Journal of the American Association for Laboratory Animal Science
Copyright 2019
by the American Association for Laboratory Animal Science

Vol 58, No 2
March 2019
Pages 178–183

Influence of Rater Training on Inter- and Intrarater Reliability When Using the Rat Grimace Scale

Emily Q Zhang,[1,†] Vivian SY Leung,[2,†] and Daniel SJ Pang[2,*]

Rodent grimace scales facilitate assessment of ongoing pain. Reported rater training using these scales varies considerably and may contribute to the observed variability in interrater reliability. This study evaluated the effect of training on interrater reliability with the Rat Grimace Scale (RGS). Two training sets (42 and 150 images) were prepared from acute pain models. Four trainee raters progressed through 2 rounds of training, scoring 42 images (set 1) followed by 150 images (set 2a). After each round, trainees reviewed the RGS and any problematic images with an experienced rater. The 150 images were then rescored (set 2b). Four years later, trainees rescored the 150 images (set 2c). A second group of raters (no-training group) scored the same image sets without review with the experienced rater. Inter- and intrarater reliability were evaluated by using the intraclass correlation coefficient (ICC), and ICC values were compared by using the Feldt test. In the trainee group, interrater reliability increased from moderate to very good between sets 1 and 2b and increased between sets 2a and 2b. Action units with the highest and lowest ICC at set 2b were orbital tightening and whiskers, respectively. In comparison to an experienced rater, the ICC for all trainees improved, ranging from 0.88 to 0.91 at set 2b. Four years later, very good interrater reliability was retained, and intrarater reliability was good or very good). The interrater reliability of the no-training group was moderate and did not improve from set 1 to set 2b. Training improved interrater reliability, with an associated reduction in 95%CI. In addition, training improved interrater reliability with an experienced rater, and performance was retained.

Abbreviations: ICC, intraclass correlation coefficient; RGS, rat grimace scale.

DOI: 10.30802/AALAS-JAALAS-18-000044

The effectiveness of a pain assessment scale lies in its validity (that is, it measures 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.[6,11,13,15,17,21] Increasing evidence supports that grimace scales discriminate painful and nonpainful states in a range of acute pain models and interventions.[6,11,12,17,21] However, reports conflict regarding reliability when multiple raters score images.[7,11,14,17,21] Factors contributing to this variability may include a lack of structured training and variation in individual learning curves.[4,5,20]

It is unclear what level of training is required to attain proficiency in using grimace scales. Most studies include minimal, nonspecific descriptions of training,[7,11,12,14,17,19,21] and few report any measure of reliability.[11,14,17,21] Trainees progress at different rates during training to achieve proficiency in a task;[4,14,20] therefore, in addition to training, some assessment of score reliability is necessary. The effect 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- and intrarater reliability when scoring was performed with single and multiple raters applying the RGS. We hypothesized that training would improve interrater reliability.

Received: 16 Apr 2018. Revision requested: 31 May 2018. Accepted: 14 Aug 2018.
[1]Western College of Veterinary Medicine, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, and [2]Faculty of Veterinary Medicine, Université de Montréal, Saint-Hyacinthe, Québec, Canada.
[*]Corresponding author. Email: daniel.pang@umontreal.ca
[†]These authors contributed equally to this study.

Materials and Methods

Animals and image selection. We created 2 sets of training images from images collected during an unrelated project that had received IACUC approval from the University of Calgary Health Sciences Animal Care Committee (protocol ID, AC13-0161 and AC13-0124).[6] This project used the following acute pain models: intraplantar carrageenan, intraplantar complete Freund adjuvant, and plantar incision. The RGS scores from the 3 models displayed the full spectrum of possible RGS scores (that is, 0 to 2).[6] Animals were adult (10 wk or older) male Wistar rats (n = 34) from a single commercial source (Charles River Laboratories, Senneville, Québec, Canada).

The methodology used to generate images was previously described.[21] Briefly, still images were captured from high-definition videorecordings and cropped so that only the face was visible. Each image was presented on a single slide in a presentation program (PowerPoint, version 14.0, Microsoft, Redmond, WA). Slide order was randomized, and identifying information (animal ID, time point, model) was removed.

A single person not involved with the study selected images on the basis of image quality only. We created 2 unique sets of training images, one of 42 images (set 1) and the other containing 150 images (set 2). Participants scored images (score range, 0 to 2) by using the RGS, and the average score was calculated from 4 action units: orbital tightening, nose or cheek flattening, ear change, and whisker change.

