliability when using the rat grimace scale

Emily Zhang [1] , **Vivian Leung** [2] , **Daniel SJ Pang** [Corresp. 2]

[1] Western College of Veterinary Medicine, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
[2] Department of Clinical Sciences, Université de Montréal, St-Hyacinthe, QC, Canada

Corresponding Author: Daniel SJ Pang
Email address: danielpang17@hotmail.com

Background. Rodent grimace scales facilitate evaluation of the affective component of pain and can identify a range of acute pain levels. Reported rater training in the use of these scales varies considerably and may contribute to observed variability in inter-rater reliability. This study evaluated the effect of training on inter-rater reliability with the Rat Grimace Scale (RGS). Methods. Two training sets, of 42 and 150 images, were prepared from several acute pain models. Four trainee raters, with no previous experience with the RGS, progressed through 2 rounds of training, first scoring 42 images (S1) followed by 150 images (S2a). After each round, trainees reviewed the RGS and any problematic images with an experienced rater. The 150 images were re-scored in a final round (S2b). Inter-rater reliability was evaluated using the intra-class correlation coefficient (ICC) and ICCs compared with a Feldt test. Results. Inter-rater reliability increased from moderate (ICC 0.58 [95%CI: 0.43-0.72]) to very good (ICC 0.85 [0.81-0.88]) between S1 and S2b ($p < 0.01$) with a significant increase also observed between S2a and S2b ($p < 0.01$). The ICCs for individual action units orbital tightening, ears and nose/cheek also improved from S1 to S2b ($p < 0.01$). The action units with the highest and lowest ICCs at S2b were orbital tightening (0.84 [0.80-0.87]) and whiskers (0.63 [0.57-0.70]), respectively. In comparison to an experienced rater the ICCs for all trainees improved, ranging from 0.88 to 0.91 at S2b. Discussion. Training improves inter-rater reliability between trainees, with an associated reduction in 95%CI. Additionally, training resulted in improved inter-rater reliability alongside an experienced rater. Training improves the scoring of individual action units though scoring of whiskers is more difficult that other sites. Conclusion. The beneficial effects of training potentially reduce data variability and improve experimental animal welfare.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26721v1 | CC BY 4.0 Open Access | rec: 17 Mar 2018, publ: 17 Mar 2018

The influence of rater training on inter-rater reliability when using the Rat Grimace Scale

Emily Zhang, Vivian Leung, Daniel SJ Pang
Veterinary Clinical and Diagnostic Sciences, Faculty of Veterinary Medicine, University of Calgary, Alberta, Canada.

Current addresses:
Zhang - Western College of Veterinary Medicine, University of Saskatchewan, Saskatoon, Saskatchewan, Canada.
Leung & Pang - Faculty of Veterinary Medicine, Université de Montréal, Saint-Hyacinthe, Québec, Canada.

Corresponding author:
Daniel SJ Pang
E: daniel.pang@umontreal.ca

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26721v1 | CC BY 4.0 Open Access | rec: 17 Mar 2018, publ: 17 Mar 2018

Abstract

Background. Rodent grimace scales facilitate evaluation of the affective component of pain and can identify a range of acute pain levels. Reported rater training in the use of these scales varies considerably and may contribute to observed variability in inter-rater reliability. This study evaluated the effect of training on inter-rater reliability with the Rat Grimace Scale (RGS).

Methods. Two training sets, of 42 and 150 images, were prepared from several acute pain models. Four trainee raters, with no previous experience with the RGS, progressed through 2 rounds of training, first scoring 42 images (S1) followed by 150 images (S2a). After each round, trainees reviewed the RGS and any problematic images with an experienced rater. The 150 images were re-scored in a final round (S2b). Inter-rater reliability was evaluated using the intra-class correlation coefficient (ICC) and ICCs compared with a Feldt test.

Results. Inter-rater reliability increased from moderate (ICC 0.58 [95%CI: 0.43-0.72]) to very good (ICC 0.85 [0.81-0.88]) between S1 and S2b ($p < 0.01$) with a significant increase also observed between S2a and S2b ($p < 0.01$). The ICCs for individual action units orbital tightening, ears and nose/cheek also improved from S1 to S2b ($p < 0.01$). The action units with the highest and lowest ICCs at S2b were orbital tightening (0.84 [0.80-0.87]) and whiskers (0.63 [0.57-0.70]), respectively. In comparison to an experienced rater the ICCs for all trainees improved, ranging from 0.88 to 0.91 at S2b.

