animals

Article

Reliability of the Mouse Grimace Scale in C57BL/6JRj Mice

Katharina Hohlbaum [1,*], Giuliano Mario Corte [2], Melanie Humpenöder [1], Roswitha Merle [3,†] and Christa Thöne-Reineke [1,†]

[1] Institute of Animal Welfare, Animal Behavior and Laboratory Animal Science, Department of Veterinary Medicine, Freie Universität Berlin, 14163 Berlin, Germany; Melanie.Humpenoeder@fu-berlin.de (M.H.); Thoene-Reineke.Christa@fu-berlin.de (C.T.-R.)

[2] Institute of Veterinary Anatomy, Department of Veterinary Medicine, Freie Universität Berlin, 14195 Berlin, Germany; giuliano.corte@fu-berlin.de

[3] Institute for Veterinary Epidemiology and Biostatistics, Department of Veterinary Medicine, Freie Universität Berlin, 14163 Berlin, Germany; Roswitha.Merle@fu-berlin.de

[*] Correspondence: katharina.hohlbaum@fu-berlin.de

[†] These authors contributed equally to this work.

Received: 27 July 2020; Accepted: 10 September 2020; Published: 14 September 2020

Simple Summary: The protection and welfare of laboratory animals is ethically and legally required. Any possible impairment of well-being caused by husbandry or experiment must be reduced to a minimum. A prerequisite for ameliorating animal welfare is that we recognize the current well-being state of the animals by using measuring tools, for instance, the grimace scale. Facial expressions of mice, the most frequently used laboratory animals, can be analyzed using the Mouse Grimace Scale, which consists of five action units, i.e., orbital tightening, nose and cheek bulge, ear position, and whisker change. Since it is important that the animal welfare measuring tools are reliable and objective, we investigated the reliability of the Mouse Grimace Scale. Our results indicated good agreement between all observers applying the Mouse Grimace Scale. However, when investigating the action units individually, the best agreement was achieved for orbital tightening and the poorest agreement for nose as well as cheek bulge, depending on the observers' experience levels. Against this background, we critically discuss factors that potentially influence the reliability of the Mouse Grimace Scale to improve its application.

Abstract: To maintain and foster the welfare of laboratory mice, tools that reliably measure the current state of the animals are applied in clinical assessment. One of these is the Mouse Grimace Scale (MGS), a coding system for facial expression analysis. Since there are concerns about the objectivity of the MGS, we further investigated its reliability. Four observers (two experienced and two inexperienced in use of the MGS) scored 188 images of 33 female and 31 male C57BL/6JRj mice. Images were generated prior to, 150 min, and two days after ketamine/xylazine anesthesia. The intraclass correlations coefficient (ICC = 0.851) indicated good agreement on total MGS scores between all observers when all three time points were included in the analysis. However, interrater reliability was higher in the early post-anesthetic period (ICC = 0.799) than at baseline (ICC = 0.556) and on day 2 after anesthesia (ICC = 0.329). The best agreement was achieved for orbital tightening, and the poorest agreement for nose and cheek bulge, depending on the observers' experience levels. In general, experienced observers produced scores of higher consistency when compared to inexperienced. Against this background, we critically discuss factors that potentially influence the reliability of MGS scoring.

Keywords: laboratory mice; animal welfare indicators; Mouse Grimace Scale; intraclass correlations coefficient; Fleiss' kappa; Cohen's kappa

1. Introduction

The protection and welfare of laboratory animals is ethically and legally required [1]. Any possible pain, suffering, distress, and lasting harm caused by husbandry or experiment must be reduced to a minimum [1], which also contributes to the quality of research findings [2]. A prerequisite for ameliorating animal welfare is that we recognize the current well-being state of an individual. When impaired well-being is identified, the animal can be treated accordingly.

The assessment of well-being is a serious challenge, especially in mice, the most commonly used laboratory animals in biomedical research. Mice are prey animals and therefore may often hide signs of pain, injury, and weakness when humans are present [3]. This underlines that a major current animal welfare issue is to develop and further refine tools that reliably measure the well-being of an animal. Humans have a tendency to focus on the face of an animal when assessing its current well-being [4]. Indeed, changes in facial expressions can indicate pain as shown for mice [5] and other animals such as rats [6], rabbits [7], cats [8,9], ferrets [10], sheep [11], piglets [12], cows [13], and horses [14]. However, grimacing can also be induced by other factors. For instance, the facial expression of mice can be affected by vibrissae contact, social proximity, cat odor exposure, rat exposure, aggression, or subordination in the resident–intruder test, and anesthesia [15–17]. To facilitate facial expression analysis in mice, the Mouse Grimace Scale (MGS) was developed [5] and has become a valuable tool in laboratory animal science in recent years; it is used in biomedical research as well as in cage-side clinical assessment [18,19]. The MGS includes five so-called facial action units (FAU), i.e., orbital tightening, nose bulge, cheek bulge, ear position, and whisker change [5]. The presence and intensity of these features can provide an insight into the affective state of a mouse.

Experts in laboratory animal welfare consider the MGS as an important welfare indicator [20]. However, concerns about the reliability of the MGS are frequently raised because, at this point, it is usually applied by human observers and, therefore, MGS scores are assumed to be subjective to a certain extent.

We had similar experiences in a previous study where four persons, who had a professional background in animal science but different experiences in the use of the MGS, were recruited to score images of C57BL/6JRj mice generated before and after ketamine/xylazine anesthesia [17]. After the MGS scoring in Hohlbaum et al. (2018) [17], the four observers reported that they found some of the FAU units difficult to score (personal communication).

Therefore, as a follow-up to our previous study [17], we conducted an interrater reliability analysis to determine whether the MGS data reflected the scorers' subjective impression. We also investigated interrater effects on the individual facial features and critically discussed factors that may influence the reliability of the MGS scoring.

