Received: 7 July 2020 | Accepted: 20 November 2020

DOI: 10.1111/jvim.15987

Journal of Veterinary Internal Medicine Open Access

ACVIM American College of Veterinary Internal Medicine

STANDARD ARTICLE

# Interobserver and intraobserver reliability for 2 grading systems for gastric ulcer syndrome in horses

Jessica C. Wise[1] | Edwina J.A. Wilkes[1] | Sharanne L. Raidal[1] | Gang Xie[2] | Danielle E. Crosby[1] | Josephine N. Hale[1] | Kristopher J. Hughes[1]

[1]School of Animal and Veterinary Sciences, Charles Sturt University, Wagga Wagga, New South Wales, Australia

[2]Quantitative Consultant Unit, Charles Sturt University, Wagga Wagga, New South Wales, Australia

**Correspondence**
Jessica C. Wise, Veterinary Clinical Centre, Charles Sturt University, 1 Agriculture Avenue, Wagga Wagga, NSW 2650, Australia.
Email: jwise@csu.edu.au

## Abstract

**Background:** Grading of equine gastric ulcer syndrome (EGUS) is undertaken in clinical and research settings, but the reliability of EGUS grading systems is poorly understood.

**Hypothesis/Objectives:** Investigate interobserver and intraobserver reliability of an established ordinal grading system and a novel visual analog scale (VAS), and assess the influence of observer experience.

**Animals:** Sixty deidentified gastroscopy videos.

**Methods:** Six observers (3 specialists and 3 residents) graded videos using the EGUS Council (EGUC) system and VAS. Observers graded the videos three 3 for each system, using a cross-over design with at least 1 week between each phase. The order of videos was randomized for each phase.

**Methods:** Interobserver and intraobserver reliability were estimated using Gwet's agreement coefficient with ordinal weights applied (AC2) for the EGUC system and the intraclass correlation coefficient (ICC) for the VAS.

**Results:** Using the EGUC system, interobserver reliability was substantial for squamous (AC2 = 0.69; 95% confidence interval [CI], 0.57-0.80) and glandular mucosa (AC2 = 0.72; 95% CI, 0.70-0.75), and intraobserver reliability was substantial for squamous (AC2 = 0.80; 95% CI, 0.71-0.90) and glandular mucosa (AC2 = 0.80; 95% CI, 0.74-0.86). Interobserver reliability using the VAS was moderate for squamous (ICC = 0.64; 95% CI, 0.31-0.96) and poor for glandular mucosa (ICC = 0.35; 95% CI, 0.06-0.64), and intraobserver reliability was moderate for squamous (ICC = 0.74; 95% CI, 0.62-0.86) and glandular mucosa (ICC = 0.56; 95% CI, 0.39-0.72).

**Conclusions and Clinical Importance:** The EGUC system had acceptable intraobserver and interobserver reliability and performed well regardless of observer experience. Familiarity and observer experience improved reliability of the VAS.

**Abbreviations:** 95% CI, 95% confidence interval; AC2, Gwet's coefficient of agreement with ordinal weights applied; EGUC, Equine Gastric Ulcer Council; EGUS, equine gastric ulcer syndrome; PS, practitioner's simplified; VAS, visual analog scale.

KEYWORDS
EGUS, grading systems, horse, reliability, VAS

## 1 | INTRODUCTION

Equine gastric ulcer syndrome (EGUS) is the most common disorder of the equine stomach.[1] Grading of EGUS lesions may inform treatment selection, comparison of the efficacy of different treatments, and the impact of husbandry protocols on ulcer healing.[1-6] For grading systems to be useful, good inter- and intraobserver reliability are required[7] to facilitate comparisons of effects of treatments within and between studies and assessment when multiple clinicians are involved in case management.

A simple gastric ulcer lesion grading system based on an ordinal scale (0-4) was described in 1999 by the Equine Gastric Ulcer Council (EGUC).[3] This grading system can be applied to the squamous and glandular mucosa of the equine stomach.[1,3] Other grading systems for EGUS have been described, including a number/severity system,[2] the practitioner's simplified (PS) scoring system,[2,8,9] and ordinal systems based on ulcer depth and surface area.[10,11] The EGUC grading system has higher interobserver reliability compared with the number/severity system,[12] and currently is recommended for assessment of the squamous mucosa.[1] However, there is no consensus on or uniformity in the use of the EGUC grading system, and uncertainty exists for glandular mucosa assessment.[1]

There are limitations in the assessment of disease when severity varies along a continuum, as occurs in EGUS, because ordinal grading systems require strict categorization of severity according to predetermined criteria or definitions. Visual analog scales (VAS) are used in complex clinical contexts to facilitate assessment of subjective characteristics that cannot be directly measured and allow users to integrate multiple variables into a single continuous variable. Previously, VAS have been used for grading of gastrointestinal lesions in humans[13,14] and may provide advantages over an ordinal scale-based approach for assessment of gastric ulceration in horses, including collection of continuous data, which allows for different statistical analysis options.

Our aims were to investigate (a) interobserver and intraobserver reliability of the EGUC grading system and a novel VAS for assessment of squamous and glandular gastric mucosal lesions and (b) the influence of observer experience on the outcomes for both systems. We hypothesized that the use of a VAS would result in superior estimates of reliability compared to the EGUC grading system and that experienced observers would have higher reliability for grading of gastric lesions than would less-experienced observers.