Discussion. Training improves inter-rater reliability between trainees, with an associated reduction in 95%CI. Additionally, training resulted in improved inter-rater reliability alongside an experienced rater. Training improves the scoring of individual action units though scoring of whiskers is more difficult that other sites.

Conclusion. The beneficial effects of training potentially reduce data variability and improve experimental animal welfare.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26721v1 | CC BY 4.0 Open Access | rec: 17 Mar 2018, publ: 17 Mar 2018

Introduction

The effectiveness of a pain assessment scale lies in its validity (does a scale measure what is intended) and reliability (measurement error). Rodent grimace scales have renewed interest in measuring the affective component of pain and have been promoted as a means of overcoming the shortfalls of nociceptive threshold testing (Mogil & Crager, 2004; Langford, Bailey & Chanda et al., 2010; Sotocinal, Sorge & Zaloum et al., 2011; Oliver, De Rantere & Ritchie et al., 2014; De Rantere, Schuster & Reimer et al., 2016). There is increasing evidence that grimace scales discriminate painful and non-painful states in a range of acute pain models and interventions (Langford, Bailey & Chanda et al., 2010; Sotocinal, Sorge & Zaloum et al., 2011; Oliver, De Rantere & Ritchie et al., 2014; De Rantere, Schuster & Reimer et al., 2016; Leach, Klaus & Miller et al., 2012). However, there are conflicting reports regarding reliability when multiple raters score images (Langford, Bailey & Chanda et al., 2010; Sotocinal, Sorge & Zaloum et al., 2011; Oliver, De Rantere & Ritchie et al., 2014; Faller, McAndrew & Schneider et al., 2015; Mittal, Gupta & Lamarre et al., 2016). Factors contributing to this variability may include a lack of structured training and variation in individual learning curves (Campbell, Hecker & Biau et al. 2014; de Oliveira Filho, 2002; Roughan & Flecknell, 2006).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26721v1 | CC BY 4.0 Open Access | rec: 17 Mar 2018, publ: 17 Mar 2018

It is unclear what level of training is required to attain proficiency in using grimace scales. Most studies include minimal, non-specific descriptions of training (Langford, Bailey & Chanda et al., 2010; Sotocinal, Sorge & Zaloum et al., 2011; Oliver, De Rantere & Ritchie et al., 2014; Leach, Klaus & Miller et al., 2012; Faller, McAndrew & Schneider et al., 2015; Mittal, Gupta & Lamarre et al., 2016) and few report any measure of reliability (Langford, Bailey & Chanda et al., 2010; Sotocinal, Sorge & Zaloum et al., 2011; Oliver, De Rantere & Ritchie et al., 2014; Mittal, Gupta & Lamarre et al., 2016). Trainees progress at different rates during training to achieve proficiency in a task (Mittal, Gupta & Lamarre et al., 2016; Campbell, Hecker & Biau et al. 2014; Roughan & Flecknell, 2006); therefore, in addition to training, some assessment of score reliability is necessary. The impact of training on scoring reliability with the Rat Grimace Scale (RGS) has not been formally evaluated. The objective of this study was to assess the effect of training on inter-rater reliability when scoring was performed with single and multiple raters applying the RGS. We hypothesized that training would improve inter-rater reliability.

Materials and Methods

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26721v1 | CC BY 4.0 Open Access | rec: 17 Mar 2018, publ: 17 Mar 2018

Two sets of training images were created from images collected during an unrelated project that had received institutional animal care and use committee approval from the University of Calgary Health Sciences Animal Care Committee (protocol IDs: AC13-0161 and AC13-0124)(De Rantere, Schuster & Reimer et al., 2016). This project used the following acute pain models: intraplantar carrageenan or Complete Freund's adjuvant or plantar incision. Animals were adult (> 10 weeks old) male Wistar (n = 34) rats, from a commercial source (Charles River Laboratories, Canada).

The methodology used to generate images was as previously described (Sotocinal, Sorge & Zaloum et al., 2011). Briefly, still images were captured from high-definition video-recordings and cropped so that only the face was visible. Each image was presented on a single slide in presentation software (Microsoft PowerPoint, version 14.0, Microsoft Corporation, Redmond, WA, USA). Slide order was randomized and identifying information (animal ID, time point, model) removed.

