

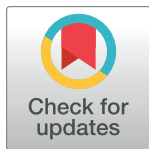
RESEARCH ARTICLE

Using photographs for rating severity degrees of clinical appearance in research mice enables valid discrimination of extreme but not mild and moderate conditions: A pilot study

Johanne C. Krueger^{1,2*}, Maren Boecker³, Siegfried Gauggel³, Andre Bleich⁴, Rene H. Tolba¹

1 Institute for Laboratory Animal Science and Experimental Surgery, RWTH Aachen University, Faculty of Medicine, Aachen, Germany, **2** Animal Welfare Unit, University of Bonn, Bonn, Germany, **3** Institute for Medical Psychology and Medical Sociology, RWTH Aachen University, Faculty of Medicine, Aachen, Germany, **4** Institute for Laboratory Animal Science, Hannover Medical School, Hannover, Germany

* jkrueger@uni-bonn.de



OPEN ACCESS

Citation: Krueger JC, Boecker M, Gauggel S, Bleich A, Tolba RH (2023) Using photographs for rating severity degrees of clinical appearance in research mice enables valid discrimination of extreme but not mild and moderate conditions: A pilot study. PLoS ONE 18(11): e0287965. <https://doi.org/10.1371/journal.pone.0287965>

Editor: David Morton, University of Birmingham, UNITED KINGDOM

Received: July 20, 2022

Accepted: June 19, 2023

Published: November 2, 2023

Copyright: © 2023 Krueger et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript and its [Supporting information](#) files.

Funding: This study was supported in part by the German Research Foundation (Deutsche Forschungsgemeinschaft DFG; FOR-2591, TO 542/5-1, TO 542/6-1; 2016; BL953/11-1) without the funders being involved in the study design, data collection and analysis, decision to publish or the preparation of the article.

Abstract

To ensure good animal welfare in laboratory research and in stockbreeding severity ratings of the animals' wellbeing are essential. The current study investigated how valid raters can evaluate different severity degrees of clinical appearance and how ratings might be influenced by factors other than the severity itself. Ninety-seven people rated the severity degree (none, mild, moderate, or severe) of the clinical appearance of mice seen in eight different images. The images also differed in the perspective in which they had been taken (entire mouse or head only). The raters differed with regard to their experience of working with laboratory animals and were subsequently divided into three groups—beginners, advanced, professionals. Generalisability theory was applied to examine the contribution of the different rater (raters themselves and experience) and image facets (actual degree of severity and perspective) to the overall data variability. The images showing the extreme severity degrees were rated more homogenously and more precisely than were the images showing the intermediate degrees, as compared to the reference scores. The largest source of variance was the actual degree of severity, accounting for 56.6% of the total variance. Considering only the images showing the extreme severity degrees, this percentage rose to 91.6%, accounting almost exclusively for the found variance. In considering only the intermediate severity degrees, the actual degree of severity did not contribute to variance at all. The remaining variance was due to the raters and the interactions between raters, the actual degree of severity and the perspective. The experience of the raters did not account for any variance. Training in the assessment of severity degrees seems necessary to enhance detection of the intermediate degrees of severity, especially when images are used. In addition, good training material should be developed and evaluated to optimise teaching and to minimise wrong assessments.

Competing interests: The authors have no competing interests to disclose.

Introduction

The use of laboratory animals is an inherent part of research in medicine. Nevertheless, there have been heated debates for many years regarding this topic. Discussions are influenced by moral and emotional aspects, of which the core question is whether humans are justified to harm an animal, causing pain, suffering or distress in order to benefit their own species.

The protection of animals and ensuring good animal welfare, defined as the 'physical and mental aspects of quality of life and extending beyond the absence of disease' [1, 2], has become one of the constitutional aims in Germany and is also indicated in the European Directive 2010/63. As defined in the European Directive, institutions are obliged to carry out severity classifications of procedures 'on the basis of estimated levels of pain, suffering, distress and lasting harm' being inflicted on the animals [3]. Moreover, not only these estimated (prospective) levels shall be assigned to the different procedures, but the entity of anxiety, affective internal/emotional state as well as the 'actual severity of the pain, suffering, distress and lasting harm' experienced by the animal [3, 4] shown through varying reactions of the experimental insult or stimuli. The classification of the procedures' severity being inflicted on the animal as well as their reaction towards that stimuli are divided into: 'non-recovery', 'mild', 'moderate', and 'severe'. Morton [5] recently suggested not only to classify a procedure's severity before conducting the experiment, but to standardly evaluate throughout and after the experiment whether the severity classification had been appropriately chosen with regards to the severity actually seen in the animals during the conduction of the experiment.

To determine and detect changes in the animals' wellbeing there are different methods available (e.g. through deviations in physiology, clinic; behavioural changes; alteration of biochemistry, biomarkers) [4]. Morton and Griffiths [6] first described a scheme, with species specific parameters and certain assigned scores. Some of the parameters were objective, meaning that they can be directly measured leaving no scope for interpretation (e.g. heart rate, body weight), whereas others were more subjective such as appearance and behaviour [6, 7]. These latter have to be rated and their rating might not only depend on the actual appearance and behaviour but also on characteristics of the raters themselves (e.g. perception and interpretation because of the experience with the species, overall experience with animal experiments or personal bias) [8–10] or on the material and setting used for assessment (assessment instrument, live observation, video or photo material). Regarding live observation, the presence of the raters might change the behaviour of the animals—causing distress on the animals, as well as prey animals might disguise their behaviour in the presence of humans [1, 11, 12].

