Journal of Animal Science, 2022, **100**, 1–8
https://doi.org/10.1093/jas/skac040
Advance access publication 10 February 2022
Animal Behavior and Cognition

Cattle acclimate more substantially to repeated handling when confined individually in a pen than when assessed as a group

Jamie T. Parham,[†] Amy E. Tanner,[‡] Sarah R. Blevins,[‡] Mark L. Wahlberg,[‡] and Ronald M. Lewis[†,‡,1,]

[†]Department of Animal Science, University of Nebraska-Lincoln, Lincoln, NE 68583, USA
[‡]Department of Animal and Poultry Sciences, Virginia Tech, Blacksburg, VA 24061, USA

[1]Corresponding author: rlewis5@unl.edu

Abstract

Chute (**CS**) and exit (**ES**) scores are common subjective methods used to evaluate temperament in cattle production systems. A pen test, which allows behavior to be observed in a non-restrained setting, may also be an effective method to evaluate temperament by allowing more variation among animals to be expressed. The merit of pen scores in assisting producers in evaluating temperament is equivocal. The objectives of this study were to validate the usefulness of a pen score in delineating temperamental cattle and to determine whether these behavioral scores change under repeated and routine management. Over 3 consecutive years, a factorial design of two measurement protocols (frequent [F], infrequent [IN]) and three recording periods was used. The F measurements were collected over 3 consecutive days and IN measurements only on day 1 within a recording period. Each year, 20 mostly Angus commercial *Bos taurus* heifers were randomly assigned to each protocol. Behavior was measured using a CS, ES, and exit velocity. Body temperature and heart rate also were recorded. A fecal and blood sample were collected and analyzed for levels of various metabolites including glucose concentration and serum cortisol. Following routine handling, each heifers' response to 30 s of exposure to a human stressor was recorded both individually and in groups of four. An individual (**IPS**) and group (**GPS**) pen scores were assigned from 1 (docile) to 6 (aggressive). For all heifers, protocol, event, and their interaction, were compared on the first day of an event. For F heifers, event and day within event were instead used. Body weight was included as a covariate, with sire and year fitted as random effects. Reliability of IPS and GPS were determined using a kappa (K) coefficient. Both IPS and GPS were reliably assigned (K = 0.64 and 0.44 for IPS and GPS, respectively) and positively correlated with body temperature, heart rate, glucose, and serum cortisol ($r = 0.28$ to 0.37). Furthermore, F heifers acclimated to repeated handling in an individual pen setting ($P < 0.05$) while acclimation to handling within groups was not evident ($P > 0.14$). IPS provides a reliable evaluation of temperament in a non-restrained setting that is indicative of an animal's response to stress and may be useful when attempting to make phenotypic selection decisions. However, temperamental heifers became calmer with repeated gentle handling.

Lay Summary

Chute and exit scores are common subjective methods used to evaluate temperament in cattle production systems. A pen test, which allows behavior to be observed in a non-restrained setting, may also be an effective method to evaluate temperament by allowing more variation among animals to be expressed. However, the merit of pen scores in assisting producers in evaluating temperament has yet to be discerned. Therefore, the objectives of this study were to validate the usefulness of pen scores in delineating temperamental cattle and to determine whether these behavioral scores change under repeated and routine management. Pen scores collected on heifers either individually or as a group could be assigned reliably and were indicative of an animal's response to stress during normal handling practices. Temperamental heifers, when handled more frequently, acclimated to repeated handling in an individual pen setting but not in a group. Therefore, regardless of method, when cattle are excitable during their first handling experience, more than one observation of temperament may be beneficial before assessing temperament.