2. Methods

2.1. Animals and Handling Methods

A total number of 64 adult C57BL/6JRj mice (33 females, 31 males; 10–13 weeks of age) were obtained from Janvier Labs (Saint-Berthevin Cedex, France) [17]. Sample size was calculated for our previous study on the impact of repeated ketamine/xylazine anesthesia on the well-being of mice and can be found in Hohlbaum et al. [17]. Animals were kept in a conventional (non-SPF) facility and were free of all viral, bacterial, and parasitic pathogens listed in the FELASA (Federation for Laboratory Animal Science Associations) recommendations [21]. Females were kept in groups of three to five animals in Makrolon type IV cages (55 × 33 × 20 cm); males had to be kept individually in Makrolon type III cages (42 × 26 × 15 cm) because of aggressive behavior [17]. The cages were filled with fine wooden bedding material (LIGNOCEL® 3-4 S, J. Rettenmaier & Söhne GmbH + Co. KG, Rosenberg, Germany), nestlets (Ancare, UK agents, Lillico, United Kingdom), a red plastic house (length: 100 mm, width: 90 mm, height: 55 mm; ZOONLAB GmbH, Castrop-Rauxel, Germany), and metal tunnels (length: 125 mm, diameter: 50 mm; one tunnel in Makrolon type III cages, two tunnels in Makrolon type IV cages). The animals were maintained under standard conditions (room temperature: 22 ± 2 °C; relative humidity: 55 ± 10%) on a light:dark cycle of 12:12 h of artificial light with a 5 min twilight transition phase (lights on from 6:00 a.m. to 6:00 p.m.) [17]. Pelleted mouse diet was fed ad libitum (Ssniff rat/mouse maintenance, Spezialdiäten GmbH, Soest, Germany) [17]. The mice had free access to tap water [17]. Animals were handled using a combination of the tunnel and cup techniques [22]. After the experiment was completed, female animals were rehomed; male mice were either used for educational purposes or castrated and resocialized in stable groups to be rehomed.

2.2. Anesthesia

Detailed description of anesthetic and monitoring procedures is available in Hohlbaum et al. [17]. In brief, a dosage of 80 mg/kg ketamine and 16 mg/kg xylazine (Ketavet® 100 mg/mL, Zoetis Deutschland GmbH, Berlin, Germany; Rompun® 2%, Bayer Vital GmbH, Leverkusen, Germany) was administered intraperitoneally at a volume of 10 µL/g body weight using 27 $\frac{3}{4}$ gauge needles [17]. After the loss of the righting reflex, mice were placed in abdominal position on a heating pad. The temperature of the heating pad could be adjusted from 30–42 °C. Artificial tears (Artelac® Splash MDO®, Bausch & Lomb GmbH, Berlin, Germany) were administered to the eyes [17]. Reflexes and vital parameters were carefully monitored during anesthesia, as reported in Hohlbaum et al. [17].

The mice were either anesthetized once (13 females, 13 males) or six times (i.e., every three to four days; 13 females, 12 males). In the latter case, images generated after their last anesthesia were examined. Control animals (7 females, 6 males) did not receive anesthesia. Since the effects of the different anesthesia regimes on the well-being of the animals had already been investigated in Hohlbaum et al. [17] and are outside the focus of this article, we do not differentiate between the treatment groups in the present article.

2.3. Mouse Photography and Grimace Scale Scoring

MGS scores were obtained from photographs (n = 188; four female mice were excluded on day 2 due to technical malfunction of the camera) generated two days prior to anesthesia (baseline), at 150 min, and two days after anesthesia, as previously described [17]. In brief, photographs were generated in a custom-made box (22 × 29 × 39 cm, three white walls, one transparent wall, 0.5 cm bedding material including soiled bedding) using a high definition camera (Canon EOS 350D, Canon Inc., Tokyo, Japan) [23].

The observers were blinded to the time point of image generation. Among them were three veterinarians and one animal technician. Due to their professional backgrounds, all of them were highly familiar with animal behavior. Scorers 1 and 2 had no previous experience in using the MGS scale, scorers 3 and 4 had already gathered experience in MGS scoring for approximately eleven months. They received electronic files containing the mouse images [23] and scored them according to the MGS at a time that was convenient to them and at their own pace (i.e., the scorers did not simultaneously evaluate the images). In brief, the five FAU (i.e., orbital tightening, ear position, whisker change, nose bulge, and cheek bulge,) were scored on a 3-point scale from 0 to 2 (0 = not present, 1 = moderately present, 2 = obviously present) [5]. An additional score "not assessable" was introduced for those cases in which a facial action unit could not be scored. Each observer scored 188 images that were then further analyzed as follows: The scores of the five facial action units were added up (overall MGS scores ranging from 0 to 10) and averaged (mean MGS scores), whereby facial action units scored "not assessable" were excluded. To calculate the mean MGS difference score, baseline MGS scores were subtracted from MGS scores obtained after anesthesia.

It should be noted that the score "not assessable" was included for scores of the individual FAUs, but excluded for overall MGS scores, mean MGS scores, as well as mean MGS difference scores.

2.4. Statistical Analysis

To assess the interrater reliability between the four scorers, weighted Fleiss' kappa (κ) statistics or intraclass correlation coefficients (ICC) were used, depending on whether the data were measured on an ordinal or continuous scale [24]. The single facial action units were scored on a 3-point scale, resulting in ordinal data. Hence, the weighted Fleiss' kappa was the most appropriate reliability parameter. However, since other working groups reported the ICC instead of the Fleiss' kappa, we also analyzed the ICC (model: two-way random model with observer as random factor, type: absolute agreement) for each facial action unit, in order to be able to compare the data. Overall MGS scores, mean MGS scores, and mean MGS differences scores were considered to be continuous and were analyzed using the ICC (model: two-way random model with observer as random factor, type: absolute agreement). Interrater reliability for the two inexperienced and the two experienced scorers was determined using Cohen's kappa.