Another problem associated with ratings of the animals' wellbeing is that not all severity degrees might be equally easy to assess correctly and consistently, especially when focusing on subjective parameters. In the literature, there are a few studies available on the assessment of different severity degrees. They consistently show that the extremes of a scale are more likely to be detected correctly than are the degrees in between [6, 13]. Moreover, e.g. when dealing with locomotion scores in cows, slight deviations in the grading are not well recognized [14–16]. Schlageter-Tello et al. [14] showed that in a Likert-type scale with five rating categories, the intermediate levels had low intra- and inter-observer agreement, in contrast to the extreme levels of the grading ranging from no to severe. Thus, cows showing slight deviations in the gait were difficult to detect. Also, Garcia et al. [13] investigated mobility scoring in dairy cows, showing that the highest agreement in the ratings was achieved in the extremes of the scale.

Therefore, we hypothesised that when assessing clinical appearance in laboratory animals, namely mice, the extremes on the scale, 'none' and 'severe', could be detected more consistently and precisely than could the severity degrees in between ('mild' and 'moderate'). However, to disentangle the contribution of different sources of variance from the severity ratings

of clinical appearance, characteristics of the images and raters were systematically combined: people with varying degrees of experience with laboratory animals were asked to rate the degree of severity of photographed mice according to the varying categories 'mild', 'moderate', 'severe' and additionally 'none' for an animal without any deviation from its normal state.

Material and methods

The study was ethically approved by the local ethics committee (*Ethik-Kommission der Medizinischen Fakultät der Rheinisch-Westfälischen Technischen Hochschule Aachen*) under the internal number EK 230/22.

Rater

Ninety-seven people from six different laboratory animal science related locations throughout Germany were asked to participate in the study and rated the severity degree of the clinical appearance of eight photographed mice. Each respondent completed and returned a printed version of the questionnaire ([S1 Questionnaire](#)). No payment or other incentive was given.

The study was performed during January and February 2019. In total 97 raters, of which 45.4% (44/97) were male and 54.6% (53/97) female, participated.

For practical reasons regarding the application of the G-theory software for which equal-sized rater experience groups were needed, 13 raters had to be excluded from further analyses, leaving a final total of 84 raters evaluated in the study (more details can be found in the Statistical Methods section).

Besides the severity rating of the mice, the raters had to respond to a few sociodemographic questions, including questions related to laboratory animal experience, since the raters, had different backgrounds in working with laboratory animals regarding profession and work experience.

These questions focused on differentiating between whether participation in animal experiments was carried out, the involvement in taking care of animals and the experience with types of animals (large, rodents or both) ([S1 Questionnaire](#)). Based on the response to these questions, the participants were divided by the study supervisor into three different categories, namely, 'beginners' (B; n = 37, but reduced to n = 28 [male: n = 13 (46.4%); female: n = 15 (53.6%)]), 'advanced' (A; n = 32, but reduced to n = 28 [male: n = 13 (46.4%); female: n = 15 (53.6%)]), and 'professionals' (P; n = 28 [male: n = 10 (35.7%); female: n = 18 (64.3%)]) ([Table 1](#)).

Images

Eight different images of mice ([Fig 1](#)) were presented for evaluation to the raters. The images were taken from the publication, 'Predictive Observation-Based Endpoint Criteria for Mice Receiving Total Body Irradiation' [17], with the permission of the publishers of the *Journal of Comparative Medicine*. The eight images varied according to two attributes: the severity of the clinical appearance of the pictured mouse and the perspective in which the respective mouse was photographed. Regarding the actual degree of severity shown in the pictured mouse, four different degrees were visualised on the images (none, mild, moderate and severe). Each degree of severity was displayed twice—once with images of the entire mouse and another time with images showing just the head of the mouse.

The actual degree shown in the pictures of the mice was determined as follows: Two laboratory animal experts, one of them with over 20 years of experience, the other being lab animal veterinarian, evaluated each image in consensus. Therefore, the original parameters assigned

Table 1. Classification of raters by using their self-declared answers regarding participation in animal experiments/care taking as well as on animal-handling experience with different species (large laboratory animals and rats and mice) into three consecutive groups (B: Beginners, A: Advanced, P: Professionals).

Participation in Animal Experiments	Participation in Animal Care Taking	Experience with Large Laboratory Animals	Experience with Rats and Mice	Rater Category	n	Cumulative n
<1 year	<1 year	Yes or no	Yes or no	B	28	28
1–5 years	<5 years	Yes or no	Yes	A	10	
<5 years	1–5 years	Yes or no	Yes	A	16	
<5 years	1–5 years	Yes	No	A	1	
>5 years	<5 years	Yes	No	A	1	
>5 years	Irrelevant*	Yes or no	Yes	P	11	28
Irrelevant**	>5 years	Yes or no	Yes	P	17	

B, beginners; A, advanced; P, professionals;

*Because of their long experience in animal experiments, people were classified as professionals regardless of their amount of experience in animal care taking.

** Because of their long experience in animal care, people were classified as professionals regardless of their amount of experience in animal experiments.

<https://doi.org/10.1371/journal.pone.0287965.t001>

[17] were used to get a first impression (proof of concept) and one additional parameter (orbital tightening) was added to get an overall score. However, the literature has provided numerous ways to assess and grade the clinical presentation of mice [6, 17–19]. Therefore, the two experts chose two additional parameters (Mouse Grimace Scale [MGS] and appearance; supplementary material [S1 Table]) to determine the actual degree of severity shown in the images. All parameters were combined, and the overall degree of severity (1–none, 2–mild, 3–moderate, and 4–severe) was assigned to each mouse seen in the images. The agreed-on results, ‘reference scores’ from here onwards, are shown in Fig 1. The parameters used to assign the degree of severity are presented in S1 Table.