Key words: acclimation, beef cattle, pen score, temperament

Abbreviations: BUN, blood urea nitrogen; CK, creatine kinase; CS, chute score; ES, exit score; EV, exit velocity; GPS, group pen score; ICC, intra-class correlation; IPS, individual pen score; K, Fleiss' kappa coefficient; NEFA, nonesterified fatty acids

Introduction

To effectively select for docility in cattle, criteria measured must be indicative of behavior during normal handling practices (Fordyce et al., 1982). Common methods proposed to quantify temperament are based on behavior when cattle are restrained in (chute score [CS; Tulloh, 1961]) and exiting from (exit score [ES; BIF Guidelines, 2002]; exit velocity [EV; Burrow et al., 1988]) the chute. CS and ES can be quickly and reliably assessed regardless of the assessors' prior experience (Parham et al., 2019a), and are correlated with objective measurements of stress (Parham et al., 2021).

Another less common evaluation of temperament has been pen score, which can be assessed either individually or as a small group. When an animal (or group of animals) was placed into a pen, its (or their) behavioral responses to the presence of a human inside and outside of their flight zone

Received October 25, 2021 Accepted February 8 2022.

were recorded using a five-point scale (Hammond et al., 1996; King et al., 2006). While individual pen scores (**IPS**) allowed the observer to quantify the temperament of the animal when placed in a secluded situation, several studies demonstrated a positive influence of group social environment on the behavioral response of farm animals (Gringard et al., 2000). For pen score to be an effective method to evaluate temperament in cattle, it not only must be indicative of physiological stress, but also reliably assessed and truly representative of an animal's temperament.

The correlations of pen score with CS and EV ranged from 0.24 to 0.75 (Cooke et al., 2009; Turner et al., 2011). However, their associations with objective measures of stress are less well known. Objective physiological measures have been considered useful indicators of stress that are not affected by observer bias, which may occur when assigning subjective scores (Hoppe et al., 2010). Cooke et al. (2009) reported a correlation between pen score and plasma cortisol of 0.44. Parham et al. (2019b) found that cattle acclimate to repeated routine management while restrained in a chute. This routine management included animal handling such as weight measurement, jugular blood draw, and rectal palpation. The reliability of multiple observers' assignments of pen score on a given animal, and the consistency of pen scores over time with repeated handling, has yet to be determined.

The first objective of this study was to assess the reliability of pen score, and its relationship with other objective and subjective measures of stress, in cattle shortly following weaning. The second objective was to assess whether their pen score changed following repeated routine handling. It was hypothesized that both IPS and group pen score (**GPS**) were correlated with physiological measures of stress as well as other subjective measures including CS and ES in a predictable direction. We also hypothesized that the values of IPS and GPS would reflect acclimation to calm, gentle handling.

Materials and Methods

All procedures and protocols used in this study were approved by the Institutional Animal Care and Use Committee at Virginia Tech, Virginia, USA.

Experimental animals

This study builds on previous work evaluating temperament in heifers shortly after weaning when confined in a squeeze chute (Parham et al., 2019a, 2019b, 2021). Across 3 consecutive years, 120 commercial *Bos taurus* (75% Angus or more) spring-born heifer calves (n = 40 per year) were reared at the Virginia Tech Shenandoah Valley Agricultural Research and Extension Center in Steeles Tavern, Virginia, USA with their respective dams until weaning (185 ± 11 d in age). Calves were exposed to working facilities on two separate occasions up through weaning: at approximately 2 ½ mo in association with breeding of their dams and at approximately 6 mo when weaned. At each event, they were vaccinated and dewormed. The facilities included a series of holding pens. An adjoining alleyway led to a curved alley, which then led to a weigh crate and separate squeeze chute (Pearson Manual Chute).

Once separated from their dams at weaning, the calves completed a 1-wk fence line weaning period. The heifers were then transported to Virginia Tech Kentland farm, Virginia, USA. They were placed in a single management group on grass for approximately 1 mo prior to the start of the current study.

The heifers were daughters of 21 sires, ranging from 1 to 23 daughters per sire, selected to establish divergent larger and smaller frame size offspring. Cows were bred within line to Angus sires correspondingly selected as larger or smaller based on their mature cow height estimated breeding value (Vargas Jurado et al., 2015).