## 2 | MATERIALS AND METHODS

### 2.1 | Horses

Sixty prerecorded, deidentified gastroscopy videos, obtained from horses during unrelated research projects, were used. For inclusion, visualization of the greater curvature, margo plicatus, lesser curvature, glandular mucosa, and pyloric antrum was required. Videos were selected by a single technician who did not participate in the study, and attempted to include an even distribution of lesion severity, based on gastric mucosal appearance.

### 2.2 | Grading systems

Two grading systems were used: the EGUC system[1] (Table 1) and a novel VAS (Figure 1). The VAS was a 10 cm line anchored at both ends with words descriptive of the maximal and minimal extremes of the dimension being measured.[15] The VAS is used as a 100-point continuous scale. Separate scores were recorded for the squamous and glandular mucosa for both systems.

### 2.3 | Observers

Six observers were included: 3 specialists in equine medicine and 3 residents in equine disciplines (medicine, surgery and sports medicine). The observers graded the videos 3 times for each system. The grading systems were used alternatively in a cross-over design with at least 1 week between each of the 6 phases of the study. For each phase, the order of videos was randomized to avoid pattern recognition that might contribute to measurement bias and influence study validity.

### 2.4 | Statistical analysis

All statistical analyses were performed using R[a] Statistical Software (R version 3.6.0 [2019]). For the EGUC system, intra- and

**TABLE 1** The Equine Gastric Ulcer Council (EGUC) 5-point ordinal grading system for grading squamous and glandular gastric disease

| Grade | Squamous mucosa | Glandular mucosa |
|---|---|---|
| 0 | The epithelium is intact and there is no appearance of hyperkeratosis | The epithelium is intact and there is no appearance of hyperemia |
| 1 | The mucosa is intact, but there are areas of hyperkeratosis | The epithelium is intact, but there are areas of hyperemia |
| 2 | Small, single or multifocal lesions | Small, single, or multifocal lesions |
| 3 | Large single or extensive superficial lesions | Large single or extensive superficial lesions |
| 4 | Extensive lesions with areas of apparent deep ulceration | Extensive lesions with area of apparent deep ulceration |

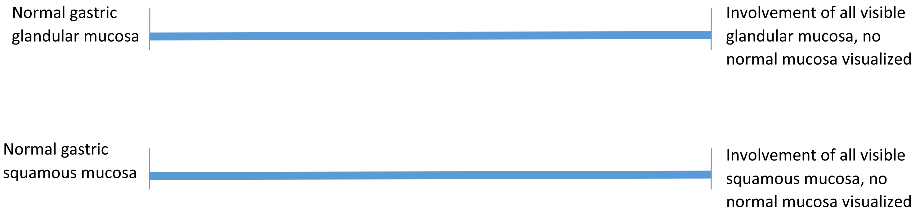
Normal gastric glandular mucosa
Involvement of all visible glandular mucosa, no normal mucosa visualized
Normal gastric squamous mucosa
Involvement of all visible squamous mucosa, no normal mucosa visualized

**FIGURE 1** The visual analog scoring system for grading the appearance of squamous and glandular gastric mucosa

interobserver reliability were assessed using Gewt's coefficient of agreement with ordinal weighting applied (AC2). Interpretation of AC2 was derived from a previously proposed system[16]: ≤0.20: poor, 0.21 to 0.40: fair, 0.41 to 0.60: moderate, 0.61 to 0.80: substantial, and 0.81 to 1.0: excellent reliability.

For the VAS, observer reliability was estimated by calculation of the intraclass correlation coefficient (ICC) based on a mean rating (k = 6), absolute agreement, 2-way mixed effects model, and 95% confidence interval (CI). The benchmarking of ICC values was adapted from previous studies[17,18]: <0.50: poor, 0.50 to 0.75: moderate, 0.76 to 0.90: good, and >0.9: excellent reliability.

For both the squamous and glandular mucosa, interobserver reliability coefficients were calculated for each of the 3 phases for each grading system, and the mean and 95% CI were calculated. Interobserver reliability was calculated for the 3 experienced observers (observers 1-3) and the 3 less-experienced observers (observers 4-6), and the mean and 95% CI were calculated for observer groups across the 3 phases.

The intraobserver reliability coefficients for the 6 observers were calculated from ratings obtained over the 3 phases of the study for each grading system. The mean and 95% CI for the reliability coefficients were calculated for all observers, experienced observers, and less-experienced observers.

# 3 | RESULTS

## 3.1 | EGUS Council grading system

### 3.1.1 | Interobserver reliability

Results of the analyses of interobserver reliability of the EGUC system for squamous and glandular gastric mucosa are provided in Figure 2 and Supplementary Item 1. Substantial interobserver reliability was found for grading of squamous (mean AC2, 0.69; 95% CI, 0.57-0.80) and glandular mucosa (mean AC2, 0.72; 95% CI, 0.70-0.75). Minimal difference was found in interobserver reliability of squamous or glandular mucosa over the 3 phases (Figure 2). Overall, experience had limited influence on interobserver reliability for grading of squamous or glandular mucosa (Figure 3, Supplementary Item 2), but experienced observers had higher reliability (AC2, 0.77; 95% CI, 0.70-0.83) than did less-experienced observers (AC2, 0.62; 95% CI, 0.52-0.72) for glandular mucosal grading in Phase 1. Experienced observers demonstrated improvement in interobserver reliability when grading squamous mucosa between Phase 1 (AC2, 0.53; 95% CI, 0.41-0.64) and Phase 3 (AC2, 0.74; 95% CI, 0.66-0.81).