Training protocol. None of the 4 trainee raters recruited had previous experience with the RGS. All trainee raters were female undergraduate and graduate students (age, 20 to 25 y) who were studying veterinary medicine, biology (n = 2), or health sciences and were recruited when they joined the research group as project students. No trainee raters had previous experience with rats, either as experimental animals or pets, before beginning training. The experienced rater (DP) had applied the RGS for several years, successfully identifying painful interventions by using established models (a form of construct validity; known-group discrimination),[3,6,17] and adoption of the RGS method within the research group of the experienced rater was supported—through informal evaluation of scoring performance—by assistance from the Mogil laboratory (McGill University), who developed the mouse and rat grimace scales.[11,21]

All trainee raters followed the same scoring protocol (Figure 1): set 1 images were scored independently by each trainee, who used the provided training manuals.[18,21] Trainee raters were encouraged to record comments regarding any images they found difficult to score. After scoring set 1, trainee raters (as a group) reviewed their scores with the experienced rater, 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 standardization of scoring images assigned a score of 0 or 2. Disagreement in scores was tolerated, provided differences between raters did not exceed 1 scale point. The standard of scoring was set by the experienced rater, after establishment of the technique within the laboratory with the support of the Mogil laboratory (McGill University). Once review of set 1 scoring was complete, trainee raters independently scored set 2 images and noted comments as before (set 2a). The set 2 image set was added when more images were available. After a facilitated group discussion with the experienced rater (as done for set 1), the trainee raters independently scored the set 2 images a second time (set 2b). Approximately 15 to 30 images were reviewed during group discussions, with 2 to 3 wk between reviews. Intrarater reliability was assessed by asking the trainee raters to rescore independently— with access to the training manual —set 2 images (set 2c). Set 2c scoring took place 4 y after initial training. The order of the images was randomized from set 2b. At the time of set 2c scoring, trainee rater 1 had not used the RGS in 10 mo, and trainee raters 3 and 4 had not used it in 3 y; trainee rater 2 was still in the research group and actively using the RGS. All trainee raters were asked whether they remembered any previous scores or images from the data set.

No-training group. A second group of raters (no training) was recruited to assess whether repeated scoring of images (with access to the training manual) without associated group discussions was sufficient to achieve scoring proficiency. We recruited 8 raters, 6 of whom completing image scoring (raters 5 through 10; 1 man, 5 women; age, 24 to 26 y). Rater 7 was the only rater aware of the RGS but had never been trained to use the scale. None of the raters had previous experiences with rats, as either experimental animals or pets. All raters had a science education background: undergraduate degree in zoology (n = 3), a veterinarian (n = 1) or a veterinary student (n = 1), or in a master's program in integrative biology (n = 1).

These no-training raters scored the same image sets as the trainee raters (sets 1, 2a, and 2b), with access to the same training manuals, but there were no group reviews or discussion of images at any time during the scoring process (Figure 1).

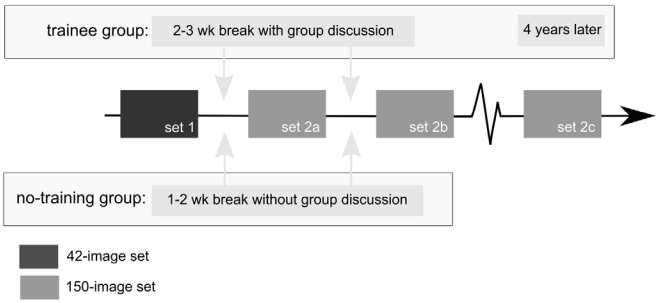

Figure 1. Timeline of training protocol. Two image sets of 42 and 150 images (set 1 and set 2, respectively) were scored independently by all trainee raters and no-training raters. For trainee raters, set 1, set 2a and set 2b were scored with 2 to 3 wk of break between sets. During the break, a group discussion with the experienced rater took place to discuss inconsistencies. After each scoring session, the scores from each individual trainee rater was compared with those from the experienced rater to assess interrater reliability. Four years later, the 150-image set was randomized and rescored (set 2c) by all trainee raters. Their scores were compared with the experienced rater's and with their own scores from set 2b to assess inter- and intrarater reliability, respectively. For no-training raters, set 1, set 2a, and set 2b were scored with a 1- to 2-wk break between sets. These raters never participated in any discussion, and their scores were also compared with the experienced rater to assess interrater reliability.