Images were selected based on image quality alone, by an individual not involved with the study. Two unique sets of training images were created, of 42 (S1) and 150 (S2) images. Images were scored using the RGS (scale range 0-2 for each action unit) and the average score calculated from four action units: orbital tightening, nose/cheek flattening, ear changes, and whisker change.

None of the 4 trainee raters recruited had previous experience with the RGS. All raters were female undergraduate and graduate students (age range 20-25 years), studying veterinary medicine, biology (n = 2) and health sciences and recruited when joining the research group as project students.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26721v1 | CC BY 4.0 Open Access | rec: 17 Mar 2018, publ: 17 Mar 2018

All raters followed the same scoring protocol: S1 images were scored independently by each individual, using the training manual provided by Sotocinal et al. (2011) (which contains prototypic images of each action unit at each score), with additional images from our laboratory (Supplementary Data_S1). Raters were encouraged to record comments for any images they found difficult to score. Following S1 scoring, raters reviewed their scores as a group with an experienced rater (DP), discussing recorded comments and areas of inconsistency. Images with the most variation between raters were selected for review. The primary goal of the discussion was to improve standardisation of scoring images assigned a score of 0 or 2. Disagreement in scores was tolerated provided differences between raters did not exceed 1 point on the scale. The standard of scoring was set by the experienced rater, following establishment of the technique within the laboratory with the support of the Mogil laboratory (McGill University). Once review of S1 scoring was complete, S2 images were scored independently by each individual and comments recorded as before (S2a). The S2 image set was then scored independently a second time (S2b) after a facilitated group discussion with the experienced rater (as per the S1 image set discussion). Approximately 15-30 images were reviewed during group discussions, with 2-3 weeks between reviews.

Intraclass correlation coefficients (ICCs, MedCalc version 12.6.1.0, MedCalc Software, Ostend, Belgium) were calculated to measure the reliability of RGS scoring between raters for the individual action unit scores and average RGS scores. An absolute model was used for the ICC calculation and single measure reported. This was done for each dataset (S1, S2a, S2b). ICCs were also calculated for the comparison between individual rater scores and those of the experienced rater (DP) to determine reliability of an individual rater. Calculated ICCs were compared with a Feldt test (critical F set at alpha = 0.01 and differences considered significant if the observed F value was greater than the critical F value) (Feldt, Woodruff & Salih, 1987; Kuzmic, 2015). Interpretation of the ICC followed the same divisions as used previously: ''very good'' (0.81–1.0), ''good'' (0.61–0.80), ''moderate'' (0.41–0.60), ''fair'' (0.21–0.40), ''poor'' (< 0.20)(Oliver, De Rantere & Ritchie et al., 2014). During the training process, raters were said to be proficient when calculated ICCs ± 95%CI overlapped with those published in a study reporting inter-rater reliability(Oliver, De Rantere & Ritchie et al., 2014). To assess the potential impact of scores memorised during group discussion between S2a and S2b introducing bias in to the ICC calculation for S2b, images with the greatest scoring variability at S2a (those with a difference of 2 points between any 2 raters and therefore the most likely to have been discussed) were removed and the ICCs for S2b recalculated. Data are presented as ICC (± 95%CI) and a corrected p value for multiple comparisons of ≤ 0.017 was considered significant.
Scoring accuracy was assessed by comparing scores for images collected at baseline and 6-9 hours after treatment (when a peak in RGS scores could be

expected for the models studied (De Rantere, Schuster & Reimer et al., 2016); paired t test with alpha set at 0.05). The datasets generated from this study are available in the Harvard Dataverse repository (Pang, 2018).

Results

Four raters completed the study. All training images were scored by every rater, and all scores included in the subsequent analysis.

Training was associated with a progressive improvement in inter-rater reliability and narrowing 95%CI (Fig. 1). The first training round (S1) resulted in a moderate ICC for the average RGS scores, with wide 95%CI (0.58 [0.43-0.72]). The increase in average RGS ICC between S1 and S2a (0.68 [0.58-0.76]) was not statistically significant ($F_{0.01;149,41} = 1.88$, observed F = 1.31, $p > 0.05$). A significant improvement was observed at S2b (0.85 [0.81-0.88]) compared with S1 (observed F = 2.8) and S2a ($F_{0.01;149,149} = 1.47$, observed F = 2.13, $p < 0.01$ for both comparisons). The resultant S2b ICC was classified as very good and comparable with published values (Fig. 1)(Oliver, De Rantere & Ritchie et al., 2014).