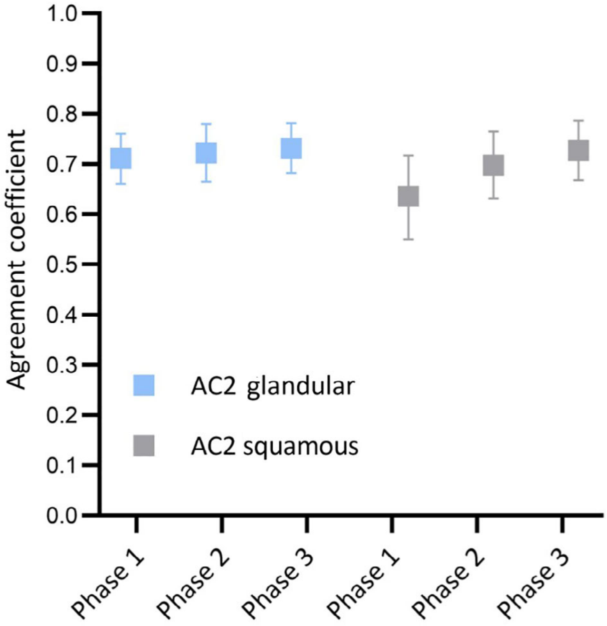

**FIGURE 2** Results of Gwet's coefficient of agreement with ordinal weighting applied (AC2) for interobserver reliability of observers grading squamous and glandular gastric mucosa using the Equine Gastric Ulcer Council (EGUC) system on 3 occasions. The figure is presented as mean and 95% CI

### 3.1.2 | Intraobserver reliability

The estimates of intraobserver reliability of the EGUC system for squamous and glandular mucosa are provided in Table 2. Overall, substantial intraobserver reliability was found for assessment of both the squamous (AC2, 0.80; 95% CI, 0.71-0.90) and glandular mucosa (AC2, 0.80; 95% CI, 0.74-0.86) using the EGUC grading system. Experienced observers had excellent and substantial intraobserver reliability for EGUC system grading of squamous (mean AC2, 0.83; 95% CI, 0.75-0.92) and glandular gastric mucosa (mean AC2, 0.79; 95% CI, 0.73-0.85), respectively. Less-experienced observers demonstrated substantial and excellent intraobserver reliability when grading squamous (mean AC2, 0.77; 95% CI, 0.45-1.0) and glandular mucosa (mean AC2, 0.82; 95% CI, 0.62-1.0), respectively. For individual observers, experience had minimal influence on intraobserver reliability: pairwise comparisons did not identify differences, with the exception of a lower AC2 result for Observer 5 compared to Observers 1 and 3 for squamous mucosal grading (Table 2).

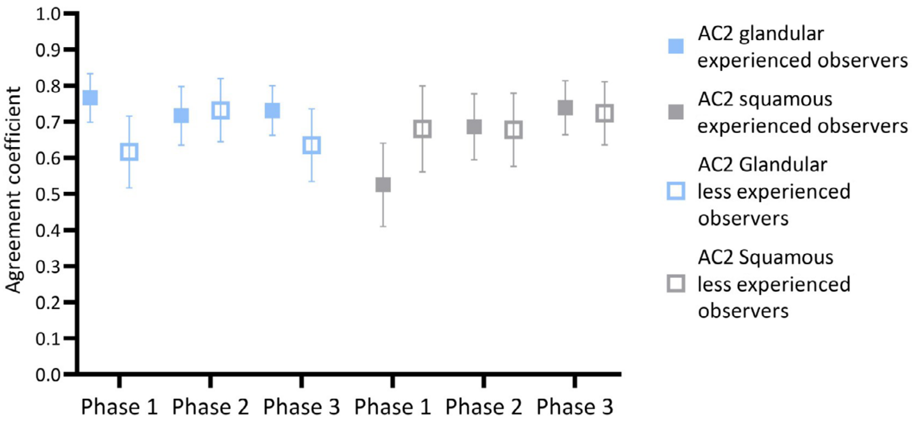

**FIGURE 3** Results of Gwet's coefficient of agreement with ordinal weighting (AC2) comparing the interobserver reliability of experienced observers (specialists in equine medicine) and less-experienced observers (residents in equine disciplines) grading squamous and glandular gastric mucosa using the Equine Gastric Ulcer Council (EGUC) system on 3 occasions. The figure is presented as mean and 95% CI

**TABLE 2** Results of Gwet's coefficient of agreement with ordinal weighting (AC2) for the intraobserver reliability of scoring of glandular and squamous gastric mucosa with the Equine Gastric Ulcer Council (EGUC) grading system. The mean AC2 has been calculated for the intraobserver reliability of experienced observers (specialists in equine medicine) and less experienced observers (residents in equine disciplines)

| | Glandular mucosa | | | Squamous mucosa | | |
|---|---|---|---|---|---|---|
| | | 95% CI | | | 95% CI | |
| | AC2 | Lower limit | Upper limit | AC2 | Lower limit | Upper limit |
| Experienced | | | | | | |
| Observer 1 | 0.76 | 0.66 | 0.87 | 0.83 | 0.74 | 0.92 |
| Observer 2 | 0.81 | 0.76 | 0.87 | 0.80 | 0.73 | 0.87 |
| Observer 3 | 0.79 | 0.70 | 0.87 | 0.87 | 0.83 | 0.91 |
| Mean (n = 3) | 0.79 | 0.73 | 0.85 | 0.83 | 0.75 | 0.92 |
| Less experienced observers | | | | | | |
| Observer 4 | 0.89 | 0.85 | 0.93 | 0.83 | 0.78 | 0.88 |
| Observer 5 | 0.73 | 0.64 | 0.82 | 0.62 | 0.50 | 0.74 |
| Observer 6 | 0.84 | 0.75 | 0.93 | 0.85 | 0.80 | 0.91 |
| Mean (n = 3) | 0.82 | 0.62 | 1.0 | 0.77 | 0.45 | 1.0 |
| Overall mean (n = 6) | 0.80 | 0.74 | 0.86 | 0.80 | 0.71 | 0.90 |