Statistics. Intraclass correlation coefficients (ICC; version 12.6.1.0, MedCalc Software, Ostend, Belgium) were calculated to measure the reliability of RGS scoring between and within raters for the individual action unit scores and average RGS scores. An absolute model was used for the ICC calculation and single-measure ICC reported, for each dataset (set 1, set 2a, set 2b and set 2c), and for both groups of raters (trainee and no-training). In addition, ICC were calculated for the comparison between the scores assigned by each individual trainee or no-training rater and those of the experienced rater (DP) to determine the reliability of individual raters. Comparisons were preestablished: by using Feldt tests, calculated ICC were compared between set 1 and set 2a, set 1 and set 2b, set 2a and set 2b, and set 2b and set 2c (critical F set at $\alpha = 0.01$; differences were considered significant when the observed F value was greater than the critical F value).[8,10] In addition, ICC were calculated between each trainee and no-training rater's own scores (set 2b and set 2c) to assess intrarater reliability over time. Interpretation of the ICC followed the same divisions as used previously: very good, 0.81 to 1.0; good, 0.61 to 0.80; moderate, 0.41 to 0.60; fair, 0.21 to 0.40; and poor, less than 0.20.[17] During the training process, trainee raters were said to be proficient when calculated ICC ± 95%CI overlapped with those published in a study reporting interrater reliability[17] and when they had obtained an ICC of at least 0.80.[9] To assess the potential effect of scores being memorized during group discussion between set 2a and set 2b and thus introducing bias in to the ICC calculation for set 2b, images with the greatest scoring variability at set 2a (those with a difference of 2 points between any 2 raters and therefore the most likely to have been discussed) were removed, and the ICC for set 2b recalculated. Data are presented as ICC (± 95%CI), and a P value of less than or equal to 0.017 (that is, corrected for multiple comparisons) was considered significant. Scoring accuracy was assessed by comparing the experienced rater's scores for images collected at baseline and 6 to 9 h after treatment (when a peak in RGS scores could be expected for the models studied;[6] paired t test with α set at 0.05) from the set 2 images. The datasets generated from this study and training manual are available in the Harvard Dataverse repository.[18]

Results

Four trainee raters and 6 no-training raters completed the study. All training images were scored by every rater, and all scores were included in the subsequent analysis.

Interrater reliability of trainee raters. Training was associated with a progressive improvement in interrater reliability and narrowing 95%CI (Figure 2). The first training round (set 1) 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 set 1 and set 2a (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 set 2b (0.85 [0.81–0.88]) compared with set 1 (observed F = 2.8) and set 2a ($F_{0.01;149,149}$ = 1.47, observed F = 2.13, $P < 0.01$ for both comparisons). The resultant set 2b ICC was classified as very good and comparable with published values (Figure 2).[17]

A similar pattern of improvement was observed in the scores for individual action units (Table 1). Significant increases in ICC were observed between set 1 and set 2b for orbital tightening (observed F = 1.94), ear changes (observed F = 2.14) and nose or cheek flattening (observed F = 2.21, $P < 0.01$ for all comparisons) but not whisker changes (observed F = 1.65, $P > 0.05$). Significant increases in ICC also occurred between set 2a and set 2b for orbital tightening (observed F = 1.81), ear changes (observed F = 1.96), and nose or cheek flattening (observed F = 1.72, $P < 0.01$ for 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. After training of trainee raters, ICC for individual action units fell within the good or very good range (Table 1).

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

In set 2a, 28 images (19%) led to score differences of 2 points between raters. Removing these scores had minimal effect on the recalculated ICC for set 2b (average RGS scores, 0.85 [0.81–0.88] and 0.86 [0.83–0.89] for 150 and 122 images, respectively).

RGS scores significantly increased between baseline (n = 41; 0.45 ± 0.07) and 6 to 9 h after treatment (n = 29; 0.92 ± 0.08; $P < 0.001$; 95%CI of mean difference, 0.27 to 0.68), at which time the mean RGS score exceeded a published analgesic intervention threshold.[17]

When the images were rescored 4 y after initial training (set 2c), the ICC was good for the averaged RGS scores (0.80 [0.76–0.84]), and proficiency was maintained from set 2b (observed F = 1.33, $P > 0.01$). Between set 2b and set 2c, there were no significant differences for nose or cheek flattening (observed F = 1.24, $P > 0.05$), whisker changes (observed F = 1.24, $P > 0.05$), and ear changes (observed F = 1.42, $P > 0.01$; Table 1). However, interrater reliability from set 2b was not maintained and decreased significantly for orbital tightening (observed F = 1.50, $P < 0.01$). All trainee raters maintained similar proficiency as the experienced rater (observed F < 1.31, $P > 0.05$) except for trainee rater 4 (observed F = 2.20, $P < 0.01$; Table 2).