ICC values were evaluated according to Koo and Li [25]:

- <0.50 poor interrater reliability
- 0.50–0.75 moderate interrater reliability
- 0.75–0.90 good interrater reliability
- >0.90 excellent interrater reliability

Fleiss' kappa (κ) and Cohen's kappa was interpreted using guidelines stated by Landis and Koch [26]:

- <0 poor interrater reliability
- 0.00–0.20 slight interrater reliability
- 0.21–0.40 fair interrater reliability
- 0.41–0.60 moderate interrater reliability
- 0.61–0.80 substantial interrater reliability
- 0.81–1 almost perfect interrater reliability

Statistical analysis was performed with IBM SPSS Version 25 and Version 26 (IBM Corporation, Armonk, NY, USA). An extension of IBM SPSS was used for Fleiss' kappa statistics [27]. Differences between facial action units were analyzed using Friedman's two-way analysis of variance by ranks with a significance level of 0.05.

2.5. Ethical Approval and Informed Consent

The study was performed according to the guidelines of the German Animal Welfare Act and the Directive 2010/63/EU for the protection of animals used for scientific purposes. Maintenance of mice and all animal experimentation were approved by the Berlin State Authority ("Landesamt für Gesundheit und Soziales", permit number: G0053/15). The purpose of the study was to investigate the impact of repeated ketamine/xylazine anesthesia on the well-being of C57BL/6JRj mice [17]. In the present work, we investigated and discussed the MGS scores obtained in this study in more detail.

3. Results

For overall MGS scores (i.e., total MGS scores, mean MGS scores, and mean MGS difference score; Figure 1), the intraclass correlation coefficient (ICC) revealed good agreement (Table 1) when scores of each mouse obtained at baseline, 150 min, and two days after anesthesia were included in the analysis. When the time points were analyzed separately, interrater reliability deteriorated at baseline (moderate) and on day 2 after anesthesia (poor), but remained good at 150 min after anesthesia (Table 1). While experienced scorers reached good agreement in overall MGS scores over all time points, those from inexperienced scorers indicated moderate to good agreement (Table 1). The ICC values over all time points did not improve by excluding the scores of any scorer. Figure 2 illustrates the reliability of the mean MGS scores. Interestingly, the MGS scores obtained from scorers 1 and 2 appeared to be higher than those generated by scorers 3 and 4 (Figure 2).