A similar pattern of improvement was observed in the scores of individual action units (Table 1). Significant increases in ICCs were observed between S1 and S2b for orbital tightening (observed F = 1.94), ear changes (observed F = 2.14) and nose/cheek flattening (observed F = 2.21, $p < 0.01$ all comparisons), but not whisker changes (observed F = 1.65, $p > 0.05$). And between S2a and S2b: orbital tightening (observed F = 1.81), ear changes (observed F = 1.96) and nose/cheek flattening (observed F = 1.72, $p < 0.01$ all comparisons), but not whisker changes (observed F = 1.35, $p > 0.05$). At all stages, orbital tightening had the highest ICC, improving from 0.69 to 0.84. Following training, ICCs for individual action units fell within the good or very good range (Table 1).

Comparing individual rater performance against the experienced rater showed considerable variation following the first training round with ICCs ranging from fair to good. All trainee raters showed improvement with training (Table 2).

There were 28 images (19%) with score differences between raters of 2 points at S2a. Removing these scores had a minimal effect on the recalculated ICCs for S2b; the 95%CI of the ICCs overlapped with those for the full data set (Supplementary Data_S2).

There was a significant increase in RGS scores between baseline (n = 41, 0.45 ± 0.07) and 6-9 hours after treatment (n = 29, 0.92 ± 0.08, $p < 0.001$, 95%CI of mean difference 0.27 to 0.68, Supplementary Data_S3).

Discussion

Little is known regarding the need for, or role of, rater training in the use of rodent grimace scales. Where training has been described, it ranges from reviewing the grimace scale training manuals (Leach, Klaus & Miller et al., 2012; Faller, McAndrew & Schneider et al., 2015) to a single training session of variable length (Langford, Bailey & Chanda et al., 2010; Sotocinal, Sorge & Zaloum et al., 2011; Oliver, De Rantere & Ritchie et al., 2014; De Rantere, Schuster & Reimer et al., 2016) or multiple training sessions (Mittal, Gupta & Lamarre et al., 2016). Few studies describe an assessment of reliability (Langford, Bailey & Chanda et al., 2010; Sotocinal, Sorge & Zaloum et al., 2011; Oliver, De Rantere & Ritchie et al., 2014; Mittal, Gupta & Lamarre et al., 2016). The results of this study show that an assessment of reliability is necessary to confirm that training will lead to proficiency.

Our results suggest that reliability is limited when training is limited to reviewing the training manual, improving when feedback and discussion with an experienced rater are included.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26721v1 | CC BY 4.0 Open Access | rec: 17 Mar 2018, publ: 17 Mar 2018

The rate at which individuals achieve proficiency in a task is highly variable and, as such, it is erroneous to assume that training guarantees proficiency. Neither a single training session nor repeated attempts at a task ensure proficiency (Campbell, Hecker & Biau et al. 2014; de Oliveira Filho, 2002; Roughan & Flecknell, 2006). A simple method of evaluating rater proficiency is to assess inter-rater reliability (Streiner & Norman, 2008). This provides assurance that variability in scoring is at an acceptable level and enables rogue raters to be identified (Mittal, Gupta & Lamarre et al., 2016; Brondani, Mama & Luna; 2013). Identification of such raters during training allows for further testing and assessment or removal from participation in scoring (Mittal, Gupta & Lamarre et al., 2016). Ensuring reliability will reduce data variability and consequently, animal use. An alternative approach is to use a single rater; however, it is still useful to compare the performance of a single rater against that of an experienced rater, or a standard set of scores, to confirm reliability and consistency over time (Oliver, De Rantere & Ritchie et al., 2014). The presence of systematic bias may negatively affect data interpretation and pain management (Faller, McAndrew & Schneider et al., 2015).

Orbital tightening had the highest associated ICC following the initial round of scoring, which was maintained throughout training. In contrast, the reliability of whisker scoring remained relatively low throughout training. These results support previous findings that assessing the whisker change action unit is more difficult for raters than orbital tightening (Oliver, De Rantere & Ritchie et al., 2014).