## 3.2 | Visual analog scale

### 3.2.1 | Interobserver reliability

The estimates of the interobserver reliability of the VAS for grading squamous and glandular gastric mucosa are provided in Figure 4 and Supplementary Item 3. Overall, the interobserver reliability of the VAS was moderate for squamous mucosal grading (mean ICC, 0.64; 95% CI, 0.31-0.96) and poor for glandular mucosal grading (mean ICC, 0.35; 95% CI, 0.06-0.64). Interobserver reliability was higher for grading of the squamous mucosa than for glandular mucosa in Phase 2 (squamous ICC, 0.64; 95% CI, 0.53-0.73; glandular ICC, 0.26; 95% CI, 0.15-0.40) and Phase 3 (squamous ICC, 0.77; 95% CI, 0.69-0.84; glandular ICC, 0.32; 95% CI, 0.20-0.47), largely because of increasing reliability of squamous mucosal grading over time (Figure 4; Supplementary Item 3).

Overall, experience had an effect on the interobserver reliability of the VAS. For both squamous and glandular mucosal grading, reliability coefficients for experienced observers were higher than those of less-experienced observers for all phases (Figure 5; Supplementary Item 4), most notably for grading of the glandular mucosa in Phase 2 (Figure 5). Both experienced and less-experienced observers demonstrated improvement in reliability of grading squamous mucosa using the VAS from Phase 1 to Phase 3, whereas, overall, interobserver reliability for grading of the glandular mucosa was poor and did not improve, regardless of experience (Figure 5).

### 3.2.2 | Intraobserver reliability

The estimates of intraobserver reliability of the VAS for grading of squamous and glandular gastric mucosa are provided in Table 3. Overall, intraobserver reliability using the VAS system was good and moderate for grading of the squamous (mean ICC, 0.74; 95% CI, 0.62-0.86) and glandular mucosa (mean ICC, 0.56; 95% CI, 0.39-0.72), respectively. By group, experienced observers had good and moderate intraobserver reliability for VAS grading of squamous (mean ICC, 0.83; 95% CI, 0.55-1.0) and glandular mucosa (mean ICC, 0.65; 95% CI, 0.50-0.80), respectively, whereas less-experienced observers had moderate reliability when grading squamous mucosa (mean ICC, 0.67; 95% CI, 0.46-0.86) and poor reliability for glandular mucosal grading (mean AC2, 0.46; 95% CI, 0.00-0.92). Individual pair-wise comparisons indicated some differences with experience.

For squamous mucosal grading, observers 4 and 6 had lower reliability than did observers 1-3, whereas for glandular grading, observer 5 had lower reliability than did observers 1 and 3, and observer 4 had lower reliability than did observer 3.

## 4 | DISCUSSION

We comprehensively investigated interobserver and intraobserver reliability of the EGUC system and introduced a novel VAS for scoring the endoscopic appearance of the equine stomach. Overall, the EGUC system had substantial interobserver and intraobserver reliability for grading of both squamous and glandular mucosa, and reliability was minimally influenced by experience. The reliability of the VAS was more variable, with poor reliability for grading glandular mucosa, and was influenced by observer experience and familiarity with the system.

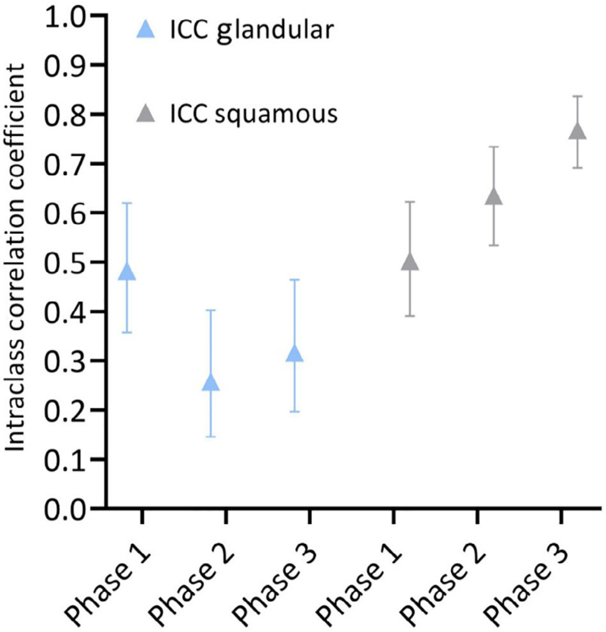

**FIGURE 4** Results of the intraclass correlation coefficient (ICC), 1 way model, for the interobserver reliability of observers grading squamous and glandular gastric mucosa using the visual analog scale (VAS) on 3 occasions. The figure is presented as mean and 95% CI