The letters seen on the images (Fig 1) were covered by a white square to avoid prediction and expectation bias regarding the degree of severity. The raters were asked to rate the degree of severity for each of the images using a 4-point rating scale (1 = none, 2 = mild, 3 = moderate and 4 = severe). They were not informed about how many times each degree of severity was displayed, neither of the parameters used to score the animals. This was part of the study design in order to clarify their intuitive evaluation. The instructions of the questionnaire read as: “Please select one of the following terms, describing the severity of the animals shown in the pictures below. Please only choose one term for each picture: “Non”; “Mild”; “Moderate”; “Severe”” (S1 Questionnaire). The order of the images in the questionnaire was randomised using the randomizing software available under <http://www.randomizer.org/> [20].

Statistical methods

Descriptive statistical analysis was performed. Means (M) and standard deviations (SDs) were derived for each image and rater group separately to get an overall impression of the data distribution and potential deviations from the respective reference score. P-values were computed using one-way ANOVA. Additionally, a boxplot diagram showing the score distributions for each image across all raters was created.

As it had been assumed that it might be easier to rate the extremes of the scale as compared to the middle degrees of severity, the mean absolute deviations from the reference scores and their SDs were calculated. Levene’s test was performed to test for variance homogeneity. Welch-test was performed to compare the mean absolute deviations of the images showing the

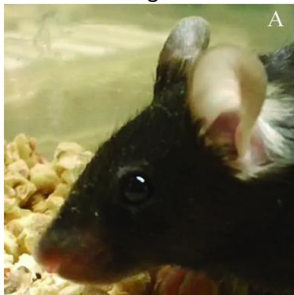
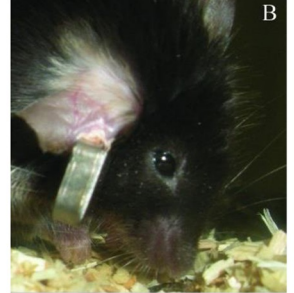
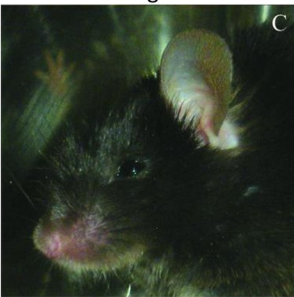
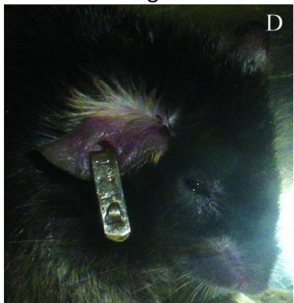
Severity Degree of clinical appearance	Perspective	
	Entire Mouse	Head
None	Image 2 	Image 8 
Mild	Image 5 	Image 3 
Moderate	Image 4 	Image 1 
Severe	Image 6 	Image 7 

Fig 1. Photographs of mice used to assess different severity degrees of clinical appearance in a questionnaire. The position of the images was assigned randomly. The numbers shown in the Table match the position of the photograph in the questionnaire.

<https://doi.org/10.1371/journal.pone.0287965.g001>

extremes of the scale as compared with the intermediate ones. SPSS Statistics Version 25 (IBM Corporation, Armonk, NY, USA) was used for statistics and graphing.

Generalisability theory

In the present study, generalisability analyses were conducted to estimate the contribution of different sources of variance to the severity ratings of the pictured mice (Table 2) [21, 22]. Generalisability theory (G-theory) can be seen as an extension of classical test theory, enabling the researcher to deal with multiple sources of errors simultaneously. The EduG 6.1 software [23] was employed for this purpose. Before the analyses were conducted, 13 raters had to be excluded. This was done as the facet ‘rater’ was nested in the facet ‘experience group’, and analysis with the EduG 6.1 software required that each group had the same number of raters. Consequently, the number of raters of the B- and A-groups had to be reduced to 28, which was the number of the smallest group, the P-group. As one of the research questions of the present study was whether the experience of the raters influenced the severity rating of the pictured mice, the raters to be excluded were selected in such a way that the difference between the experience groups was maximised (e.g. the nine raters of the A-group with the most experience were excluded). Additionally, a second line of analyses was conducted, this time with the 28 raters of the A- and T-groups randomly drawn. The results of the latter analyses will not be reported, as no differences were obtained.

As described above, the raters had to judge the degree of severity in each of the eight images using a Likert-type 4-point rating scale (1 = none, 2 = mild, 3 = moderate and 4 = severe). The images differed according to the two facets, the ‘actual severity degree of clinical appearance of the pictured mouse’ (S; 4 levels) and ‘perspective’ (P; 2 levels) (Table 2), such that each $S \times P$ combination was presented by one image. The raters differed in terms of ‘experience’ (E; 3 levels) and the ‘raters’ themselves, which were nested in the experience groups (R:E; 28 levels). A random-effect nested measurement design ($S \times P \times E \times R:E$) was applied where S, P and E were treated as fixed and R:E as random.

Variance components were computed for each of the facets and for their interactions. EduG estimated the variance components by applying a Whimbey’s correction to conventional ANOVA estimates, which accounts for the type of sampling involved (i.e. random, fixed or random finite). This calculation of variance components determined how much variance is due to the actual degree of severity of the pictured mice, the perspective the photo was taken, the raters’ experience level and the raters themselves. This procedure was first carried out for all eight images and subsequently in two further and separate analyses for the extreme severity levels and the middle severity levels.

Table 2. Potential factors influencing the rating of the clinical presentation in photographs of mice investigated with generalisability study.