Intrarater reliability of trainee raters. The ability of a trainee rater to score reliably over time was good or very good, with ICC ranging from 0.78 to 0.86 for the average RGS (Table 3). The intrarater reliability of individual action units ranged from moderate to very good, depending on the action unit and trainee rater. Two trainee raters (2 and 4) reported that they did not recognize any images or remember previous scores. The remaining trainee raters (1 and 3) reported recognizing a few images but did not remember scores.

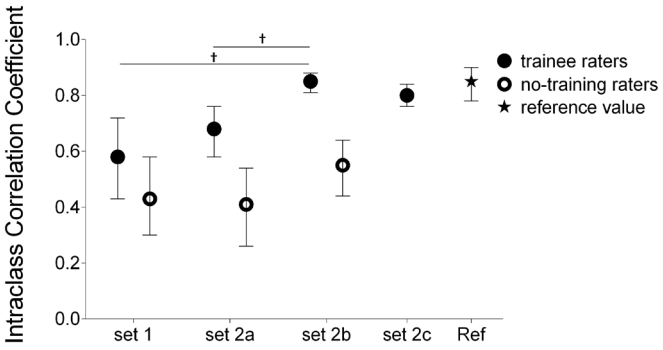

Figure 2. Average group ICC for each of the 4 datasets (mean and 95%CI) with reference values (from reference 17 [Ref]). †, $P < 0.01$.

Interrater reliability of no-training raters. In the no-training group, repeated scoring of images did not result in significant improvement of interrater reliability (Figure 2). Agreement between raters was moderate during each stage of scoring, with no significant improvement observed from set 1 (0.43 [0.30–0.58]) to set 2a (0.41 [0.26–0.54]; $F_{0.01;\ 149;\ 41}$ = 1.88, observed F = 1.04, $P > 0.05$), from set 1 to set 2b (0.55 [0.44–0.64]; $F_{0.01;\ 149;\ 41}$ = 1.88, observed F =1.27, $P > 0.05$), or from set 2a to set 2b ($F_{0.01;\ 149;\ 149}$ = 1.47, observed F =1.31, $P > 0.05$).

This lack of improvement also was observed in regard to individual action units (Table 4). Some improvements for individual raters were observed when their scores were comparable to the experienced rater's, but none of the raters achieved very good agreement with the experienced rater (Table 5). Rater 6 improved from set 1 to set 2a (observed F = 1.97, $P < 0.01$), and raters 7 and 8 improved from set 2a to set 2b (observed F = 1.58, $P < 0.01$; observed F = 1.57, $P < 0.01$).

Discussion

Our results suggest that scoring reliability is minimal when raters review the training manual and score images without the opportunity for feedback and discussion. In contrast, improvement occurs when feedback and discussion with an experienced rater is included. The high level of reliability and proficiency due to training can be maintained for several years.

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 grimace scale training manuals[7,12] to a single training session of variable length[6,11,17,19,21]or multiple training sessions.[14] Few studies describe an assessment of reliability.[11,14,17,21] The results of our current study show that an assessment of reliability is necessary to confirm that training leads to proficiency as well as standardized scoring. Our study also showed that the inclusion of group discussion as part of training is beneficial. Although repeated exposure without discussion does have some benefits, as demonstrated by the increased reliability among the individual raters from the no-training group, this improvement is variable between raters and limited.