A limitation of this study was re-scoring the 150 image set in the final training round, with the potential for memorised scores assigned during the group discussion following the second training round being applied rather than a rater scoring independently. We feel this is unlikely due to the large number of images scored, the similar appearance of rodent faces from similar strains, the time elapsed between review rounds, the small number of images reviewed during group discussion and the nature of the group discussion, where disagreement between raters was acceptable. The minimal difference in ICCs after removal of the 28 image scores supports this assertion.

Images for training were selected on the basis of quality rather than to allow comparison between treatment groups. This limits any assessment of construct validity but the comparison of baseline and predicted peak pain periods indicates that accuracy was preserved.

Conclusion

These data show that reliance on access to the available manuals for rater training may be insufficient. Formal training improves inter-rater reliability and is likely to reduce data variability if rater proficiency is assessed before embarking on data collection. Collaborative training between research groups would ensure similar levels of rater proficiency and improve the reproducibility of research. Inclusion of clear descriptions of rater training and assessment would help in evaluating study results.

Acknowledgements

The authors wish to thank Susana Sotocinal of the Mogil Laboratory, University of McGill, for invaluable assistance in establishing the Rat Grimace Scale in our laboratory and reviewing the selection of images in our training manual (Supplementary Data_S1), and Kent Hecker and Grace Kwong (University of Calgary) for statistical advice.

References

Brondani JT, Mama KR, Luna SP, Wright BD, Niyom S, Ambrosio J, Vogel PR, Padovani CR. 2013. Validation of the English version of the UNESP-Botucatu multidimensional composite pain scale for assessing postoperative pain in cats. *BMC Vet Res* 9:143. DOI:10.1186/1746-6148-9-143.

Campbell RD, Hecker KG, Biau DJ, Pang DS. 2014. Student attainment of proficiency in a clinical skill: the assessment of individual learning curves. *PLoS One* 9:e88526. DOI: 10.1371/journal.pone.0088526.

de Oliveira Filho GR. 2002. The construction of learning curves for basic skills in anesthetic procedures: an application for the cumulative sum method. *Anesth Analg* 95:411-416.

De Rantere, D, Schuster CJ, Reimer JN, Pang DS. 2016. The relationship between the Rat Grimace Scale and mechanical hypersensitivity testing in three experimental pain models. *Eur J Pain* 20:417–426. DOI: 10.1002/ejp.742.

Faller KM, McAndrew DJ, Schneider JE, Lygate CA. 2015. Refinement of analgesia following thoracotomy and experimental myocardial infarction using the Mouse Grimace Scale. *Exp Physiol* 100:164–172. DOI: 10.1113/expphysiol.2014.083139.

Feldt LS, Woodruff DJ and Salih FA. 1987. Statistical inference for coefficient alpha. *Appl Psychol Meas* 11:93-103.

Kuzmic P. 2015. Critical values of F-statistics. *Available at http://www.biokin.com/tools/f-critical.html* (accessed 26 February 2018).

Langford DJ, Bailey AL, Chanda ML, Clarke SE, Drummond TE, Echols S, Glick S, Ingrao J, Klassen-Ross T, Lacroix-Fralish ML, Matsumiya L, Sorge RE, Sotocinal SG, Tabaka JM, Wong D, van den Maagdenberg AM, Ferrari MD, Craig KD, Mogil JS. 2010. Coding of facial expressions of pain in the laboratory mouse. *Nat Methods* 7:447–449. DOI: 10.1038/nmeth.1445.

Leach MC, Klaus K, Miller AL, Scotto di Perrotolo M, Sotocinal SG, Flecknell PA. 2012. The assessment of post-vasectomy pain in mice using behaviour and the Mouse Grimace Scale. *PLoS One* 7:e35656. DOI: 10.1371/journal.pone.0035656.

Mittal A, Gupta M, Lamarre Y, Jahagirdar B, Gupta K. 2016. Quantification of pain in sickle mice using facial expressions and body measurements. *Blood Cells Mol Dis* 57:58–66. DOI: 10.1016/j.bcmd.2015.12.006.

Mogil JS and Crager SE. 2004. What should we be measuring in behavioral studies of chronic pain in animals? *J Pain* 112:12–15. DOI: 10.1016/j.pain.2004.09.028.