In our study, the EGUC system demonstrated substantial interobserver reliability and substantial to excellent intraobserver reliability. These results are consistent with the findings of an earlier study in which good interobserver agreement of the EGUC system was reported.[12] Similarly, ordinal grading systems are used for the assessment of lameness, heart murmurs and ataxia in horses, and moderate to substantial interobserver and intraobserver reliability and agreement for these systems have been reported.[19-22] Given widespread application of ordinal grading systems in veterinary clinical practice and research, determination of intra- and interobserver agreement and reliability of each system is important. Although agreement reflects the extent to which scores, ratings or diagnoses are identical, reliability is the ratio of variability between scores or ratings of the same patients to the total variability of all scores in the sample and represents the ability of a measurement to differentiate between patients.[23] Both agreement and reliability are important for the development of rating scales and conduct of clinical studies, and provide information on the error inherent in measurement, rating, or diagnosis.[23] Although agreement is desirable for binary decisions, such as whether to institute treatment or not, our results indicate good ability of observers to distinguish between ulcer severity when using the EGUC system, which remains important in clinical and research settings because it indicates that this system can be used for comparison among studies and assessment of animal responses to treatment and management changes.

There was minimal influence of experience on the interobserver or intraobserver reliability of the EGUC system. Within the group of less-experienced observers, 2 of the 3 observers had no previous experience in using the EGUC grading system. The experienced observers all were specialists in equine internal medicine, with extensive clinical and research experience using the EGUC grading system. Our findings emphasize that interobserver and intraobserver reliability of the EGUC system is not affected when used by observers unfamiliar with the grading system, or by observers experienced using the system. Furthermore, interobserver reliability was not different when grading squamous or glandular mucosa. Intraobserver reliability was

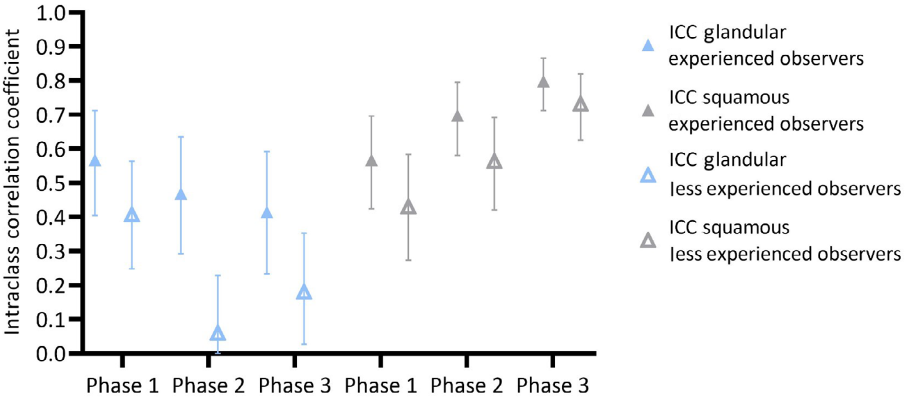

**FIGURE 5** Results of the intraclass correlation coefficient (ICC), 1 way model, comparing the interobserver reliability of experienced observers (specialists in equine medicine) and less-experienced observers (residents in equine disciplines) grading squamous and glandular gastric mucosa using the visual analog scale (VAS) on 3 occasions. The figure is presented as mean and 95% CI

**TABLE 3** Results of the intraclass correlation coefficient (ICC), one way model, for the intraobserver reliability of grading squamous and glandular mucosa using the novel visual analog scale. The mean ICC has been calculated for experienced observers (specialists in equine medicine) and less-experienced observers (residents in equine disciplines)

| | Glandular mucosa | | | Squamous mucosa | | |
|---|---|---|---|---|---|---|
| | | 95% CI | | | 95% CI | |
| | ICC | Lower limit | Upper limit | ICC | Lower limit | Upper limit |
| Experienced | | | | | | |
| Observer 1 | 0.68 | 0.53 | 0.79 | 0.75 | 0.64 | 0.83 |
| Observer 2 | 0.58 | 0.41 | 0.72 | 0.85 | 0.78 | 0.90 |
| Observer 3 | 0.69 | 0.56 | 0.80 | 0.89 | 0.83 | 0.93 |
| Mean (n = 3) | 0.65 | 0.50 | 0.80 | 0.83 | 0.65 | 1.0 |
| Less experienced observers | | | | | | |
| Observer 4 | 0.41 | 0.25 | 0.56 | 0.59 | 0.45 | 0.71 |
| Observer 5 | 0.31 | 0.15 | 0.48 | 0.75 | 0.65 | 0.83 |
| Observer 6 | 0.67 | 0.54 | 0.77 | 0.64 | 0.52 | 0.75 |
| Mean (n = 3) | 0.46 | 0.0 | 0.92 | 0.67 | 0.46 | 0.86 |
| Overall mean (n = 6) | 0.56 | 0.39 | 0.72 | 0.74 | 0.62 | 0.86 |

slightly better than interobserver reliability, possibly reflecting different interpretation of the EGUC system scale among individuals, but good ability of individual observers to repeatedly apply the grading system scale in the same way. Intraobserver reliability has been reported to be higher than interobserver reliability for other ordinal grading systems,[19,24,25] which may reflect consistency in the interpretation or application of the grading system within observers, but differences in interpretation of the grading system among observers.[26] Differences in interpretation of a grading system have been speculated to be affected by clinical experience and opinions of the disorder being assessed.[25] However, the impact of experience on interobserver or intraobserver reliability of the EGUC grading system was minimal in our study.