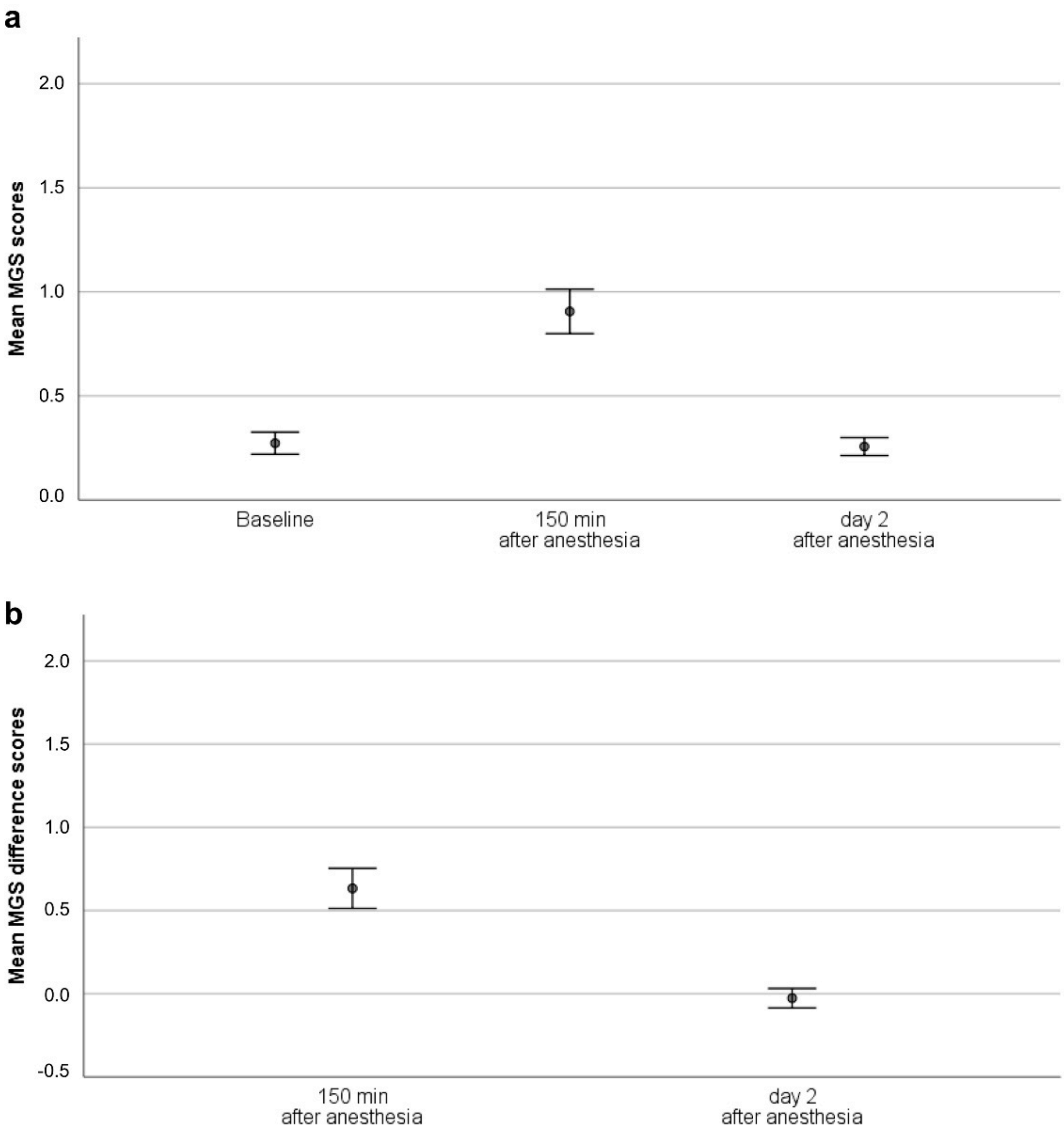

Figure 1. MGS scores (mean, 95% confidence interval). (**a**) Mean MGS scores, (**b**) Mean MGS difference scores. Please note that data were derived from animals that received either no, one, or six anesthesia. Since the present study focuses on MGS reliability, treatment groups are not differentiated. The effects of the anesthesia regimes on the well-being of mice in the different treatment groups was investigated in a previous study by Hohlbaum et al. [17]. Abbreviation: Mouse Grimace Scale (MGS).

Table 1. Interrater reliability of overall MGS scores.

	Total MGS Scores [1]	Mean MGS Scores [1]	Mean MGS Difference Scores [2]
Interrater Reliability between All Scorers			
ICC over all time points [3]	0.851	0.845	0.855
ICC at baseline	0.556	0.522	-
ICC at 150 min after anesthesia	0.799	0.802	0.813
ICC on day 2 after anesthesia	0.329	0.317	0.433
Interrater Reliability between Inexperienced Scorers			
ICC over all time points [3]	0.749	0.757	0.736
Interrater Reliability between Experienced Scorers			
ICC over all time points [3]	0.824	0.821	0.811
Interrater Reliability between Three Scorers, Excluding One Person at a Time			
ICC over all time points [3], excluding scorer #1 (inexperienced)	0.812	0.797	0.791
ICC over all time points [3], excluding scorer #2 (inexperienced)	0.826	0.822	0.855
ICC over all time points [3], excluding scorer #3 (experienced)	0.814	0.808	0.816
ICC over all time points [3], excluding scorer #4 (experienced)	0.791	0.786	0.799

[1] Scores obtained at baseline, 150 min, and two days after anesthesia. [2] Scores obtained at 150 min, and two days after anesthesia. Baseline scores were subtracted from the score generated after anesthesia. [3] Three scores of each mouse were included in the analysis (i.e., one score each was obtained from images generated at baseline, 150 min, and two days after anesthesia). Abbreviations: Intraclass correlation (ICC), Mouse Grimace Scale (MGS).

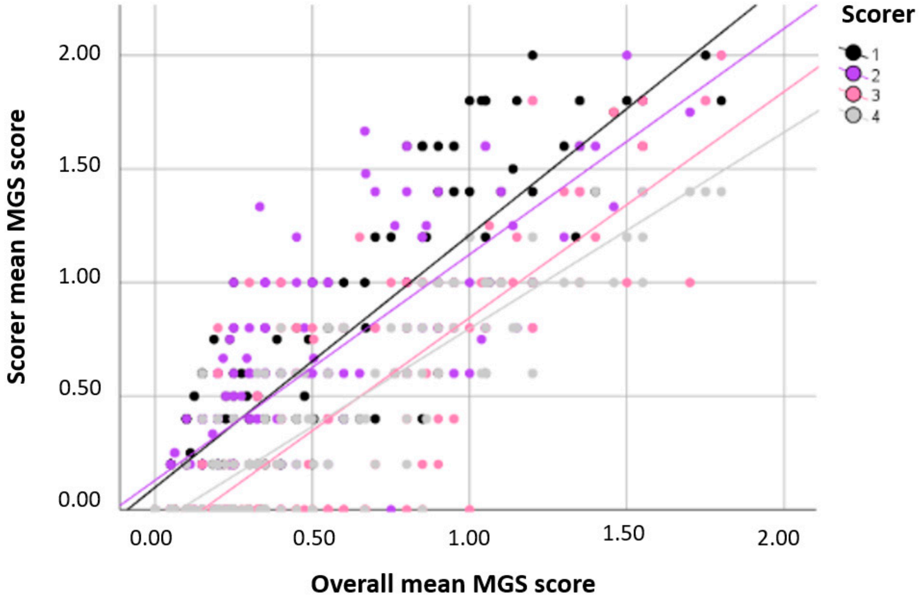

Figure 2. Reliability analysis of the mean MGS scores. The mean MGS scores of each of the four scorers were compared to the mean MGS scores of all scorers. Three scores for each mouse were included in the analysis (i.e., one score each was obtained from images generated at baseline, 150 min, and two days after anesthesia). Abbreviation: Mouse Grimace Scale (MGS).

Fleiss' kappa indicated substantial agreement between all four observers for orbital tightening, and fair agreement for ear position as well as whisker change, while slight agreement was reached for nose and cheek bulge, when scores of all time points were included in the analysis (Table 2). In some cases, the interrater reliability of the individual FAU scores differed between the three time points at which the images were acquired. Fleiss' kappa revealed a higher consistency in scores of orbital tightening and whisker change at 150 min after anesthesia (orbital tightening: moderate; whisker change: fair) than at baseline or on day 2 (orbital tightening: fair; whisker change: slight) (Table 2). The consistency in nose and cheek bulge scores were slight at baseline and 150 min after anesthesia but poor on day 2. Independent of the time point, ear position scores showed slight agreement.

Table 2. Interrater reliability of each facial action unit.