The rate at which trainees achieve proficiency in a task is highly variable and, as such, it is erroneous to assume that participating in training guarantees proficiency. Neither a single training session nor repeated attempts at a task ensure proficiency.[4,5,20] The length and intensity of training should depend on the difficulty of the mastering the tool and the proficiency of the trainee.[9] In addition, proficiency should not be assumed when a rater feels confident using a scale after training.[1] Instead, it is important to test the actual proficiency of raters, and a

Table 1. Group ICC for each of the datasets

Action unit	Set 1	Set 2a	Set 2b	Set 2c	Reference value
Orbital tightening	0.69 (0.56–0.80)[a]	0.71 (0.63–0.78)[b]	0.84 (0.80–0.87)[a,b,c]	0.76 (0.70–0.81)[c]	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,c]	0.60 (0.51–0.68)[c]	0.62 (0.51–0.72)
Nose or cheek flattening	0.36 (0.21–0.52)[a]	0.50 (0.41–0.58)[b]	0.71 (0.65–0.76)[a,b]	0.64 (0.57–0.70)	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)	0.54 (0.45–0.62)	0.52 (0.39–0.63)

Set 1, set 2a, and set 2b are the first, second and third training round, respectively. Set 2c was scored 4 y after initial training. ICC scores are divided as: very good, 0.81–1.0; good, 0.61–0.80; moderate, 0.41–0.60; fair, 0.21–0.40; and poor, less than 0.20. Data are given as ICCsingle (95%CI). Within a row, identical superscript letters indicate significant ($P < 0.01$) differences between the different training rounds. Reference values and ICC score divisions are from reference 17.

Table 2. Agreement of each individual trainee rater when compared with an experienced rater (DP)

Image set	Rater 1 compared with DP	Rater 2 compared with DP	Rater 3 compared with DP	Rater 4 compared with DP
Set 1	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]
Set 2a	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]
Set 2b	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,c]
Set 2c	0.87 (0.82–0.90)	0.86 (0.82–0.90)	0.86 (0.80–0.90)	0.78 (0.71–0.83)[c]

ICC scores are divided as: very good, 0.81–1.0; good, 0.61–0.80; moderate, 0.41–0.60; fair, 0.21–0.40; and poor, less than 0.20. Data are given as ICCsingle (95%CI). Within a column, identical superscript letters indicate significant ($P < 0.01$) differences. Reference values and ICC score divisions are from reference 17.

Table 3. ICC for intrarater reliability for each individual trainee rater 4 y after initial training

Action unit	Rater 1	Rater 2	Rater 3	Rater 4
Average	0.85 (0.78–0.90)	0.86 (0.82–0.90)	0.86 (0.79–0.90)	0.78 (0.71–0.84)
Orbital tightening	0.72 (0.53–0.82)	0.86 (0.82–0.90)	0.85 (0.78–0.89)	0.75 (0.63–0.83)
Ear changes	0.68 (0.48–0.80)	0.49 (0.11–0.70)	0.74 (0.66–0.81)	0.71 (0.61–0.79)
Nose or cheek flattening	0.64 (0.53–0.73)	0.68 (0.56–0.77)	0.74 (0.60–0.82)	0.63 (0.53–0.72)
Whisker change	0.77 (0.70–0.83)	0.69 (0.55–0.78)	0.53 (0.27–0.69)	0.47 (0.34–0.59)

ICC scores are divided as: very good, 0.81–1.0; good, 0.61–0.80; moderate, 0.41–0.60; fair, 0.21–0.40; and poor, less than 0.20. Data are given as ICCsingle (95%CI). Reference values and ICC score divisions are from reference 17.

Table 4. Group ICC for each of the datasets for the no-training group

Action Unit	Set 1	Set 2a	Set 2b	Reference values
Orbital tightening	0.48 (0.35–0.62)	0.65 (0.58–0.71)	0.71 (0.65–0.76)	0.92 (0.89–0.95)
Ear changes	0.24 (0.14–0.38)	0.35 (0.25–0.46)	0.35 (0.24–0.46)	0.62 (0.51–0.72)
Nose/Cheek flattening	0.35 (0.23–0.50)	0.17 (0.09– 0.26)	0.35 (0.27–0.43)	0.62 (0.51–0.72)
Whisker change	0.19 (0.09 –0.32)	0.23 (0.16–0.32)	0.25 (0.18–0.33)	0.52 (0.39–0.63)

Set 1, set 2a, and set 2b are the first, second and third training round, respectively. ICC scores are divided as: very good, 0.81–1.0; good, 0.61–0.80; moderate, 0.41–0.60; fair, 0.21–0.40); and poor, less than 0.20. Data are given as ICCsingle (95%CI). Reference values and ICC score divisions are from reference 17.