Experimental design and data collection

Prior to the start of the study, the three individuals (observers) with responsibility for assigning the subjective temperament scores throughout the experiment studied and discussed the ethograms to be used, described later (Tulloh, 1961; BIF Guidelines, 2002; King et al., 2006). Throughout the morning of the first day of observations, the three observers discussed the score to be assigned to each animal at each stage of the evaluation. At the midpoint and end of that first day, these observers again discussed the scoring system to confirm their consistency. Thereafter, at the start of each day's measurements, the scoring systems were again reviewed.

As described in Parham et al. (2019a), in each of 3 yr, heifers were randomly assigned to one of two measurement protocols (F, IN) within their dam's frame size category (larger, smaller) and sire. Data were collected across three recording periods, termed an "event," each 1 mo apart (i.e., event 1 [October], 2 [November], 3 [December]) starting on the second Monday or Tuesday of October each year. Heifers within the F measurement protocol were observed 3 consecutive days within each event (month), while the heifers in the IN measurement protocol were evaluated on only the first day of an event. Day within event was designated by $d_{i,j}$, where i was the event and j was the day within an event.

On day 1 of each event (i.e.,$d_{1,1}$, $d_{2,1}$, $d_{3,1}$) every year, all 40 heifers were moved into a holding pen. Four heifers were randomly drawn from the group and herded into the cattle-handling facility, regardless of measurement protocol. The facility consisted of a small holding pen narrowing into a curved alley that led to a weigh crate and separate squeeze chute. One at a time, each heifer was calmly moved through the alley into the weigh crate. Once weighed, each heifer was then moved into the squeeze chute (Priefert Model S04) where their head was secured in the head gate and the sides of the chute left opened with no restriction on the body.

On a given day, the three observers simultaneously recorded a CS (1 = docile to 6 = aggressive; Tulloh, 1961) during the first 15 s of restraint followed by a heart rate and rectal temperature. Each heifer's head was further secured to the side using a halter rope so a jugular blood sample could be taken into separate ethylenediaminetetraacetic acid, lithium-heparinized, and plain serum tubes and stored on ice until plasma and serum were collected via centrifugation. Fecal samples were also taken and double-wrapped in prelabeled aluminum foil and snap-frozen in liquid nitrogen. The feces were stored at –20 °C until analyzed for fecal cortisol levels.

Upon release, an ES (1 = docile to 5 = aggressive; BIF Guidelines, 2002) was recorded by the same individuals. An EV (Burrow et al., 1988) was also measured using electronic timers (Polaris Multi-Event Timer) over a 2 m distance, beginning 1 m from the head of the chute.

The heifer was then calmly walked to a 12 × 6 m pen where they were exposed to the same human stressor. The human stressor wore similar insulated clothing to the other study

participants and did their best to wear the same clothing during each cattle-handling experience. Once closing the gate behind, this person walked into and stood in the center of the pen for 30 s. During this time, IPS were assigned independently by the same observers using an ethogram based on King et al. (2006), provided in Table 1.

Each heifer was then moved into a 12 × 8 m pen separated from the individual pen by an alleyway with no direct and limited visual contact. Heifers remained in the group pen until four heifers had been moved through the working facilities and placed together. Once in a group of four, the same ethogram (adapted from King et al., 2006; Table 1) was used to assign a GPS to each heifer individually by the same observers, based on their reaction to the same human stressor. In this case, the person walked to the center of the pen, paused, and then continued diagonally in the direction of the group of heifers before returning to the center of the pen. This process was repeated until all 40 heifers had been observed.

Each year on days 2 and 3 of each event (i.e.,$d_{1,2}$, $d_{2,2}$, $d_{3,2}$, $d_{1,3}$, $d_{2,3}$, $d_{3,3}$), the same measurements were again recorded on all heifers assigned to the F protocol (n = 20). The exception was that no blood sample was collected on day 2. However, researchers still simulated a "mock" blood sample, including tapping the occluded jugular but no needle prick, for consistency in experience.