	Orbital Tightening	Nose Bulge	Cheek Bulge	Ear Position	Whisker Change
Interrater Reliability between All Scorers					
Fleiss' kappa over all time points [1]	0.664	0.093	0.125	0.285	0.279
Intraclass correlation over all time points [1,2]	0.876	0.156	0.146	0.542	0.519
Fleiss' kappa at baseline	0.245	0.104	0.032	0.164	0.066
Intraclass correlation at baseline [2]	0.416	0.117	0.041	0.173	0.174
Fleiss' kappa at 150 min after anesthesia	0.531	0.052	0.083	0.197	0.220
Intraclass correlation at 150 min after anesthesia [2]	0.900	0.243	0.115	0.674	0.721
Fleiss' kappa on day 2 after anesthesia	0.357	−0.089	−0.031	0.074	0.096
Intraclass correlation on day 2 after anesthesia [2]	0.695	0.098	0.102	0.282	0.450
Interrater Reliability between Inexperienced Scorers					
Cohen's kappa over all time points [1]	0.708	0.197	0.013	0.280	0.244
Interrater Reliability between Experienced Scorers					
Cohen's kappa over all time points [1]	0.629	0.355	0.372	0.484	0.428

The score "not assessable" was included in the analysis. [1] Three scores of each mouse were included in the analysis (i.e., one score each was obtained from images generated at baseline, 150 min, and two days after anesthesia). [2] To be able to compare our data with other studies, we calculated the ICC in addition to Fleiss' kappa. Abbreviation: Intraclass correlation (ICC).

Interrater reliability was separately analyzed for inexperienced and experienced observers using Cohen's kappa. Both inexperienced and experienced scorers achieved substantial agreement for orbital tightening. Moderate agreement between experienced observers for ear position as well as whisker change and fair agreement for nose and cheek bulge were achieved, whereas agreement between inexperienced observers was only fair for ear position as well as whisker change and slight for cheek and nose bulge.

4. Discussion

Due to ethical, scientific, and legal issues [1,2], it is crucial to develop scientifically sound tools to assess the well-being of mice so that those with disturbed well-being can be identified and treated accordingly. This implies that the tools can be easily applied in practice and reliably measure welfare. Since there are concerns about the objectivity of the MGS, we took a closer look at its reliability when scoring images of C57BL/6JRj mice generated prior to and after ketamine/xylazine anesthesia. We found good agreement between observers for overall MGS scores (i.e., including orbital tightening, nose bulge, cheek bulge, ear position, and whisker change), though the interrater reliability depended on the affective state of the mice, the observers' experience levels, and the scored facial features. The best agreement was achieved for orbital tightening, the poorest agreement for nose and cheek bulge.

The MGS was originally developed to identify acute pain from the facial expressions of mice [5], however, grimacing can also be induced by other stress factors [15–17]. In the present study, we assume that changes in facial expressions were primarily due to pharmacological effects of the injection anesthetics [17]. The fact that the mice did not experience acute pain is not relevant for this study, because it should not influence the interrater reliability. For instance, if a certain FAU does not change after anesthesia, the scorers should be able to recognize the lack of facial expression changes and assign appropriate scores. Nevertheless, against the background that anesthesia may cause lower MGS scores than acute pain, i.e., less obvious facial expression changes, it is conceivable that interrater reliability may be higher for faces of mice experiencing acute pain.

Our interrater reliability of mean MGS scores in C57BL/6JRj mice (ICC = 0.845 over all time points) was similar to that found by the MGS inventors (ICC = 0.90) in CD-1 mice [5] and other working groups in CD-1 mice (ICC = 0.8405) [28], BALB/c mice (ICC ranging from 0.75 to 0.84) [29], or BALB/cAnNCrl mice (ICC ranging from 0.7 to 0.84) [30]. In cats, the ICC between four persons applying the Feline Grimace Scale was 0.84 [31].

In the present study, the ICC values comparing overall MGS scores acquired from inexperienced or experienced observers did not improve by excluding the scores of any scorer, suggesting that the observers evaluated the images consistently.

Interestingly, images generated at 150 min after anesthesia were scored with a higher consistency when compared to baseline images or those acquired on day 2 after anesthesia. This phenomenon was particularly noticeable regarding orbital tightening as well as whisker change and indicated that facial expressions of C57BL/6JRj mice with undisturbed well-being, i.e., at baseline or on day 2, seemed more difficult to score. Vice versa, changes in facial expressions caused by the experimental treatment were easier recognizable for the observers, as expected above.

To investigate the reported difficulty of scoring certain facial features in the images of C57BL/6JRj mice, we analyzed the interrater reliability separately for the four observers and each FAU using Fleiss' kappa and found widely divergent results. The highest agreement was observed for orbital tightening, followed by ear position and whisker change, whereas cheek and nose bulge scorings were less homogeneous among the four observers. To be able to compare our findings to Costa et al. [28], who described the ICC instead of Fleiss' kappa for all FAU, we also calculated the ICC. It was not possible to measure the Fleiss' kappa for the data given in Costa et al. because in their study not all observers scored all images on the MGS. While the interrater reliability was similar for orbital tightening (ICC = 0.858) and only slightly different for ear position (ICC = 0.756), Costa et al. measured much higher values for nose (ICC = 0.557) and cheek bulge (ICC = 0.644).

This discrepancy may be explained by different factors that can either be observer-dependent (experience, training), method-dependent (image quality, mutual interference of FAU, level of difficulty to score a FAU), animal-dependent (recognizability of FAU due to fur color), or treatment-dependent (height of MGS scores).