Table 5. Agreement of each individual no training rater when compared with an experienced rater (DP)

Image set	Rater 5 compared with DP	Rater 6 compared with DP	Rater 7 compared with DP	Rater 8 compared with DP	Rater 9 compared with DP	Rater 10 compared with DP
Set 1	0.63 (0.40–0.78)	0.37 (0.07–0.60)[a]	0.57 (0.33–0.74)	0.33 (0.04–0.57)	0.56 (0.24–0.75)	0.57 (0.25–0.76)
Set 2a	0.72 (0.60–0.81)	0.68 (0.58–0.76)[a]	0.51 (0.06–0.73)[a]	0.12 (–0.06–0.30)[a]	0.63 (0.28–0.80)	0.67 (0.45–0.79)
Set 2b	0.68 (0.57–0.77)	0.65 (0.54–0.74)	0.69 (0.41–0.82)[a]	0.41 (0.05–0.64)[a]	0.73 (0.56–0.82)	0.68 (0.51–0.78)

Set 1, set 2a, and set 2b are the first, second and third training round, respectively. ICC scores are divided as: very good, 0.81–1.0; good, 0.61–0.80; moderate, 0.41–0.60; fair, 0.21–0.40); and poor, less than 0.20. Data are given as ICCsingle (95%CI). Within a column, identical superscript letters indicate significant ($P < 0.01$) differences between the different training rounds. Reference values and ICC score divisions are from reference 17.

simple approach is to assess interrater reliability.[22] This practice provides assurance that scoring has reached the desired standard, that variability is at an acceptable level and enables rogue raters to be identified.[2,14] Identification of rogue raters during training allows for further testing and assessment or removal from participation in scoring.[14,16] Ensuring reliability and standardizing scoring will reduce data variability and consequently, animal use. An alternative approach is to use a single

rater; however, it is still helpful to compare the performance of a single rater with that of an experienced rater or a standard set of scores, to confirm reliability and consistency over time.[17] The presence of systematic bias may negatively affect data interpretation and pain management.[7]

Orbital tightening had the highest associated ICC after 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 is orbital tightening.[17]

Four years after training, with variable use of the RGS during this time, the inter- and intrarater reliability of the average RGS was maintained. This outcome indicates that raters can retain scoring proficiency and score consistently relative to each other and themselves and achieve the standard set by the experienced rater. This finding agrees with a previous study showing that a single rater maintained scoring reliability after a break of 6 mo.[17] Nevertheless, the reductions in ICC that we observed for one of the action units indicate that some degree of retraining may be beneficial.

A recent description of a successful machine learning approach to the MGS highlights the potential for simplifying the standard method of facial image acquisition and scoring.[23] This advance could greatly shorten what is currently a relatively slow process and allow for the scoring of large numbers of animals in a short period of time, an advance over real-time scoring.[13] However, the need for proficient human raters remains necessary to classify those images that cannot currently be scored by machine with a high degree of confidence.[23]

A limitation of our current study was rescoring the 150-image set in the final training round, with the potential for the application of scores memorized during the group discussion after the second training round being applied rather than a rater scoring independently. We feel that this bias is unlikely due to the large number of images scored, similar appearance of rodent faces from similar strains, time elapsed between review rounds, few images reviewed during group discussion, and nature of the group discussion, where disagreement between raters was acceptable. The minimal difference in ICC after removal of the 28 images with scores that differed by 2 points between raters supports this assertion, as well as the maintained quality of scores after 4 y. A further limitation is the generalizability of these findings, based on 4 trainee raters and 6 no-training raters, to a larger population. These results highlight the risk of assuming that some form of training in the use of the RGS (and perhaps other facial expression scales) is unnecessary and should serve to encourage users to regularly evaluate scoring reliability and accuracy. In more general terms, scale performance is specific to the population and context studied, so that performance—when applied by different raters or in a different context—should be formally evaluated.[22]

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.

In conclusion, these data show that reliance on access to the available manuals for rater training in the RGS may be insufficient. Formal training that includes group discussion with an experienced rater improves interrater 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. Lastly, once raters achieve proficiency, it can be maintained over several years even without scoring during the intervening period.

Acknowledgments

We thank Susana Sotocinal (Mogil Laboratory, McGill University) for invaluable assistance in establishing the Rat Grimace Scale in our laboratory and reviewing the selection of images in our training manual; Kent Hecker and Grace Kwong (University of Calgary) for statistical advice; and Audrey Pang, Chelsea Schuster, Elfleda Chik, Jesse Tong, Julie Reimer, Shi Jie Zhou, Vimanda Chow, and Winnie Yang for participating in the study.