Between events, the F heifers were grazed together in a pasture adjacent to the working facilities. The remaining IN heifers were returned to their original pasture. After day 3 of recording, all 40 heifers were mixed into a single management group until the next event. Across the 3 yr, 2 of the 120 heifers, 1 from F and the other IN, were removed from the study due to lameness.

Physiology

Fecal analysis

Frozen fecal samples were thawed at room temperature and weighed into 0.5 g samples and mixed with 5 mL of 80% methanol. Samples were then centrifuged for 15 min at 2,500 × *g*, and the supernatant removed and stored at –20 °C. Samples obtained were analyzed in duplicate for cortisol metabolites using the Siemens Coat-A-Count Cortisol Radioimmunoassay (Siemens Medical Solutions Diagnostics, Los Angeles, CA, USA) according to manufacturer's instructions.

Laboratory analytical error was assessed as the ratio of the absolute value of the difference between duplicates and their mean. If that value exceeded 0.10 (or 10%), the analyses were repeated until the standard was achieved. The same protocols were applied to the serum analyses.

Serum analysis

Serum cortisol concentrations were measured in duplicate using the Siemens Coat-A-Count Cortisol Radioimmunoassay (Siemens Medical Solutions Diagnostics). Concentrations of the serum chemistry metabolites of blood urea nitrogen (**BUN**), creatine kinase (**CK**), and glucose were measured in duplicate using the QuantiChrom Urea Assay, EnzyChrom Creatine Kinase Assay, and QuantiChrom Glucose Assay Kits (BioAssay Systems, Hayward, CA, USA), respectively, using a clear bottom 96-well plate and plate reader according to manufacturer's instructions. With regard to CK, if calculated activity was higher than 300 units per liter (U/L), the sample was diluted 1:10 in 0.9% saline and repeated, per manufacturer's instructions. Serum concentrations of nonesterified fatty acids (**NEFA**) were determined using the Wako HR series NEFA-HR(2) microtiter assay (Wako Diagnostics, Richmond, VA, USA).

Statistical analyses

Interobserver reliability, or the consistency among the three observers of both IPS and GPS, was calculated using Fleiss' kappa coefficient (K) and an intra-class correlation (**ICC**). All reliability calculations were carried out using the irr package (Gamer et al., 2012) in R.

Pearson correlations were calculated between IPS or GPS with CS, ES, EV, and all other physiological indicators of behavior for heifers on the first day of each event ($d_{1,1}$, $d_{2,1}$, and$d_{3,1}$) in R (R Core Team, 2013). Correlations were first calculated for F and IN separately. Confidence intervals were derived from standard errors estimated using a Fisher's z-transformation (Fisher, 1921). In most cases (more than 79%), coefficients did not differ in magnitude between protocols. Therefore, the Pearson correlations were estimated for the combined data. The value of the coefficient was tested for whether it was significantly different from zero ($P < 0.05$; n = 118), assuming a linear relationship between variables. Correlations of IPS and GPS with CS, ES, and EV appeared to increase over time and were reported separately.

Table 1. Pen score ethogram used to measure temperament in heifers both individually and in a group

Pen score	Individual pen score description[1]	Group pen score description[1]
1. Docile	Walks slowly, can be approached slowly, not excited by humans	Walks slowly, can be approached slowly, not excited by humans
2. Slightly restless	Aware of humans, head up, moves away from approaching human, runs fence line, stops and looks around	Aware of humans, head up, moves slowly away from approaching human
3. Restless	Constantly runs along fence line, head up	Runs along fences stands in corner if humans stay away
4. Nervous	Agitated, runs along fence line, head up, looking for a way of escape, and will run if humans come closer, stops before hitting gates and fences, avoids humans	Runs along fences, head up and will run if humans come closer, stops before hitting gates and fences, avoids humans
5. Very nervous	Runs, head high and very aware of humans, may run into fences and gates, flighty	Runs, stays in back of the group, head high and very aware of humans, may run into fences and gates
6. Wild (aggressive)	Excited, runs into fences, runs over anything in its path	Excited, runs into fences, runs over anything in its path

[1]Adapted from King et al. (2006).