	Facet Effect (With Facet Levels)	Meaning (If Facet Accounts Substantially for Variance)
Characteristics of images	Severity (S) (none, mild, moderate, severe)	The actual degree of severity shown by the pictured mouse ² influences the severity rating of the pictured mouse. (This should be by far the largest source of variance.)
	Perspective (P) (full body, head)	The perspective in which the mouse is presented influences the severity rating of the pictured mouse.
Characteristics of raters	Rater experience (E) (beginner, advanced, professional)	The degree of experience regarding the work with mice influences the severity rating of the pictured mouse.
	Rater (R) ¹	The severity rating of the pictured mouse varies across the raters. (A very small variance component for the R:E facet would reflect a high inter-rater reliability.)

¹Nested in experience group, as each rater can only belong to one experience group;

²as rated by experts (see the Material and Methods section).

Results

Descriptive statistics for the images

The descriptive statistics for each image and experience group are presented in Table 3. No substantial differences were observed among the different rater groups in terms of the mean severity ratings. A significant difference between the groups for each image was only found for Image 7 ($p = 0.02$). The mean absolute deviations from the reference scores across all raters varied between 0.06 (Image 7) and 0.25 (Image 6) for the images showing the extremes none and severe), whereas it differed between 0.56 (Image 5) and 0.89 (Image 3) for the images showing intermediate (mild and moderate) degrees of severity. The Levene's test revealed variance heterogeneity for the mean deviations from the reference scores regarding the ratings of the extreme compared to the intermediate degrees of severity ($p < 0.000$). Thus, a Welch test was performed for further analysis, showing a significant difference ($p < 0.000$). Fig 2 also shows that the middle categories were rated less correctly and also less consistently, as indicated by the higher variance found for the intermediate-degree images.

Generalisability theory

The results of the descriptive statistics were strongly supported by the generalisability analyses (Table 4). Differences from the overall grand mean, which is the mean rating across all images and raters, were highest in the severity (S) facet, followed by the rater nested in the experience groups (R:E) facet. The differences were negligible for the experience (E) facet. The significant differences in the severity facet were especially due to the relatively more accurate scoring and differentiation of the extreme severity levels from the 'mild' and 'moderate' levels. The raters were not able to differentiate distinctly between the latter two levels.

The results of the generalisability analyses are summarised in S2A Table for the rating of all eight images. It includes the classical ANOVA estimates for the severity, perspective, experience and rater experience group facets and their interactions. The '%' column shows the proportion of the variance that is attributable to each of the different sources of variance. As expected and desired, by far the largest degree of data variance was due to the severity (S facet; 56.3%). Although the extent of experience (E facet) did not contribute at all to the variance, 5.3% could be attributed to the rater experience group (R:E facet). An additional 11.6%, 5.9% and 18.1% were attributed to each of the interactions of the rater (R facet) with the

Table 3. Results of clinical appearance severity scoring based on rater experience presented as mean, Standard Deviation (SD), p-values and absolute deviation from the reference score.

Degree of Severity	Perspective	Beginners (n = 28)		Advanced (n = 28)		Professionals (n = 28)		p-value (ANOVA)	Absolute Deviation from the Reference Score	
		Mean	SD	Mean	SD	Mean	SD		Mean	SD
None	Entire mouse	1.25	0.44	1.21	0.50	1.14	0.36	0.65	0.20	0.43
	Head	1.18	0.48	1.07	0.26	1.00	0.00	0.11	0.08	0.32
Mild	Entire mouse	2.21	0.63	2.21	0.83	2.46	0.74	0.35	0.56	0.57
	Head	2.89	0.96	2.71	0.71	2.86	0.80	0.70	0.89	0.74
Moderate	Entire mouse	2.29	0.76	2.32	0.77	2.18	0.61	0.74	0.81	0.63
	Head	2.93	0.77	2.89	0.83	2.71	0.94	0.60	0.58	0.62
Severe	Entire mouse	3.71	0.54	3.79	0.42	3.75	0.44	0.85	0.25	0.46
	Head	3.82	0.48	4.00	0.00	4.00	0.00	0.02	0.06	0.28