In our study, reliability of the EGUC system was estimated using Gwet's weighted agreement coefficient (AC2). In previous studies, Gwet's AC statistics have been found to provide good estimates of intra- and interobserver reliability for categorical scoring systems in human medicine.[27-29] Gwet's AC1 is a first-order agreement coefficient that is an alternative to the kappa coefficient and adjusts the overall probability of agreement for chance agreement.[30] Although the AC1 statistic can be used for any number of raters, this coefficient is used primarily for nominal data. The second-order agreement coefficient, Gwet's AC2 statistic, is a weighted version of AC1 that adjusts for chance agreement and accounts for misclassification errors and nonabsolute agreement, and is recommended for analyzing ordinal, interval, and ratio data.[30] Other estimates of reliability, including Cohen's kappa and weighted Cohen's kappa statistics, have been used to estimate the interobserver and intraobserver reliability of ordinal grading systems. The advantage of Gwet's AC2 statistic over other estimates of agreement, including Cohen's kappa, is that it is paradox-resistant and expected to provide a more accurate estimate of observer reliability because other estimates of agreement often are influenced by the number of categories available and the proportion of subjects in each category, creating a paradox whereby a low agreement coefficient is calculated despite good reliability.[27,30] To our knowledge, ours is the first study to use Gwet's AC2 statistic to assess the reliability of the EGUC grading system. The results indicated that reliability was substantial to excellent, within and between observers, for rating of both squamous and glandular mucosal lesions using the EGUC system (ie, observers graded lesions similarly but not identically). Although observers may grade lesions similarly, differences in the interpretation of the EGUC system remain possible, which is important when applying this system to measure treatment efficacy, as has been done previously.[4,5,31-33] In some studies, a differences of 1 grade was considered a treatment effect or improvement,[4,5,33] but our findings indicate that the intraobserver reliability of the EGUC system is not perfect, requiring consideration when assessing responses to treatment. Similarly, consideration of the interobserver reliability of the EGUC system is necessary when several clinicians are involved in the assessment of treatment responses in an individual animal, because our results suggest that observers grade mucosal lesions in a similar but not identical way.

The severity of gastric lesions in our study varied along a continuum, presenting challenges for categorization using the defined grading criteria. Although the use of ordinal scales to assess changes is simpler for extreme categories, observer agreement can be more challenging for borderline categories[16] or for assessment of mild and moderate disease,[14] leading to higher variability and misclassification errors. Our study introduces the use of a VAS to grade EGUS, in an attempt to provide an alternative to the ordinal grading system. In our study, the inter- and intraobserver reliability of the VAS improved with time and there was some influence of observer experience. Because none of the observers in our study had previous experience in using the VAS, the differences in reliability between the 2 groups is likely more reflective of knowledge and clinical experience than of familiarity with the grading system. The reliability results for the VAS for experienced observers in our study were similar to those reported for clinical assessments in human dentistry (0.69-0.92)[34] and human medicine (0.77-0.91).[24] In our study, interobserver reliability of the VAS for grading of the squamous mucosa improved over time

(Figures 4 and 5), which may reflect conditioning of observers to the VAS and increasing familiarity with the system. Previously, consistency among observers using a VAS has been improved by consensus meetings[24] and by the use of guide points (anchors) adjacent to the scale.[13] Further investigation into the value of training clinicians in the use of VAS, and whether reliability of this system is improved with repeated utilization of the system by observers, is warranted. Our results suggest that inter- and intraobserver reliability of grading squamous gastric mucosa with the VAS could be improved using these techniques. For the VAS, inter- and intraobserver reliability were better for grading squamous mucosa that glandular mucosa. These results likely reflect an observer's ability to grade severity of squamous gastric mucosal lesions and difficulties in interpreting glandular lesions and application of hierarchical grading systems.[1] Although we tried to include a broad spectrum of squamous and glandular disease in the study, most of the included gastroscopy videos featured mild to moderate glandular disease, and very few chronic severe gastric lesions were available. The poor inter- and intraobserver reliability found when using the VAS to grade glandular gastric lesions may be compounded by the included observers' inexperience with using the VAS grading system, as well as the complexity of interpreting mild to moderate glandular gastric lesions. The poor reliability of the VAS also may be explained by increased categories (continuous scale) when compared with the EGUC system. Similarly, in a previous study, the N/S system, which contains a higher number of categories, had a poorer reliability when compared with the EGUC.[12] The poor reliability of both the VAS and N/S systems in comparison with the EGUC may reflect increased ease of reliability with fewer categories.

Visual analog scales have been established as valid and reliable in a range of clinical and research applications.[15] In our study, the novel VAS used was designed as previously recommended[15] using a 10 cm line with words descriptive of the maximal and minimal extremes of the dimension being measured. The 10 cm line was used as a 100-point continuous scale and data were used for estimation of observer reliability. Intermediate points were not used in the VAS to avoid false clustering of scores around an intermediate point or numbers.[15] The use of a VAS results in collection of continuous data, permitting a wider range of statistical analysis options and the potential for higher power and sensitivity of outcome rankings.[34] The location and dispersion of scores might give information on the extent to which the observer takes advantage of the length of the scale,[13] but this was not evaluated in our study.