Langford et al. previously demonstrated that the observers' experience levels in the use of the MGS can have critical influence on the accuracy of the MGS [5]. An accuracy of 97% and 81% were determined for experienced and inexperienced scorers, respectively [5]. The accuracy was shown to deteriorate when low instead of high resolution images were presented to the observers [5], which emphasized the importance of image quality for the MGS application. In our study, two inexperienced observers and two persons with 11-month experience in MGS scoring were recruited to score images with a very high resolution (3.456 × 2.304 pixels). A separate analysis of the interrater reliability of experienced and inexperienced scorers revealed that the first achieved higher agreements for overall MGS scores and for almost all FAU, except for whisker change. Experienced persons from our study reached similar ICC values for orbital tightening, nose bulge, cheek bulge, and ear position when compared to Costa et al. [28]. From this, we hypothesized that consistency in MGS scoring may increase with the observers' experience and, therefore, the level of experience can strongly affect the interrater correlation. However, Roughan and Sevenoaks found equal or even greater consistency in the scores generated for BALB/c mice by inexperienced observers when compared to experienced observers [30]. Novices may have performed worse in our study than in the latter work because the MGS manual and MGS poster provided by the NC3Rs (National Centre of Replacement, Refinement, and Reduction, United Kingdom of Great Britain) is limited to white-furred mice [5,32]. Therefore, the inexperienced scorers in our study were confronted with the difficulty of transferring facial expressions of white-furred to black-furred mice.

Another interesting observation to note is that the inexperienced scorers generally assigned higher scores to the mouse images. This may be explained by initial uncertainties in MGS scorings, which may tempt them to orient towards the intensity of easily accessible facial features when applying the MGS. Thus, the score of a FAU can easily be biased by the intensity of other FAU. This is not surprising, as humans have a tendency to focus on particular facial features in order to assess the affective state of another person, whereby the features of interest depend on the respective state [33]. A template covering all the other facial features may solve this problem but would increase the labor costs for MGS scoring.

Furthermore, we deem the training process important and postulate that the method used to introduce new scorers to the MGS may significantly influence interrater reliability. In our study, the two experienced scorers received one-to-one training; the two inexperienced scorers were trained together and had the opportunity to discuss their scores with the trainer as well as the other trainee. Against this background, we assumed the group training would produce a higher agreement between the scorers, which was not true in our study. The interrater reliability may be improved if trainees perform multiple training sessions, and if scores from the first round are reviewed and discussed with an expert in the second round [34]. Apart from this, Zhang et al. demonstrated a beneficial effect for the group discussion about the grimace scale scores that were assigned to the images [34]. As the grimace scale training varies between laboratories, ranging from reviewing the grimace scale manual to a single or multiple training sessions of varying length and intensity (e.g., one-to-one or group training including discussions), the effectiveness of the respective training and the actual proficiency should be tested [5,34–37].

Since agreement between inexperienced observers was highest for orbital tightening and lowest for nose as well as cheek bulge, we suggested that the former was the easiest and the latter were the hardest to evaluate. A reason for this may be the degree of detail recognition, which is higher for those facial features colored in contrast to the fur. In C57BL/6JRj mice, the ears are lighter- and eyes are darker-colored than the fur and the whiskers contrast well against the whitish background. The background of the images acquired in our study was white, because we utilized a custom-made photography cube with one clear and three white walls. In contrast to ears, eyes, and whiskers, the cheeks and nose bridge have the same color as the faces of the mice. In other species such as cats and horses, it has also been reported that the identification of some FAU is difficult in black-furred animals [8,14,31].

Evaluations of the interrater reliability in Rat Grimace Scale scoring revealed that orbital tightening had the highest and whisker change had the lowest ICC value [34]. Probably due to the difficulty in clearly recognizing and evaluating certain FAU, some working groups excluded nose and cheek bulge [19] or whisker change [38,39] from facial expression analysis in mice. Similar observations were made in cats: the highest ICC value was found for orbital tightening and the lowest for muzzle tension as well as whisker change [31]. Considering these challenges of scoring certain FAU, grimace scales that were developed more recently often comprise less parameters than earlier grimace scales.

Another strategy to consider the differences in interrater reliability would be to add different weights to the five FAU, depending on the degree of agreement.

The questions that have arisen cannot be investigated based on our data, as the number of scorers is too low and we did not follow up their progress in MGS scoring. To further improve the reliability of MGS, the issues raised need to be addressed. A further approach to generate more objective and reliable data is to standardize the experimental set-up for MGS scoring, including image/video generation, training, as well as scoring sessions, and to automatize the facial expression analysis of mice [40–42].

5. Conclusions

Coming back to our initial questions: Did the MGS data represent the observers' subjective impressions that some of the FAU units were difficult to score? In summary, we found that nose and cheek bulge appeared to be FAU with potential difficulties, especially for the inexperienced observers in our study. Moreover, facial expressions that were affected by the anesthetic treatment were scored more consistently than those without pharmacological influence. However, although our analysis is based on a large image dataset, we could only recruit four scorers, which is why the conclusion drawn in this article is limited and should be verified by further studies.

Author Contributions: Conceptualization: K.H., R.M., C.T.-R.; Project Administration: R.M., C.T.-R.; Funding Acquisition: C.T.-R.; MGS Scoring: G.M.C., M.H., C.T.-R.; Data Analysis: K.H., R.M.; Writing—Original Draft Preparation: K.H.; Writing—Review & Editing: K.H., G.M.C., M.H., R.M., C.T.-R. All authors have read and agreed to the published version of the manuscript.

Funding: Mouse Grimace Scale scores were obtained from a project of the Berlin-Brandenburg research platform BB3R (www.bb3r.de) that was funded by the German Federal Ministry of Education and Research (grant number: 031A262A) (www.bmbf.de/en/index.html).

Acknowledgments: The authors would like to thank Luise Grace Klass for proofreading the manuscript and Sabine Jacobs for assisting in MGS scoring. The publication of this article was funded by Freie Universität Berlin.

Conflicts of Interest: The authors declare no competing interests.