<https://doi.org/10.1371/journal.pone.0287965.t003>

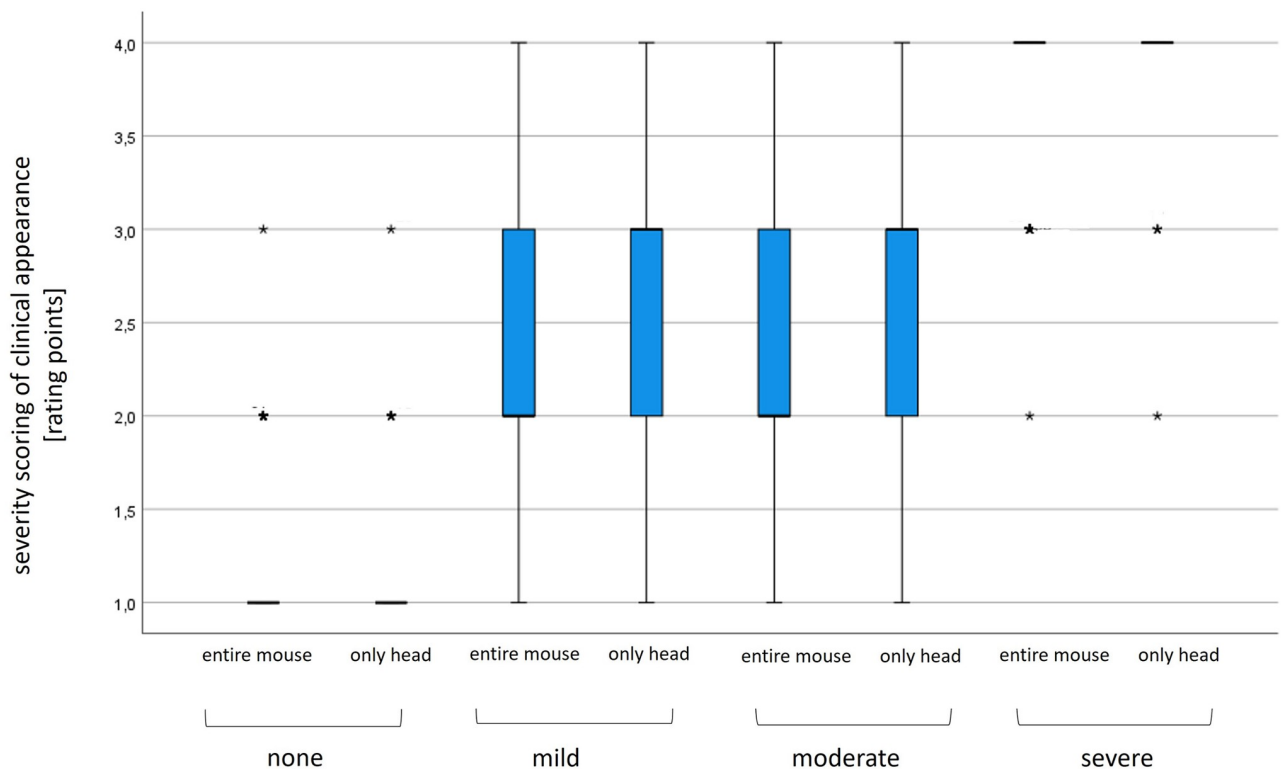


Fig 2. Box and whisker plot for all severity ratings of clinic appearance. Results are presented over all raters (n = 84) regardless of their laboratory animal experience and for each picture individually. *Single ratings of raters.

<https://doi.org/10.1371/journal.pone.0287965.g002>

Table 4. Descriptive statistics of the main effects (severity degree of clinical appearance of pictured mouse, perspective, experience, and rater) when rating the degree of severity of clinical appearance in pictured mice analysed using generalisability theory.

Main effects		Mean (SD)	$\Delta_{\text{to grand mean}}$
Severity degree of clinical appearance of pictured mouse	Grand mean	2.53 (1.15)	0
	None (1)	1.14 (.38)	1.39
	Mild (2)	2.56 (.82)	-0.03
	Moderate (3)	2.55 (.83)	-0.02
	Severe	3.85 (.39)	-1.32
	Perspective		
	Entire mouse	2.38 (1.09)	0.15
	Head	2.67 (1.20)	-0.14
	Experience		
	Beginner	2.54 (1.13)	-0.01
Rater*	Advanced	2.53 (1.16)	0
	Professional	2.51 (1.16)	0.02
	Rater _{min}	1.88 (1.05)	0.65
	...		
	Rater _{max}	3.13 (1.27)	-0.60

SD, standard deviation; Δ , difference; bold, values of special interest.

*As there were 84 raters only the two raters with either the minimum or maximum mean scores across the eight images are shown with both having the biggest Δ to grand mean.

<https://doi.org/10.1371/journal.pone.0287965.t004>

severity facet and perspective facet and to the multiple interactions of the rater experience group, severity and perspective (R:E, S and P) facets, respectively.

The results of the separate analyses for the extreme severity levels (none and severe) and the middle severity levels (mild and moderate) are also shown in [S2B and S2C Table](#). Although in the analysis of the extreme severity levels the severity facet was predominantly the unique source of variance (91.6%), it did not contribute at all to the variance in the middle severity levels. Here the variance was due to the rater experience group facet (R:E 13.2%), perspective facet (6.6%) and the interactions of the rater experience group (R:E) facet with the perspective facet (15.6%) and severity facet (23%). As in the analysis for all eight images, the extent of experience (E facet) did not contribute at all to the variance in both analyses.

Discussion

The aim of the study was to determine how well the different degrees of severity can be differentiated by raters and whether the extremes of a rating scale are easier to determine than are the intermediate degrees. Additionally, we investigated the extent that the rating might be influenced by factors other than the presented degree of severity. These factors included the characteristics of the raters (experience) and of the items being rated (perspective of the mice on the images), as well as their interactions.

The extremes of the scale were found easier to rate, and if only the extreme levels were considered, the actual degree of severity was almost entirely the unique source of variance. However, the degree of severity did not contribute to the variance in the intermediate degrees at all. Furthermore, the raters' experience had no influence on the rating, only in combination with other facets.

The study confirmed a higher homogeneity for the ratings of the extreme degrees of the scale (none and severe) compared with the intermediate degrees of severity (mild and moderate) and a smaller deviation from the reference scores. These findings were also supported by the generalisability analysis, where the resulting variance of the ratings of the extremes of the severity scale was nearly completely due to the degree of severity itself (Severity facet; 91.6%). In contrast, the severity degree of clinical appearance of the pictured mouse did not contribute to the variance in the ratings of the intermediate degrees of severity. Here, the interaction between the different facets was decisive, as well as the facet of the raters nested in experience and the facet of the perspective. These results are in line with previous research, especially with regard to farm animals, since no literature on laboratory animals was found. Garcia et al. [13] found that in a video-based mobility scoring of dairy cows, the extremes of the scoring scale were easier to assess than were the grades in between. In addition, clinically detectable lame animals were more easily identified by most of the raters. In contrast, a differentiation between physiological gait and subclinical altered moving patterns seems to be more difficult [16]. This indicates that slight alterations to the normal state usually being scored in the grades in between the extremes are more difficult to be scored precisely and consistently even by experienced or trained raters [14, 15].