References

1. **Björn A, Pudas-Tähkä SM, Salanterä S, Axelin A.** 2017. Video education for critical care nurses to assess pain with a behavioral pain assessment tool: a descriptive comparative study. Intensive Crit Care Nurs **42:**68–74. https://doi.org/10.1016/j.iccn.2017.02.010.
2. **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:**1–15. https://doi.org/10.1186/1746-6148-9-143.
3. **Calvo G, Holden E, Reid J, Scott EM, Firth A, Bell A, Robertson S, Nolan AM.** 2014. Development of a behaviour-based measurement tool with defined intervention level for assessing acute pain in cats. J Small Anim Pract **55:**622–629. https://doi.org/10.1111/jsap.12280.
4. **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:**1–5. https://doi.org/10.1371/journal.pone.0088526.
5. **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.
6. **De Rantere D, Schuster CJ, Reimer JN, Pang DS.** 2016. The relationship between the Rat Grimace Scale and mechanical hypersensitivity testing in 3 experimental pain models. Eur J Pain **20:**417–426. https://doi.org/10.1002/ejp.742.
7. **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. https://doi.org/10.1113/expphysiol.2014.083139.
8. **Feldt LS, Woodruff DJ, Salih FA.** 1987. Statistical inference for coefficient α. Appl Psychol Meas **11:**93–103. https://doi.org/10.1177/014662168701100107.
9. **Haidet KK, Tate J, Divirgilio-Thomas D, Kolanowski A, Happ MB.** 2009. Methods to improve reliability of video-recorded behavioral data. Res Nurs Health **32:**465–474. https://doi.org/10.1002/nur.20334.
10. **Kuzmic P.** [Internet]. 2015. Critical values of F-statistics. [Cited 26 February 2018]. Available at http://www.biokin.com/tools/f-critical.html.
11. **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. https://doi.org/10.1038/nmeth.1455.
12. **Leach MC, Klaus K, Miller AL, Scotto di Perrotolo M, Sotocinal SG, Flecknell PA.** 2012. The assessment of postvasectomy pain in mice using behaviour and the Mouse Grimace Scale. PLoS One **7:**1–9. https://doi.org/10.1371/journal.pone.0035656.
13. **Leung V, Zhang E, Pang DSJ.** 2016. Real-time application of the Rat Grimace Scale as a welfare refinement in laboratory rats. Sci Rep **6:**1–12. https://doi.org/10.1038/srep31667.

14. **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. https://doi.org/10.1016/j.bcmd.2015.12.006.
15. **Mogil JS, Crager SE.** 2004. What should we be measuring in behavioral studies of chronic pain in animals? Pain **112:**12–15. https://doi.org/10.1016/j.pain.2004.09.028.
16. **Mullard J, Berger JM, Ellis AD, Dyson S.** 2017. Development of an ethogram to describe facial expressions in ridden horses (FEReq). J Vet Behav **18:**7–12. https://doi.org/10.1016/j.jveb.2016.11.005.
17. **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:**1–7. https://doi.org/10.1371/journal.pone.0097882.
18. **Pang DS**. [Internet]. 2018. Rat Grimace Scale rater training data 1.0. [Cited 16 April 2018]. Available at https://doi.org/10.7910/DVN/57K7PE.
19. **Philips BH, Weisshaar CL, Winkelstein BA.** 2017. Use of the Rat Grimace Scale to evaluate neuropathic pain in a model of cervical radiculopathy. Comp Med **67:**34–42.
20. **Roughan JV, Flecknell PA.** 2006. Training in behaviour-based postoperative pain scoring in rats—an evaluation based on improved recognition of analgesic requirements. Appl Anim Behav Sci **96:**327–342. https://doi.org/10.1016/j.applanim.2005.06.012.
21. **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:**1–10.
22. **Streiner DL, Norman GR.** 2008. Reliability. p 167–210. In: Streiner DL, Norman GR, editors. Health measurement scales: a practical guide to their development and use. New York (NY): Oxford University Press.
23. **Tuttle AH, Molinaro MJ, Jethwa JF, Sotocinal SG, Prieto JC, Styner MA, Mogil JS, Zylka MJ.** 2018. A deep neural network to assess spontaneous pain from mouse facial expressions. Mol Pain **14:**1–9. https://doi.org/10.1177/1744806918763658.