Oliver V, De Rantere D, Ritchie R, Chisholm J, Kecker KG, Pang DS. 2014. Psychometric assessment of the Rat Grimace Scale and development of an analgesic intervention score. *PLoS One* 9:e97882. DOI: 10.1371.journal.pone.0097882.

Pang DS. 2018. Rat Grimace Scale rater training data 1.0. *Available at https://doi.org/10.7910/DVN/57K7PE* (accessed 26 February 2018).

Roughan JV and Flecknell PA. 2016. Training in behaviour-based post-operative pain scoring in rats - an evaluation based on improved recognition of analgesic requirements. *Appl Anim Behav Sci* 96:327-342. DOI: 10.1016/j.applanim.2005.06.012.

Sotocinal SG, Sorge RE, Zaloum A, Tuttle AH, Martin LJ, Wieskopf JS, Mapplebeck JC, Wei P, Zhan S, Zhang S, McDougall JJ, King OD, Mogil JS. 2011. The Rat Grimace Scale: a partially automated method for quantifying pain in the laboratory rat via facial expressions. *Mol Pain* 7:55. DOI: 10.1186/1744-8069-7-55.

Streiner DL and Norman GR. 2008. Reliability. In: Streiner DL and Normal GR, eds. *Health Measurement Scales*. New York: Oxford University Press, 167–210.

Legends

Figure 1. Average group ICCs for each of the three datasets (with 95%CI) with reference values (Oliver, De Rantere & Ritchie et al., 2014).

Supplementary Data_S3. Bar graph (mean ± SEM) showing RGS scores at baseline (n = 41 images) and 6-9 hours after treatment (n = 29 images: intraplantar Complete Freund’s Adjuvant; n = 19 images, plantar incision; n = 10 images). Broken horizontal line indicates derived analgesia intervention threshold (Oliver, De Rantere & Ritchie et al., 2014).

Table 1(on next page)

Group Intra-class Correlation Coefficients (ICC) for each of the datasets.

S1, S2a and S2b are the first, second and third training round, respectively. Data are ICCsingle [95%CI]. Within a row, identical superscript letters indicate significant differences between the different training rounds, $p < 0.01$. Reference values are from Oliver, De Rantere, Ritchie et al. (2014).

Table 1. Group Intra-class Correlation Coefficients (ICC) for each of the datasets. S1, S2a and S2b are the first, second and third training round, respectively. Data are ICCsingle [95%CI]. Within a row, identical superscript letters indicate significant differences between the different training rounds, $p < 0.01$. Reference values are from Oliver, De Rantere, Ritchie et al. (2014).

Action Unit	S1	S2a	S2b	Reference values
Orbital tightening	0.69 [0.56-0.80][a]	0.71 [0.63-0.78][b]	0.84 [0.80-0.87][a,b]	0.92 [0.89-0.95]
Ear changes	0.40 [0.25-0.56][a]	0.45 [0.35-0.54][b]	0.72 [0.66-0.77][a,b]	0.62 [0.51-0.72]
Nose/Cheek flattening	0.36 [0.21-0.52][a]	0.50 [0.41-0.58][b]	0.71 [0.65-0.76][a,b]	0.62 [0.51-0.72]
Whisker change	0.39 [0.26-0.55]	0.50 [0.42-0.58]	0.63 [0.57-0.70]	0.52 [0.39-0.63]

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26721v1 | CC BY 4.0 Open Access | rec: 17 Mar 2018, publ: 17 Mar 2018

Table 2(on next page)

Agreement of individual raters when compared to an experienced rater (DP).

Data are ICCsingle [95%CI]. Within a column, matching superscript letters indicate significant differences ($p < 0.01$).