Focusing on the influencing factors upon the ratings, the generalisability analysis indicated that the actual degree of severity shown in the pictured mice was as expected and, as it should be, by far, the biggest source of variance when all eight images were analysed. Other sources of variance were the raters themselves, interactions between the raters, the actual degree of severity of the pictured mice and the perspective in which the photograph was taken. The variance due to the raters and the interactions of severity and perspective with raters is attributed to the severity ratings of the photos showing the intermediate degrees of severity.

Rather unexpectedly, the perspective and its interactions also contributed to variance. Overall, the images showing only the head of the mouse were rated slightly higher regarding the severity of clinical appearance than were the photos of the same severity degree of clinical appearance showing the entire mouse. This effect was particularly pronounced for the analysis of the intermediate levels of severity. One influencing factor might have been the earclip being much more dominant in the photographs only showing the head. Raters were not familiar with this labelling method and might have defined this treatment as inappropriate and welfare relevant. This might also have influenced the ratings, since the ear position was altered because of the metal clip. The effect of focusing on one specific criterion only, in this case the ear clip, and drawing conclusions about the overall well-being of the animal seems to be a very human way of seeing things. In psychology, it is known as the halo effect, meaning that a criterion induces an assumption of how the rest (of the animal) has to look like and might lead, to a false impression of how the subject is burdened [24, 25]. This might also have been the case in this study because of the negation of the parameters 'fur appearance' and 'body position', which were not shown in detail, and the specific focus on the head of the animal.

Interestingly, the experience of the raters did not contribute to the variance in any of the analyses. Renner et al. [26] found that differences in behaviour of rats living in either enriched or in impoverished environments, were detected equally successful, regardless of the raters' experience. In line with these findings, Garcia et al. [13] showed that, in assessing cow lameness, within-observer agreements of rating did not differ, even after training, between experienced and inexperienced raters.

Nevertheless, experience and training might be useful in perceiving subtler variations in physical conditions [27], as well as for pain scoring [28] and data gathering [26]. Additionally, since the ratings are linked to what the observer expects, the observers' range of experience seems to be important [8]. Controversially, this concept was not supported by the present study, likely because of the sole use of photographs, as well as the lacking information regarding the parameters accounting for the severity assessment.

This draws the attention to the limitations of this study. First is the sole use of images instead of including living mice or videos. When thinking about material to use for these kind of studies and as training material, different circumstances must be considered. The most important one seems to be the unethical nature of conducting animal experiments related to harmful procedures where the animals show specific and repeatable signs of deviations from normal appearance just for training purposes. Also, this would not support the ethical aspect of protecting laboratory animals and would be contrary to the '3R principles' [29]. However, when using images or videos, one has to focus on the quality of the material used. In our case, we found some images that were not as good as the other ones because of improper illumination or glare ('moderate' images of the whole mouse and 'severe' images of the head of the mouse). These factors might influence the ratings, besides the perspective itself. The use of video sequences was also discussed in other publications assessing animal welfare statuses of sows and sheep [30, 31]. Nevertheless, many authors use photographs in assessing severity, especially in mice [17, 18]. Foddai et al. [32] reported that scores assigned to photographs were even less variable than were those from video sequences. Miller et al. found that live scores are significantly lower than scores from images [33]. When comparing videos with photographs some clinical signs would be more appropriately and accurately rated with videos, e.g. gait, some dynamic behaviours such as epilepsy, speed of ambulation and other movements. However, as demonstrated here other clinical signs can also be scored quite accurately with photographs [33].

Another drawback regarding the used material was the use of photographs depicting black-coloured mice since whisker position might be more difficult to score [34]. However, this

effect might be neglectable when using good contrasting backgrounds in photographs or recordings [18]. It has been suggested that a more accurate score may be achieved in the Mouse Grimace Scale in animals with dark coat colours [35].

Another limitation of this study is the assignment to the experience groups. The question arises as to how a beginner is defined. Even people who have worked less than a year with animals might have a better knowledge than people working in the field for up to 5 years which may be related to the actual time spent with animals rather than to the time spent in a job (e.g., checking on the animals and caring for them on a regular basis such as animal caretakers or post-graduate students). We recommend assessing the frequency of animal contact in future studies. Combined with years of experience, this would give a better impression of the experience level and might have a different impact on the ratings.

In addition, another influence on the rating might be background information on the parameters rated. In the present study, raters were deliberately left without information regarding the parameters used for the reference scores (S1 Table), and only the overall degree of severity was given. The purpose was to get an unbiased impression on how well the degrees of severity can be determined without any prior instructions and to widen the effect of experience that might have been found throughout the different groups.

Knowing the parameters on which the assessment should be based beforehand may produce a more precise outcome in rating the intermediate degrees of severity. The definition of terms and training previously showed an improvement in standardisation [27] and scoring [36]. A method which could also be used for defining appropriate parameters is the Qualitative Behaviour Assessment, which was e.g. used by Wemelsfelder et al. in cattle [37]. Formal training programmes are especially recommended for animal welfare assessors to reduce inter- and intra-observer variation when focusing on animal-based measures [38] and to increase reliability and agreement among raters [10, 39]. This goes in line with the results of the present study, since we strongly suggest that people working with animals have to be trained to detect and categorize mild to moderately affected mice based on their clinical appearance reliably. A huge impact on animal welfare could be obtained, since early intervention can prevent animals from reaching a 'severe' state. That way we could reduce the animals' pain, distress or lasting harm and would contribute to refine animal experiments as stated in '3R principles' [29]. Also, the comparison of the severity experienced by an animal or an interventional cohort and the severity classification of that model given in the EU Directive would benefit to the animals' welfare, since humane endpoints have to be applied once the severity is exceeded. Therefore, a new semiquantitative assessment has been developed by Morton [5]. Overall, the accurate ratings of degrees of severity need closer attention from the community and clearer parameters on which the rating decisions should be based. Furthermore, parameters, which are to be evaluated when rating the degrees of severity of animals, have to be known and clearly defined [5]. Regular teaching and training sessions are necessary and required to maintain good animal welfare, regardless of the background experience.

