Training Improves the Reliability of Temperament Assessment in Cattle

Jamie T. Parham
Jessica J. Schmidt
Ronald M. Lewis

Summary with Implications

Accurate and precise measurement of docility in cattle is paramount when including temperament as a criterion for selection. The value of training individuals in assigning a docility score was evaluated by comparing the reliability of individual assessments of temperament in beef cattle before and after various instructional methods. Preceding training, participants' assessment of cattle behavior, videoed while each heifer was restrained in a chute, was not impacted by age, gender, or pre-existing cattle handling experience. Groups of participants that received additional training were more accurate and precise in evaluating temperament, regardless of training method, compared to those without. No matter an individual's prior beef cattle experience, they benefitted from the information provided in the training material. By completing a relatively short and targeted instructional program, producers can more reliably evaluate docility in their cattle, thereby enhancing their ability to incorporate temperament into their selection decisions within their herd.

Introduction

Strong behavioral responses of cattle towards humans or any other stressor have been associated with increased risk of handler injury. Additionally, such cattle have poorer weight gain and meat-eating quality, decreased tolerance to disease, and decreased reproductive performance, with increased production costs. Because of these effects, it is not uncommon for ranchers to make selection decisions based on an animal's behavior. Therefore, accurate and precise evaluation of docility in livestock is important for improvements in animal well-being, human safety, and profitability.

An animal's temperament is often subjectively evaluated as it is relatively straightforward to accomplish while working cattle. Research using such methods, however, report inconsistent classifications among evaluators, which affects the usefulness of subjective assessments. Consistency can be quantified by both the accuracy—the closeness of a measured value to a standard or known value—and precision—the closeness of two or more measurements to each other—of a set of measurements. Accuracy and precision are formally evaluated using inter- and intraobserver reliability, respectively.

Previous research has shown that chute scores are effective methods of measuring temperament and are consistently assessed by trained individuals (*2018 Nebraska Beef Cattle Report*, pp. 75–80). To assist the beef industry in benefitting from subjective evaluation of temperament, the objective of this study was to determine the impact of various training methods on improving reliability of behavior assessment in cattle restrained in a chute.

Table 1. Participant demographics by experience, age, and gender

		Group			
Category[1]	Level	C[4]	T1[5]	T2[6]	Total[7]
Experience[2]	Experienced	13	13	13	39
	Inexperienced	18	17	16	51
Age[3]	College	18	18	17	53
	Other	13	12	12	37
Gender	Male	16	17	16	49
	Female	15	13	13	41

[1] Categories determined using participants' responses to a questionnaire completed before the start of session 1.

[2] Experienced included "Expert (I work with cattle every day)" and "Competent (I work with cattle on a regular basis)" while Inexperienced included "Inexperienced (I work with cattle from time to time)" and "No experience".

[3] Age was grouped into "college" (19 to 22) and "other" (23 and up).

[4] Participants received no training and were not provided with a self-test.

[5] Participants viewed a training video prior to session 2.

[6] Participants viewed a training video and completed a self-test prior to session 2.

[7] Only participants who completed both sessions were included.

Procedure

Ninety individuals of varying age, gender, and cattle backgrounds were recruited to participate in the study, which was conducted on the East Campus of the University of Nebraska – Lincoln. Participants arrived to the first session (S1) and completed an animal experience questionnaire designed to collect information about previous animal handling experience and general demographics. Upon completion of the questionnaire, participants were shown 28 video clips (15 sec each) of cattle restrained in a chute and were asked to score each animal's temperament on a scale of 1 (docile) to 6 (aggressive). Unbeknownst to the participants, the video clips were a repetition of 14 videos shown twice. Data were collected using Qualtrics Survey Software.

The prerecorded video clips used were obtained from an earlier study of animal behavior conducted at the Virginia Tech Kentland farm, Virginia, U.S.A. As part of their assessment, heifers were previously given a subjective chute score by three trained individuals.

Participants were assigned in a balanced way to one of three treatments based on

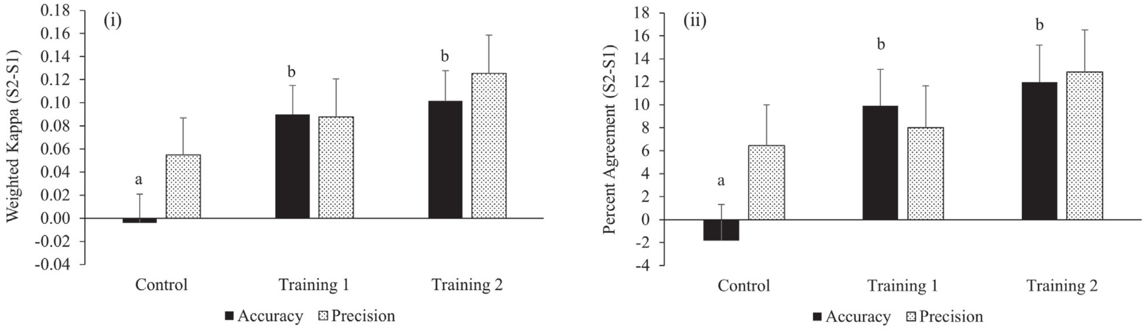

Figure 1. Comparison of accuracy (interobserver reliability) and precision (intraobserver reliability) from first (S1) to second (S2) session, shown as the difference in weighted Kappa coefficient (i) and the difference in percent agreement (ii) between sessions (S2 – S1). ^{a,b} Means with differing superscripts differ (P < 0.05).

their survey responses. They were asked to return one week later for a second session (S2) where they were shown another collection of video clips, as in S1. Assignment was based on cattle experience level (experienced, inexperienced), age (college, other), and gender (male, female). Final distribution of participants for each treatment is provided in Table 1.