IPS and GPS taken by the three observers were averaged on each day to obtain a representative score. Based on examination of residuals, the distribution of these data appeared skewed. A natural logarithm was applied with the transformed values tested for normality using the Jarque–Bera (Skewness–Kurtosis) Test (Jarque and Bera, 1980). The log-transformed data were more normal with less skewness and kurtosis. Therefore, lognormal transformed average IPS and GPS were analyzed using the GLIMMIX procedure in SAS (SAS Inst. Inc., Cary, NC, USA) fitting two separate models described later. However, to express the results on their scale of measurement, means and SE were back-transformed to the observed scale.

To compare the effect of measurement protocol on IPS and GPS, initially a 2 × 2 × 3 factorial model (model 1) was analyzed fitting protocol (F and IN), dam frame size (larger or smaller), and event (1, 2, or 3), and their two and three-way interactions, as fixed effects. Year, sire, and heifer nested within the combination of year, measurement protocol, and dam frame size, were treated as random effects. Comparisons were only made on the first day of each event ($d_{1,1}$, $d_{2,1}$, $d_{3,1}$) such that all heifers regardless of measurement protocol were present. Dam frame size never explained significant variation in the response variables ($P > 0.12$) and thus was excluded from the model. The final fixed effect model fitted therefore included protocol, event, and their interaction, with the addition of body weight as a covariate.

To measure changes in response variables over time within F, a separate model (model 2) was fitted. Event, dam frame size, their interaction, and the nested effect of day within event, were fitted as fixed effects. Heifer nested within year and dam frame size combination, as well as year and sire, were treated as random effects. Again, dam frame size did not define variation in the response variables ($P > 0.24$) and was removed from the final model fitted. However, body weight was included as a covariate.

A repeated measures analysis was initially conducted as in Parham et al. (2019b). Results from these analyses were consistent with the initial models described. Therefore, results obtained from the simpler factorial models are reported henceforth.

Results

Interobserver reliability

IPS had an average K of 0.64 and ICC of 0.92. Reliabilities of GPS were lower, with an average K of 0.44 and ICC of 0.77. However, for both IPS and GPS, K and ICC were above reported threshold values for acceptable reliability, namely >0.40 (Landis and Koch, 1977) and >0.70 (Martin and Bateson, 1993), respectively, indicating accurate evaluation.

Based on an earlier study considering the same cattle (Parham et al., 2019a), ES assessments also had acceptable reliabilities with K of 0.73 and ICC of 0.90. Reliabilities of CS evaluations were lower, yet still dependable with coefficients of 0.46 for K and 0.74 for ICC.

Relationship between measurements

Pearson correlations for IPS and GPS with objective measures of temperament are provided in Table 2. Pen scores were positively correlated with body temperature, heart rate, glucose concentration, and serum cortisol ($P < 0.001$); there was a lower positive correlation with CK ($P = 0.03$). However, correlations with BUN and fecal cortisol were not different from zero. Finally, negative correlations existed for both IPS and GPS with NEFA concentrations ($P < 0.01$). These values were consistent with correlations of these same objective measures with CS, ES, and EV previously reported by Parham et al. (2021). Table 2 also includes the average value for each measurement throughout the study. Parham et al. (2021) reported that while (these same) heifers with higher subjective scores (CS and ES) had greater physiological responses to handling ($r = 0.24$ to 0.33), there was no significant change in the concentration of metabolites over time. The exception was CK.

Pearson correlations for IPS and GPS with CS, ES, and EV for $d_{1,1}$, $d_{2,1}$, and $d_{3,1}$ separately and combined are given in Table 3. Overall, there were strong correlations between IPS and GPS with ES and EV ($P < 0.01$), as both were non-restrained measures of temperament. On the first day of observation, the correlations between IPS and GPS with CS were lower. However, they increased over time ($P < 0.05$). This pattern was not present for ES and EV, although the correlations tended to be lower on $d_{1,1}$ compared with $d_{2,1}$ and $d_{3,1}$ ($P < 0.10$).