References

1. 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**, *ABI*, L276/33–L276/79.
2. Poole, T. Happy animals make good science. *Lab. Anim.* **1997**, *31*, 116–124. [CrossRef] [PubMed]
3. Stasiak, K.L.; Maul, D.; French, E.; Hellyer, P.W.; Vandewoude, S. Species-specific assessment of pain in laboratory animals. *Contemp. Top. Lab. Anim. Sci.* **2003**, *42*, 13–20. [PubMed]
4. Leach, M.C.; Coulter, C.A.; Richardson, C.A.; Flecknell, P.A. Are we looking in the wrong place? Implications for Behavioural-based pain assessment in rabbits (oryctolagus cuniculi) and beyond? *PLoS ONE* **2011**, *6*, e13347. [CrossRef] [PubMed]
5. Langford, D.J.; Bailey, A.L.; Chanda, M.L.; Clarke, S.E.; Drummond, T.E.; Echols, S.; Glick, S.; Ingrao, J.; Klassen-Ross, T.; Lacroix-Fralish, M.L.; et al. Coding of facial expressions of pain in the laboratory mouse. *Nat. Methods* **2010**, *7*, 447–449. [CrossRef]
6. Sotocinal, S.G.; Sorge, R.E.; Zaloum, A.; Tuttle, A.H.; Martin, L.J.; Wieskopf, J.S.; Mapplebeck, J.C.; Wei, P.; Zhan, S.; Zhang, S.; et al. The rat grimace scale: A partially automated method for quantifying pain in the laboratory rat via facial expressions. *Mol. Pain* **2011**, *7*, 55. [CrossRef]
7. Keating, S.C.; Thomas, A.A.; Flecknell, P.A.; Leach, M.C. Evaluation of EMLA cream for preventing pain during tattooing of rabbits: Changes in physiological, behavioural and facial expression responses. *PLoS ONE* **2012**, *7*, e44437. [CrossRef]
8. Evangelista, M.C.; Watanabe, R.; Leung, V.S.Y.; Monteiro, B.P.; O'Toole, E.; Pang, D.S.J.; Steagall, P.V. Facial expressions of pain in cats: The development and validation of a Feline Grimace Scale. *Sci. Rep.* **2019**, *9*, 19128. [CrossRef]
9. Evangelista, M.C.; Benito, J.; Monteiro, B.P.; Watanabe, R.; Doodnaught, G.M.; Pang, D.S.J.; Steagall, P.V. Clinical applicability of the Feline Grimace Scale: Real-time versus image scoring and the influence of sedation and surgery. *PeerJ* **2020**, *8*, e8967. [CrossRef]
10. Reijgwart, M.L.; Schoemaker, N.J.; Pascuzzo, R.; Leach, M.C.; Stodel, M.; de Nies, L.; Hendriksen, C.F.M.; van der Meer, M.; Vinke, C.M.; van Zeeland, Y.R.A. The composition and initial evaluation of a grimace scale in ferrets after surgical implantation of a telemetry probe. *PLoS ONE* **2017**, *12*, e0187986. [CrossRef]
11. Hager, C.; Biernot, S.; Buettner, M.; Glage, S.; Keubler, L.M.; Held, N.; Bleich, E.M.; Otto, K.; Muller, C.W.; Decker, S.; et al. The Sheep Grimace Scale as an indicator of post-operative distress and pain in laboratory sheep. *PLoS ONE* **2017**, *12*, e0175839. [CrossRef] [PubMed]