The first group of participants served as the control (C, n = 31), receiving no training between sessions. Participants assigned to training program 1 (T1, n = 30) watched a 20-minute training video that discussed the scoring system in detail and included short video clips as illustrations. Participants assigned to training program 2 (T2, n = 29) watched the same training video as T1 but were then asked to complete a self-test consisting of 10 additional video clips. Participants assigned to T2 were then given the opportunity re-watch each clip and read an explanation regarding the scoring of each animal.

Statistical Analysis

Inter- and intraobserver reliabilities were calculated. Interobserver reliability measured accuracy by comparing an individual's score of a video clip to that of the trained experts collected the day the video was recorded. Intraobserver reliability measured precision by comparing a participant's scores when viewing the same video clip multiple times.

Using the statistical package R, reliabilities were evaluated by percent agreement (PA). The PA is the ratio of the number of times a participant's scores matched up—either the participant's score with the experts or the participant's score with themselves—with the total number of observations they provided. A PA of zero means no agreement while a PA of 100 means perfect agreement.

A further statistic, the weighted Cohen's Kappa (K) coefficient, was also obtained. The values of K vary from -1 to 1. Negative values indicate agreement is poorer than chance, a zero indicates agreement is entirely by chance, while positive values indicate agreement that is better than chance.

The effect of preexisting biases (experience level, age, and gender) on accuracy and precision during S1, and on the change in reliability between sessions, was also assessed. The SAS statistical package was used for these analyses. Least-squares means and their standard errors were obtained. The means were compared applying a Tukey's adjustment.

Results

Experience level, age, and gender had no effect on accuracy or precision when assigning chute score during S1. Individuals with prior cattle handling experience appeared to be no better or worse at assessing behavior than those without experience. Overall, accuracy (interobserver reliability) for S1 was 0.62 and 50.5% for K and PA, respectively. Precision (intraobserver reliability) for S1 was 0.66 and 56.1%, respectively.

To assess changes in accuracy and precision between sessions because of training, differences in the assigned chute scores between S1 and S2 were determined. There were still no effects of experience level, age, or gender on change between sessions (P > 0.23).

Training, however, improved the accuracy (interobserver reliability) of the assessments of temperament (P < 0.01). The values of K increased between sessions by 0.00 ± 0.02, 0.09 ± 0.03, and 0.10 ± 0.03 for C, T1, and T2, respectively. Although the two training methods improved accuracy compared to the control, the extent of that improvement did not differ between them (Figure 1). They did, however, result in final K values that were 0.68 ± 0.02 and 0.73 ± 0.02 for T1 and T2, respectively. The same outcome was observed for PA. Following the training, the PA improved to a similar extent for both training methods, with little change in the control (Figure 1). Clearly, the training video increased the accuracy of chute score assessment, regardless of treatment group. There was minimal additional benefit, however, in adding the self-test.

Conversely, precision (intraobserver reliability) increased between sessions not only for the two training methods but also for the control. That general improvement was to such an extent that size of the change did not differ among them (P > 0.31). The K values increased by 0.05 ± 0.03, 0.08 ± 0.03, and 0.13 ± 0.03 for C, T1, and T2, respectively. Increases in PA were also similar among the three groups (Figure 1). Arguably, since the increases in accuracy and precision were similar for T1 and T2, this lack of significance was due to the increase in precision within C.

Without training, the control group became more precise while, if anything, less accurate when assigning chute score; in other words, they became more consistently incorrect in their assessments of calf temperament. When chute scores are incorporated into a docility Expected Progeny Difference (EPD), less accurate evaluations of temperament are less a concern. Differences in mean scores across operations, which reflect accuracy, are accounted for in the genetic evaluation itself. In this case, increased precision is more beneficial than increased accuracy.

By viewing the training video, participants not only became more precise but also more accurate in assigning a chute score. In the commercial industry, where culling may be based on an animal's score during handling, misallocation may result in poorer decision-making. For instance, if a restless heifer (score 3) is deemed acceptable as a replacement cow but not a nervous one (score 4), those temperaments need to be accurately distinguished. Therefore, when selecting cattle based on their phenotype alone, or when comparing the temperaments of cattle across operations, scores need to be assigned both accurately and precisely.

Implications/Conclusions

Prior to training, individual assessments of temperament of beef cattle behavior while restrained in a chute were inexact. Such was the case regardless of prior cattle handling experience, age, or gender. Precise measurements are important for reliable genetic evaluations. When selecting, or culling, cattle based on their assigned chute score, accuracy also matters. Incorporation of a short training video significantly increased participants' ability to assess chute score. When producers make decisions within their operation to select for docile cattle, it is imperative that these decisions are as accurate and precise as possible. When they are, improvements in the overall temperament of a herd can be achieved more quickly. To assist those producers wishing to gain skills in assigning chute scores, the training video, as well as some additional materials, are available online at https://beef.unl.edu/learning-modules.

Jamie T. Parham, Neogen GeneSeek Operations, Lincoln, NE

Jessica J. Schmidt, undergraduate honors student, University of Nebraska–Lincoln

Ronald M. Lewis, full professor, Animal Science, University of Nebraska–Lincoln