Table 2. Pearson correlations (r) of IPS and GPS with objective measurements of temperament

Measure[2]	n	$\bar{X}^1$	Individual pen score ($\bar{X}^1 = 2.32$)		Group pen score ($\bar{X}^1 = 1.80$)	
			r	SE	r	SE
Temperature, °C	350	39.33	0.37	0.05	0.30	0.05
Heart rate, bpm	351	127.75	0.28	0.05	0.29	0.05
BUN[3,4], mg/dL	351	34.27	0.04	0.05	0.02	0.05
CK[3], units/L	351	12.94	0.10	0.05	0.14	0.05
Glucose, mg/dL	351	117.47	0.28	0.05	0.28	0.05
NEFA[3], mmol/L	351	0.35	−0.23	0.05	−0.13	0.05
Serum cortisol, ng/mL	350	44.45	0.28	0.05	0.30	0.05
Fecal cortisol[4], ng/mL	344	11.49	−0.03	0.05	−0.02	0.05

[1]Mean value for each physiological measurement and categorical score throughout the study.
[2]Details of the changes in physiological measures over time were reported elsewhere (Parham et al., 2021).
[3]BUN, blood urea nitrogen; CK, creatine kinase; NEFA, nonesterified fatty acids.
[4]Correlations not different from zero ($P > 0.05$).

Acclimation to handling

When comparing measurement protocols for change in IPS over time, there was an interaction of measurement protocol and event ($P = 0.03$). In Figure 1, the mean IPS for F and IN heifers on the first day of each event are provided. IPS did not differ between F and IN on $d_{1,1}$, but decreased in the F group from $d_{1,1}$ to $d_{2,1}$ ($P = 0.03$) and remained constant from $d_{2,1}$ to $d_{3,1}$ ($P = 0.99$). However, IPS on $d_{2,1}$ and $d_{3,1}$ in F were not different from those for IN on those same days, although their values were numerically smaller. Final IPS on $d_{3,1}$ for the F and IN groups was 1.35 ± 0.05 and 1.75 ± 0.07, respectively.

There was no effect of event, measurement protocol, or their interaction on GPS ($P > 0.26$). Mean GPS for all heifers on the first day of each event were 1.72 ± 0.05, 1.58 ± 0.04, and 1.51 ± 0.04 for $d_{1,1}$, $d_{2,1}$, and $d_{3,1}$, respectively.

When assessing change in measurements across days for F heifers, there was a decrease in IPS across both events ($P < 0.01$) and days ($P < 0.01$). IPS decreased from 1.98 ± 0.06 during event 1 to 1.53 ± 0.05 during event 2 ($P = 0.01$), but with a smaller further decline to 1.31 ± 0.04 for event 3 ($P = 0.07$). Figure 2a shows the change in IPS across days; there was a significant difference between $d_{1,1}$ and $d_{1,2}$ with all other days. The IPS at $d_{1,3}$, did not differ with those scores recorded at the second event ($d_{2,1}$, $d_{2,2}$, and $d_{2,3}$, $P > 0.05$). However, the difference became more substantial when comparing IPS at $d_{1,3}$ to the last 3 d of observation ($P < 0.05$). By $d_{3,3}$, IPS reduced to 1.29 ± 0.04, which was less than $d_{1,1}$ through $d_{2,1}$ ($P < 0.05$).

The GPS did not change for F heifers over time ($P > 0.10$), with mean values for events 1 to 3 of 1.68 ± 0.06, 1.48 ± 0.06, and 1.39 ± 0.05, respectively. Mean GPS across days is provided in Figure 2b. A low GPS of 1.69 ± 0.06 on $d_{1,1}$ left little room for any further decrease. However, GPS on $d_{3,3}$ was numerically the lowest with a value of 1.32 ± 0.05.