12. Viscardi, A.V.; Hunniford, M.; Lawlis, P.; Leach, M.; Turner, P.V. Development of a piglet grimace scale to evaluate piglet pain using facial expressions following castration and tail docking: A pilot study. *Front. Vet. Sci.* **2017**, *4*, 51. [CrossRef]
13. Gleerup, K.B.; Andersen, P.H.; Munksgaard, L.; Forkman, B. Pain evaluation in dairy cattle. *Appl. Anim. Behav. Sci.* **2015**, *171*, 25–32. [CrossRef]
14. Dalla Costa, E.; Minero, M.; Lebelt, D.; Stucke, D.; Canali, E.; Leach, M.C. Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. *PLoS ONE* **2014**, *9*, e92281. [CrossRef]
15. Hohlbaum, K.; Bert, B.; Dietze, S.; Palme, R.; Fink, H.; Thone-Reineke, C. Severity classification of repeated isoflurane anesthesia in C57BL/6JRj mice—Assessing the degree of distress. *PLoS ONE* **2017**, *12*, e0179588. [CrossRef] [PubMed]
16. Defensor, E.B.; Corley, M.J.; Blanchard, R.J.; Blanchard, D.C. Facial expressions of mice in aggressive and fearful contexts. *Physiol. Behav.* **2012**, *107*, 680–685. [CrossRef]
17. Hohlbaum, K.; Bert, B.; Dietze, S.; Palme, R.; Fink, H.; Thöne-Reineke, C. Impact of repeated anesthesia with ketamine and xylazine on the well-being of C57BL/6JRj mice. *PLoS ONE* **2018**, *13*, e0203559. [CrossRef]
18. Miller, A.L.; Leach, M.C. The mouse grimace scale: A clinically useful tool? *PLoS ONE* **2015**, *10*. [CrossRef]
19. Jirkof, P.; Tourvieille, A.; Cinelli, P.; Arras, M. Buprenorphine for pain relief in mice: Repeated injections vs. sustained-release depot formulation. *Lab. Anim.* **2015**, *49*, 177–187. [CrossRef]
20. Campos-Luna, I.; Miller, A.; Beard, A.; Leach, M. Validation of mouse welfare indicators: A delphi consultation survey. *Sci. Rep.* **2019**, *9*, 10249. [CrossRef]
21. FELASA Working Group on Revision of Guidelines for Health Monitoring of Rodents and Rabbits; Mähler Convenor, M.; Berard, M.; Feinstein, R.; Gallagher, A.; Illgen-Wilcke, B.; Pritchett-Corning, K.; Raspa, M. FELASA recommendations for the health monitoring of mouse, rat, hamster, guinea pig and rabbit colonies in breeding and experimental units. *Lab. Anim.* **2014**, *48*, 178–192.
22. Hurst, J.L.; West, R.S. Taming anxiety in laboratory mice. *Nat. Meth.* **2010**, *7*, 825–826. [CrossRef] [PubMed]
23. Hohlbaum, K.; Bert, B.; Dietze, S.; Palme, R.; Fink, H.; Thöne-Reineke, C. Systematic assessment of well-being in mice for procedures using general anesthesia. *J. Vis. Exp. JoVE* **2018**, e57046. [CrossRef] [PubMed]
24. Mandrekar, J.N. Measures of interrater agreement. *J. Thorac. Oncol.* **2011**, *6*, 6–7. [CrossRef]
25. Koo, T.K.; Li, M.Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef]
26. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
27. Nichols, D. IBM SPSS: Stats Fleiss Kappa Version 4. 2015. Available online: https://www.ibm.com/developerworks/community/files/form/anonymous/api/library/47cbe597-84f5-44c2-a739-acc6f0d3970c/document/48234a16-fb14-4bee-8c18-570319c57108/media/STATS%20FLEISS%20KAPPA.spe (accessed on 6 December 2019).
28. Dalla Costa, E.; Pascuzzo, R.; Leach, M.C.; Dai, F.; Lebelt, D.; Vantini, S.; Minero, M. Can grimace scales estimate the pain status in horses and mice? A statistical approach to identify a classifier. *PLoS ONE* **2018**, *13*, e0200339. [CrossRef]
29. Roughan, J.V.; Bertrand, H.G.; Isles, H.M. Meloxicam prevents COX-2-mediated post-surgical inflammation but not pain following laparotomy in mice. *Eur. J. Pain* **2016**, *20*, 231–240. [CrossRef]
30. Roughan, J.V.; Sevenoaks, T. Welfare and scientific considerations of tattooing and ear tagging for mouse identification. *J. Am. Assoc. Lab. Anim. Sci.* **2019**, *58*, 142–153. [CrossRef]
31. Watanabe, R.; Doodnaught, G.M.; Evangelista, M.C.; Monteiro, B.P.; Ruel, H.L.M.; Steagall, P.V. Inter-rater reliability of the feline grimace scale in cats undergoing dental extractions. *Front. Vet. Sci.* **2020**, *7*, 302. [CrossRef]
32. NC3Rs. Mouse Grimace Scale Poster. Available online: https://www.nc3rs.org.uk/grimacescales (accessed on 6 December 2019).
33. Tsang, V. Eye-tracking study on facial emotion recognition tasks in individuals with high-functioning autism spectrum disorders. *J. Autism.* **2018**, *22*, 161–170. [CrossRef] [PubMed]
34. Zhang, E.Q.; Leung, V.S.; Pang, D.S. Influence of rater training on inter-and intrarater reliability when using the Rat Grimace Scale. *J. Am. Assoc. Lab. Anim. Sci.* **2019**, *58*, 178–183. [CrossRef] [PubMed]

35. Faller, K.M.; McAndrew, D.J.; Schneider, J.E.; Lygate, C.A. Refinement of analgesia following thoracotomy and experimental myocardial infarction using the Mouse Grimace Scale. *Exp. Physiol.* **2015**, *100*, 164–172. [CrossRef]
36. Oliver, V.; De Rantere, D.; Ritchie, R.; Chisholm, J.; Hecker, K.G.; Pang, D.S. Psychometric assessment of the Rat Grimace Scale and development of an analgesic intervention score. *PLoS ONE* **2014**, *9*, e97882. [CrossRef]
37. Philips, B.H.; Weisshaar, C.L.; Winkelstein, B.A. Use of the rat grimace scale to evaluate neuropathic pain in a model of cervical radiculopathy. *Comp. Med.* **2017**, *67*, 34–42.
38. Duffy, S.S.; Perera, C.J.; Makker, P.G.; Lees, J.G.; Carrive, P.; Moalem-Taylor, G. Peripheral and central neuroinflammatory changes and pain behaviors in an animal model of multiple sclerosis. *Front. Immunol.* **2016**, *7*, 369. [CrossRef] [PubMed]
39. Leach, M.C.; Klaus, K.; Miller, A.L.; Scotto di Perrotolo, M.; Sotocinal, S.G.; Flecknell, P.A. The assessment of post-vasectomy pain in mice using behaviour and the Mouse Grimace Scale. *PLoS ONE* **2012**, *7*, e35656. [CrossRef] [PubMed]
40. Andresen, N.; Wöllhaf, M.; Hohlbaum, K.; Lewejohann, L.; Hellwich, O.; Thöne-Reineke, C.; Belik, V.J.B. Towards a fully automated surveillance of well-being status in laboratory mice using deep learning: Starting with facial expression analysis. *PLoS ONE* **2020**, *15*, e0228059. [CrossRef]
41. Ernst, L.; Kopaczka, M.; Schulz, M.; Talbot, S.R.; Struve, B.; Hager, C.; Bleich, A.; Durst, M.; Jirkof, P.; Arras, M.; et al. Semi-automated generation of pictures for the Mouse Grimace Scale: A multi-laboratory analysis (Part 2). *Lab. Anim.* **2020**, *54*, 92–98. [CrossRef]
42. Ernst, L.; Kopaczka, M.; Schulz, M.; Talbot, S.R.; Zieglowski, L.; Meyer, M.; Bruch, S.; Merhof, D.; Tolba, R.H. Improvement of the Mouse Grimace Scale set-up for implementing a semi-automated Mouse Grimace Scale scoring (Part 1). *Lab. Anim.* **2019**. [CrossRef]