The use of benchmarking reliability coefficients allows for practical application and interpretation of results. However, the margin of error associated with the reliability coefficient also should be included in interpretation of the results.[17,30] The estimates of reliability calculated for the EGUC and VAS, by the AC2 coefficient and ICC, respectively, cannot be directly compared. As such, 2 different benchmarking systems were used to reflect the 2 different statistical methods used to estimate reliability in our study. Application of only the reliability coefficient to determine the benchmark often leads to an overly optimistic characterization of the extent of reliability.[30] In our study, all measures of reliability, for both grading systems, had wide CI. The width of CI reflects the variability or precision of the calculated estimate,[35] and precision is associated with the degree of random error, which is minimized by increasing sample size.[36] In our study, the small number of observers increased random error and resulted in more imprecise estimates of reliability, reflected by the wide 95% CI.[35,36] When comparing the reliability coefficients, either between observers in intraobserver assessments, or over phases in interobserver assessments, the 95% CI overlapped, which reflects uncertainty as to whether a true difference existed (Figure 3). Conversely, a lack of overlap in 95% CI increases the likelihood of a true difference in results, such as the improvement in the interobserver reliability of grading squamous mucosa with the VAS shown in Figure 4.

A limitation of our study was the use of prerecorded videos, rather than assessment of gastric ulceration at the time of gastroscopy. This approach was necessary to ensure appropriate stratification of gastric ulcer lesions and permit repeated evaluation of unchanged lesions. In a study of human patients, good agreement between video-recorded and live colonoscopy examinations was found, although live assessment was perceived as easier.[37] In a previous study, interobserver agreement in the evaluation of lameness was higher for examination of live horses, compared to video recordings.[20] To our knowledge, comparison of live and recorded gastroscopic examinations in horses has not been performed. Glandular lesions are considered more difficult to grade than squamous lesions, and it has been suggested that the number, location and type of lesions be recorded.[1] In our study, fewer severe glandular lesions were available for inclusion than was the case for squamous lesions, and this difference may have influenced the scoring of glandular lesions by either system. Another limitation of our study is that all included observers worked in the same referral hospital, which could have influenced the estimates of interobserver reliability.

## 5 | CONCLUSION

The EGUC system for grading EGUS lesions has acceptable intraobserver and interobserver reliability and performs well regardless of clinician experience. A VAS may offer advantages in ease of use for rating of squamous mucosa, but observers should be practiced in the use of this system. Glandular lesions in horses may be more difficult to grade than squamous lesions.

### ACKNOWLEDGMENTS

No funding was received for this study. The data that support the findings of this study are openly available in "figshare" at https://figshare.com/articles/The_inter-and_intra_observer_agreement_and_repeatability_of_two_grading_systems_of_Euqine_Gastric_Ulcer_Syndrome/11944761/1. Acknowledgments to E. McCormack for her contribution in compiling the videos.

### CONFLICT OF INTEREST DECLARATION

Kristopher J. Hughes serves as Associate Editor for the Journal of Veterinary Internal Medicine. He was not involved in review of this manuscript.

## OFF-LABEL ANTIMICROBIAL DECLARATION

Authors declare no off-label use of antimicrobials.

## INSTITUTIONAL ANIMAL CARE AND USE COMMITTEE (IACUC) OR OTHER APPROVAL DECLARATION

As this study included videos from previous studies with appropriate ethical animal research approval, this study required no ethical animal or human research approval.

## HUMAN ETHICS APPROVAL DECLARATION

Authors declare human ethics approval was not needed for this study.

## ORCID

*Jessica C. Wise* https://orcid.org/0000-0003-3483-9384
*Edwina J.A. Wilkes* https://orcid.org/0000-0001-8410-5894
*Sharanne L. Raidal* https://orcid.org/0000-0001-5558-3133
*Kristopher J. Hughes* https://orcid.org/0000-0002-8405-6268