Discussion

IPS and GPS are suitable methods to assess stress in cattle when exposed to routine handling. They were reliably assessed by multiple observers and were positively correlated with body temperature, heart rate, glucose, and serum cortisol. IPS and GPS were also positively correlated with ES and EV across all days of collection; however, their correlations with CS were initially lower but increased over time reaching moderately positive values. The increase in the strength of the correlation of IPS and GPS with CS was conceivably due to acclimation. Frequently handled heifers appeared to acclimate more substantially to handling, especially across days during

Table 3. Pearson correlations of IPS and GPS with CS, ES, and EV

Day[1]	Individual pen score			Group pen score		
	CS[2]	ES[2]	EV[2]	CS	ES	EV
$d_{1,1}$	0.26 ± 0.09[a]	0.60 ± 0.07	0.45 ± 0.08[a]	0.15 ± 0.09[3,a]	0.42 ± 0.08	0.33 ± 0.09
$d_{2,1}$	0.41 ± 0.09[ab]	0.70 ± 0.07	0.55 ± 0.08[ab]	0.41 ± 0.09[b]	0.54 ± 0.08	0.41 ± 0.09
$d_{3,1}$	0.54 ± 0.08[b]	0.67 ± 0.07	0.63 ± 0.07[b]	0.47 ± 0.08[b]	0.53 ± 0.08	0.45 ± 0.08
All[4]	0.42 ± 0.05	0.65 ± 0.04	0.54 ± 0.05	0.36 ± 0.05	0.50 ± 0.05	0.40 ± 0.05

[1]Day within event is designated by $d_{i,j}$, where i is the event and j is the day within an event.
[2]CS, chute score; ES, exit score; EV, exit velocity.[3]Correlation not different from zero ($P > 0.05$).
[4]Since all combines information across 3 individual days, these correlations cannot be independently compared with those on $d_{1,1}$, $d_{2,1}$, or $d_{3,1}$.
[a,b]Means in a column with differing superscripts differ ($P < 0.05$).

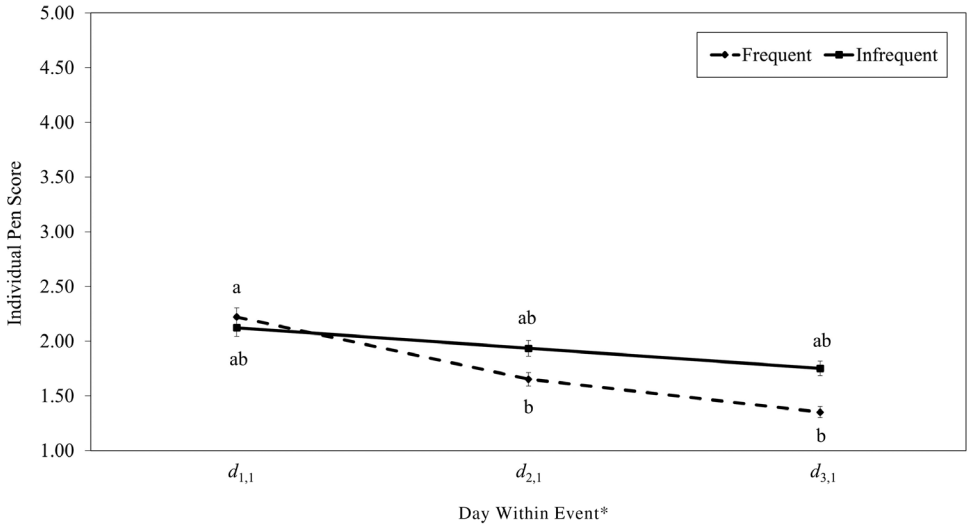

Figure 1. Change in IPS over time for F and IN handled heifers. [a,b]Means with differing letters differ ($P < 0.05$). *Day within event is designated by $d_{i,j}$, where i is the event and j is the day within an event.