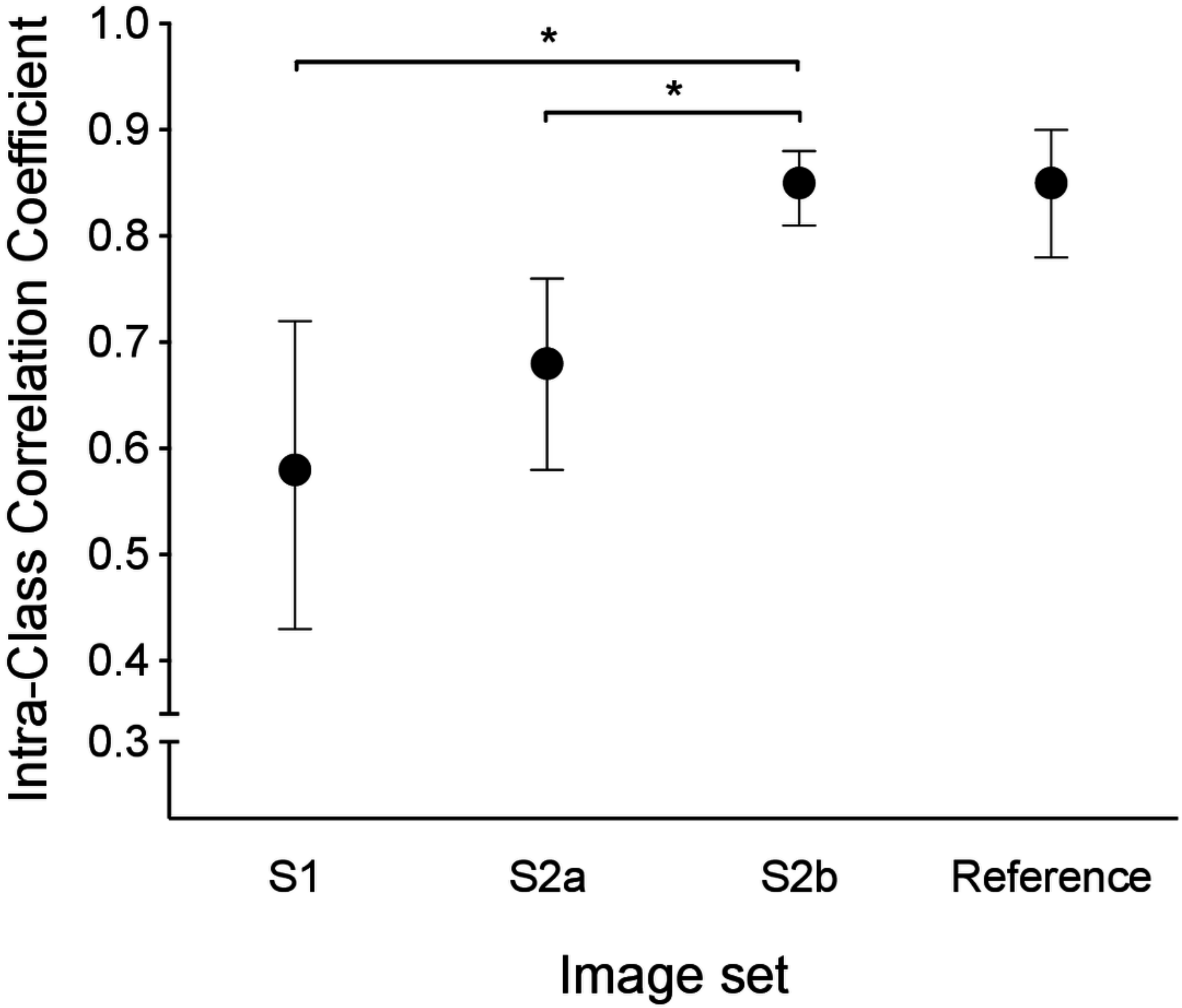
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26721v1 | CC BY 4.0 Open Access | rec: 17 Mar 2018, publ: 17 Mar 2018

Table 2: Agreement of individual raters when compared to an experienced rater (DP). Data are ICCsingle [95%CI]. Within a column, matching superscript letters indicate significant differences ($p < 0.01$).

Image set	Rater 1 vs DP	Rater 2 vs DP	Rater 3 vs DP	Rater 4 vs DP
S1	0.41 [0.06-0.66][a,b]	0.70 [0.50-0.83][a]	0.62 [0.36-0.79][a]	0.42 [0.13-0.64][a]
S2a	0.84 [0.79-0.88][a]	0.75 [0.68-0.82][b]	0.68 [0.25-0.84][b]	0.65 [0.38-0.79][b]
S2b	0.89 [0.85-0.92][b]	0.88 [0.84-0.91][a,b]	0.91 [0.88-0.94][a,b]	0.90 [0.87-0.93][a,b]

Figure 1

Average group ICCs for each of the three datasets (with 95%CI) with reference values (Oliver, De Rantere & Ritchie et al., 2014).



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26721v1 | CC BY 4.0 Open Access | rec: 17 Mar 2018, publ: 17 Mar 2018