Conclusion

This pilot study used photographed mice showing different degrees of severity in their clinical appearance to evaluate the rating behaviour of people with different expertise in laboratory animal science (beginners, advanced, professionals). The grading of (un)altered clinical appearance (none, mild, moderate, severe) resulted in highly accurate ratings for the extremes (none and severe) but considerable variability for intermediate degrees (mild and moderate). Therefore, we suggest further education, training and overall deepening of people's knowledge about animal behaviour and appearance. Training outcomes need to be tested and retested so

that a certain standard is maintained and raters do not incorporate their own expectations or individual interpretations of the criteria [40]. Further research is necessary to determine whether improvements in detailed grade explanation of scores and additional objectivity-based items should be added to differentiate the degree of severity in animal experiments more precisely, especially among intermediate degrees of severity.

Supporting information

S1 Questionnaire.

(DOCX)

S1 Table. Variables comprising the reference scores.

(DOCX)

S2 Table. ANOVA for the rating of the different degrees of severity of clinical appearance in images of mice, using the S x P x E x R:E design of the G theory with interactions.

(DOCX)

Acknowledgments

The authors would like to thank Dr. S. Talbot and S. Bruch for statistical support and Dr. M. Afify for reviewing the manuscript. Additional thanks go to the *Journal of Comparative Medicine* and to all participants of the survey.

Author Contributions

Conceptualization: Johanne C. Krueger, Rene H. Tolba.

Data curation: Johanne C. Krueger.

Formal analysis: Johanne C. Krueger, Maren Boecker.

Funding acquisition: Rene H. Tolba.

Investigation: Johanne C. Krueger.

Methodology: Johanne C. Krueger, Maren Boecker.

Resources: Rene H. Tolba.

Supervision: Andre Bleich, Rene H. Tolba.

Visualization: Johanne C. Krueger, Maren Boecker.

Writing – original draft: Johanne C. Krueger.

Writing – review & editing: Maren Boecker, Siegfried Gauggel, Andre Bleich, Rene H. Tolba.

References

1. Dwyer C.M., Welfare of sheep: Providing for welfare in an extensive environment. *Small Ruminant Research*, 2009. 86(1): p. 14–21.
2. Brambell, F., *Technical Committee to Enquire into the Welfare of Animals kept under Intensive Livestock Husbandry Systems. 1965.* Report of the technical committee to enquire into the welfare of animals kept under intensive livestock husbandry conditions. London: Her Majesty's Stationary Office.
3. Commission E., Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes. *Off. J. Eur. Union*, 2010. 50: p. 33–79.
4. Keubler L.M., et al., Where are we heading? Challenges in evidence-based severity assessment. *Laboratory animals*, 2020. 54(1): p. 50–62. <https://doi.org/10.1177/0023677219877216> PMID: 31718424