## REFERENCES

1. Sykes BW, Hewetson M, Hepburn RJ, Luthersson N, Tamzali Y. European College of Equine Internal Medicine Consensus Statement—equine gastric ulcer syndrome in adult horses. *J Vet Intern Med*. 2015;29:1288-1299.
2. Macallister CG, Andrews FM, Deegan E, et al. A scoring system for gastric ulcers in the horse. *Equine Vet J*. 1997;29:430-433.
3. Equine Gastric Ulcer Council. Recommendations for the diagnosis and treatment of equine gastric ulcer syndrome (EGUS): The Equine Gastric Ulcer Council. *Equine Vet Educat*. 1999;11:262-272.
4. Sykes BW, Sykes KM, Hallowell GD. A comparison of two doses of omeprazole in the treatment of equine gastric ulcer syndrome: a blinded, randomised, clinical trial. *Equine Vet J*. 2014;46:416-421.
5. Sykes BW, Sykes KM, Hallowell GD. A comparison of three doses of omeprazole in the treatment of equine gastric ulcer syndrome: a blinded, randomised, dose-response clinical trial. *Equine Vet J*. 2015;47:285-290.
6. MacAllister CG, Sifferman RL, McClure SR, et al. Effects of omeprazole paste on healing of spontaneous gastric ulcers in horses and foals: a field trial. *Equine Vet J*. 1999;31:77-80.
7. Fuller CJ, Bladon BM, Driver AJ. The intra- and interassessor reliability of measurement of functional outcome by lameness scoring in horses. *Vet J*. 2006;171:281-286.
8. Andrews FM, Nadeau JA. Clinical syndromes of gastric ulceration in foals and mature horses. *Equine Vet J*. 1999;29:30.
9. Andrews FM, Sifferman RL, Bernard W, et al. Efficacy of omeprazole paste in the treatment and prevention of gastric ulcers in horses. *Equine Vet J*. 1999;31:81-86.
10. Begg LM, O'Sullivan CB. The prevalence and distribution of gastric ulceration in 345 racehorses. *Aust Vet J*. 2003;81:199-201.
11. Dionne RM, Vrins A, Doucet MY, Pare J. Gastric ulcers in standardbred racehorses: prevalence, lesion description, and risk factors. *J Vet Intern Med*. 2003;17:218-222.
12. Bell RJ, Kingston JK, Mogg TD. A comparison of two scoring systems for endoscopic grading of gastric ulceration in horses. *N Z Vet J*. 2007;55:19-22.
13. Aabakken L, Larsen S, Osnes M. Visual analogue scales for endoscopic evaluation of nonsteroidal anti-inflammatory drug-induced mucosal damage in the stomach and duodenum. *Scand J Gastroenterol*. 1990;25:443-448.
14. de Lange T, Larsen S, Aabakken L. Inter-observer agreement in the assessment of endoscopic findings in ulcerative colitis. *BMC Gastroenterol*. 2004;4:9-9.
15. McCormack HM, Horne DJ, Sheather S. Clinical applications of visual analogue scales: a critical review. *Psychol Med*. 1988;18:1007-1019.
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
17. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155-163.
18. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6:284-290.
19. Fuller CJ, Bladon BM, Driver AJ, Barr ARS. The intra- and interassessor reliability of measurement of functional outcome by lameness scoring in horses. *Vet J*. 2006;171:281-286.
20. Olsen E, Dunkel B, Barker WHJ, et al. Rater agreement on gait assessment during neurologic examination of horses. *J Vet Intern Med*. 2014;28:630-638.
21. Menzies-Gow NJ, Stevens KB, Sepulveda MF, Jarvis N, Marr CM. Repeatability and reproducibility of the obel grading system for equine laminitis. *Vet Rec*. 2010;167:52-55.
22. Menzies-Gow NJ, Knowles EJ, Rogers I, Rendle DI. Validity and application of immunoturbidimetric and enzyme-linked immunosorbent assays for the measurement of adiponectin concentration in ponies. *Equine Vet J*. 2019;51:33-37.
23. Kottner J, Audige L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidem*. 2011;64:96-106.
24. Zannoni L, Savelli L, Jokubkiene L, et al. Intra- and interobserver reproducibility of assessment of Doppler ultrasound findings in adnexal masses. *Ultrasound Obstet Gynecol*. 2013;42:93-101.
25. McGivney CL, Sweeney J, David F, et al. Intra- and interobserver reliability estimates for identification and grading of upper respiratory tract abnormalities recorded in horses at rest and during overground endoscopy. *Equine Vet J*. 2017;49:433-437.
26. McGivney CL, Sweeney J, Gough KF, et al. Serial evaluation of resting and exercising overground endoscopic examination results in young Thoroughbreds with no treatment intervention. *Equine Vet J*. 2019;51:192-197.
27. Wongpakaran N, Wongpakaran T, Wedding D. A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol*. 2013;13:61.
28. Tammaa A, Fritzer N, Lozano P, et al. Interobserver agreement and accuracy of non-invasive diagnosis of endometriosis by transvaginal sonography. *Ultrasound Obstet Gynecol*. 2015;46:737-740.
29. Bignotti B, Calabrese M, Signori A, et al. Background parenchymal enhancement assessment: inter- and intra-rater reliability across breast MRI sequences. *Europ J Radiol*. 2019;114:57-61.
30. Gwet KL. *Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters*. US: Advanced Analytics, LLC; 2014.
31. Birkmann K, Junge HK, Maischberger E, Wehrli Eser M, Schwarzwald CC. Efficacy of omeprazole powder paste or enteric-coated formulation in healing of gastric ulcers in horses. *J Vet Intern Med*. 2014;28:925-933.
32. Sykes BW, Sykes K, Hallowell GD. Comparison of the effect of two doses of omeprazole on the squamous gastric mucosa in thoroughbred racehorses. *Vet Rec*. 2014;175:249.
33. Sykes BW, Sykes KM, Hallowell GD. A comparison between pre- and post exercise administration of omeprazole in the treatment of equine gastric ulcer syndrome: a blinded, randomised, clinical trial. *Equine Vet J*. 2014;46:422-426.

34. Fowler P, Bellardie H, Shaw B, et al. Reliability of a Categorical Scale (GOSLON) and a continuous scale (10-cm visual analog scale) for assessing dental arch relationships using conventional plaster and 3D digital orthodontic study models of children with complete unilateral cleft lip and palate. *Cleft Palate Craniofacial J*. 2019;56:84-89.
35. Sim J, Reid N. Statistical inference by confidence intervals: issues of interpretation and utilization. *Phys Ther*. 1999;79:186-195.
36. Akobeng AK. *Confidence Intervals and p-Values in Clinical Decision Making*. Oxford, UK: Blackwell Publishing; 2008:1004-1007.
37. Scaffidi MA, Grover SC, Carnahan H, et al. A prospective comparison of live and video-based assessments of colonoscopy performance. *Gastrointest Endosc*. 2018;87:766-775.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Wise JC, Wilkes EJA, Raidal SL, et al. Interobserver and intraobserver reliability for 2 grading systems for gastric ulcer syndrome in horses. *J Vet Intern Med*. 2021;35:571–579. https://doi.org/10.1111/jvim.15987