5. Morton D.B., A Model Framework for the Estimation of Animal "Suffering": Its Use in Predicting and Respectively Assessing the Impact of Experiments on Animals. *Animals*, 2023. 13(5): p. 800. <https://doi.org/10.3390/ani13050800> PMID: 36899657
6. Morton D.B. and Griffiths P.H.M., Guidelines on the recognition of pain, distress and discomfort in experimental animals and an hypothesis for assessment. *Vet Rec*, 1985. 116(16): p. 431–6. <https://doi.org/10.1136/vr.116.16.431> PMID: 3923690
7. Carstens E. and Moberg G.P., Recognizing pain and distress in laboratory animals. *Ilar j*, 2000. 41(2): p. 62–71. <https://doi.org/10.1093/ilar.41.2.62> PMID: 11304586
8. Meagher R.K., Observer ratings: Validity and value as a tool for animal welfare research. *Applied Animal Behaviour Science*, 2009. 119(1): p. 1–14.
9. Burghardt G.M., et al., Perspectives—minimizing observer bias in behavioral studies: a review and recommendations. *Ethology*, 2012. 118(6): p. 511–517.
10. Gibbons J., et al., A training programme to ensure high repeatability of injury scoring of dairy cows. *Animal Welfare-The UFAW Journal*, 2012. 21(3): p. 379.
11. Marsh D.M. and Hanlon T.J., Observer gender and observation bias in animal behaviour research: experimental tests with red-backed salamanders. *Animal Behaviour*, 2004. 68(6): p. 1425–1433.
12. Rosenthal, R., *Experimenter effects in behavioral research*. 1966.
13. Garcia E., et al., Experienced and inexperienced observers achieved relatively high within-observer agreement on video mobility scoring of dairy cows. *J Dairy Sci*, 2015. 98(7): p. 4560–71. <https://doi.org/10.3168/jds.2014-9266> PMID: 25935241
14. Schlageter-Tello A., et al., Effect of merging levels of locomotion scores for dairy cows on intra- and interrater reliability and agreement. *Journal of Dairy Science*, 2014. 97(9): p. 5533–5542. <https://doi.org/10.3168/jds.2014-8129> PMID: 24996266
15. Kaler J., Wassink G.J., and Green L.E., The inter- and intra-observer reliability of a locomotion scoring scale for sheep. *The Veterinary Journal*, 2009. 180(2): p. 189–194. <https://doi.org/10.1016/j.tvjl.2007.12.028> PMID: 18308594
16. Winckler C. and Willen S., The reliability and repeatability of a lameness scoring system for use as an indicator of welfare in dairy cattle. *Acta Agriculturae Scandinavica, Section A-Animal Science*, 2001. 51 (S30): p. 103–107.
17. Nunamaker E.A., et al., Predictive observation-based endpoint criteria for mice receiving total body irradiation. *Comparative medicine*, 2013. 63(4): p. 313–322. PMID: 24209966
18. Langford D.J., et al., Coding of facial expressions of pain in the laboratory mouse. *Nature Methods*, 2010. 7: p. 447. <https://doi.org/10.1038/nmeth.1455> PMID: 20453868
19. Beynen A.C., et al., Assessment of discomfort in gallstone-bearing mice: a practical example of the problems encountered in an attempt to recognize discomfort in laboratory animals. *Lab Anim*, 1987. 21 (1): p. 35–42. <https://doi.org/10.1258/002367787780740770> PMID: 3560862
20. Urbaniak, G. and S. Plous, *Research randomizer (version 4.0)[computer software]*. 2013. <http://www.randomizer.org/> (accessed June 22, 2013), 2013.
21. Cronbach L.J., Rajaratnam N., and Gleser G.C., Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 1963. 16(2): p. 137–163.
22. Brennan R.L., Generalizability theory. Generalizability theory. 2001, New York, NY, US: Springer-Verlag Publishing. xx, 538–xx, 538.
23. Group, S.S.f.R.i.E.W., *EDUG user guide*. Neuchâtel, Switzerland: IRDP, 2006.
24. Thorndike E.L., A constant error in psychological ratings. *Journal of applied psychology*, 1920. 4(1): p. 25–29.
25. Sherbino J. and Norman G., On rating angels: the halo effect and straight line scoring. *Journal of Graduate Medical Education*, 2017. 9(6): p. 721–723. <https://doi.org/10.4300/JGME-D-17-00644.1> PMID: 29270261
26. Renner M.J. and Renner C.H., Expert and novice intuitive judgments about animal behavior. *Bulletin of the Psychonomic Society*, 1993. 31(6): p. 551–552.
27. Kristensen E., et al., Within- and Across-Person Uniformity of Body Condition Scoring in Danish Holstein Cattle. *Journal of Dairy Science*, 2006. 89(9): p. 3721–3728. [https://doi.org/10.3168/jds.S0022-0302\(06\)72413-4](https://doi.org/10.3168/jds.S0022-0302(06)72413-4) PMID: 16899709
28. Roughan J.V. and Flecknell P.A., Training in behaviour-based post-operative pain scoring in rats—An evaluation based on improved recognition of analgesic requirements. *Applied Animal Behaviour Science*, 2006. 96(3): p. 327–342.
29. Russell, W.M.S., R.L. Burch, and C.W. Hume, *The principles of humane experimental technique*. Vol. 238. 1959: Methuen London.

30. Phythian C.J., et al., Inter-observer agreement, diagnostic sensitivity and specificity of animal-based indicators of young lamb welfare. *Animal*, 2013. 7(7): p. 1182–90. <https://doi.org/10.1017/S1751731113000487> PMID: 23561038
31. Nalon E., et al., Comparison of the inter-and intra-observer repeatability of three gait-scoring scales for sows. *animal*, 2014. 8(4): p. 650–659. <https://doi.org/10.1017/S1751731113002462> PMID: 24438690
32. Foddai A., et al., Evaluating observer agreement of scoring systems for foot integrity and footrot lesions in sheep. *BMC Veterinary Research*, 2012. 8(1): p. 65. <https://doi.org/10.1186/1746-6148-8-65> PMID: 22630057
33. Miller A.L. and Leach M.C., The Mouse Grimace Scale: A Clinically Useful Tool? *PLoS One*, 2015. 10(9): p. e0136000. <https://doi.org/10.1371/journal.pone.0136000> PMID: 26406227
34. Whittaker A.L., Liu Y., and Barker T.H., Methods used and application of the mouse grimace scale in biomedical research 10 years on: a scoping review. *Animals*, 2021. 11(3): p. 673. <https://doi.org/10.3390/ani11030673> PMID: 33802463
35. Ernst L., et al., Improvement of the Mouse Grimace Scale set-up for implementing a semi-automated Mouse Grimace Scale scoring (Part 1). *Laboratory animals*, 2020. 54(1): p. 83–91. <https://doi.org/10.1177/0023677219881655> PMID: 31648592
36. Engel B., et al., Assessment of observer performance in a subjective scoring system: visual classification of the gait of cows. *The Journal of Agricultural Science*, 2003. 140(3): p. 317–333.
37. Wemelsfelder, F., et al., *Qualitative behaviour assessment*. Assessment of Animal Welfare Measures for Sows, Piglets and Fattening Pigs, 2009: p. 215–224.
38. EFSA recommends use of animal-based measures when assessing welfare. *Veterinary Record*, 2012. 170(5): p. 112–112. <https://doi.org/10.1136/vr.e776> PMID: 22311310
39. Vasseur E., et al., Development and implementation of a training program to ensure high repeatability of body condition scoring of dairy cows. *J Dairy Sci*, 2013. 96(7): p. 4725–37. <https://doi.org/10.3168/jds.2012-6359> PMID: 23660141
40. Kazdin A.E., Assessing the Clinical or Applied Importance of Behavior Change through Social Validation. *Behavior Modification*, 1977. 1(4): p. 427–452.