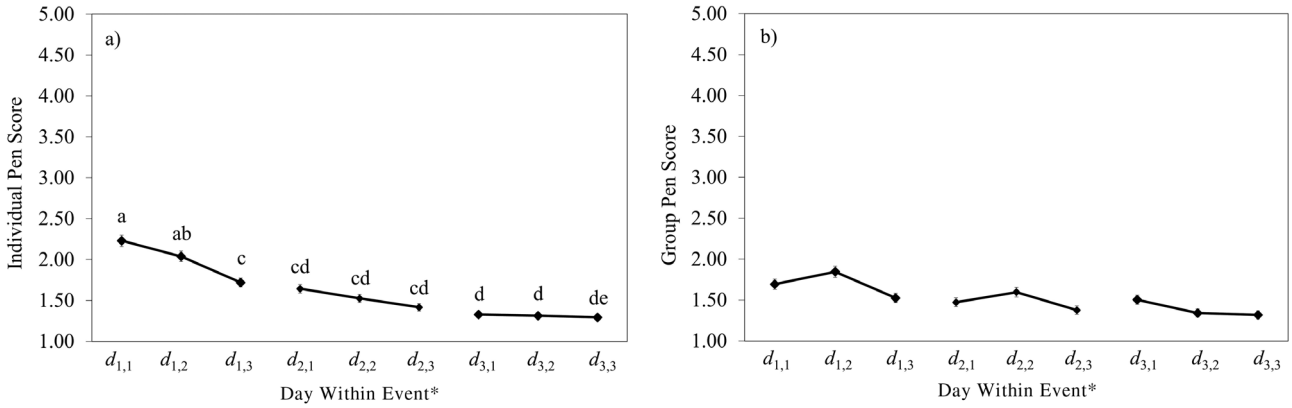

Figure 2. Change in pen scores across days for F handled heifers. Plot (a) change in IPS and (b) change in GPS. [a,b]Data points with differing letter assignments differ ($P < 0.05$). *Day within event is designated by $d_{i,j}$, where i is the event and j is the day within an event.

event 1. With lower starting values, there was no significant change in GPS over time. This possibly was due to calmer temperament when heifers were in a group setting. Since IPS resulted in more visually detectable responses to stress, it may be the more useful pen score for producers wishing to evaluate temperament in an animal. However, heifers acclimated to repeated calm handling. More than one observation of temperament therefore may be beneficial before making selection decisions. This is especially true when cattle are more excitable during their first handling experience.

Interobserver reliabilities for IPS and GPS were consistently higher than published thresholds for acceptable accuracy (Landis and Koch, 1977; Martin and Bateson, 1993). The IPS and GPS reliabilities were even higher than those for CS assigned by the same individuals (Parham et al., 2019a). Furthermore, they were positively correlated with several physiological measures known to be associated with stress in animals (Van de Water et al., 2003; Sporer et al., 2008). The strength and direction of these correlations were consistent with those reported for CS and ES by Parham et al. (2021) within the same set of animals, indicating IPS and GPS were acceptable evaluators of stress. Similarly, correlations of IPS and GPS with CS, ES, and EV were positive; the size of these correlations increased from $d_{1,1}$ to $d_{3,1}$ with CS. Lastly, correlations were consistently higher for ES than CS. This is perhaps because ES, IPS, and GPS were all non-restrained methods of evaluating temperament.

All correlations and interobserver reliabilities were higher for IPS than GPS. There are two possible explanations. First, in being herd animals the presence of social partners reduced heifers' behavioral signs of disturbance toward fear-eliciting stimulation (Boissy and Le Neindre, 1990). When secluded in an individual pen, calves spent more time standing still and were easier to handle when peers were present in an adjacent pen as compared with when no peers were present (Gringard et al., 2000). This lower responsiveness may lead to lower correlations with measures known to be indicative of stress. It also may lead to a decrease in variability of the behaviors expressed, making it more difficult to delineate temperaments among animals in a group.

The second cause of a reduction in reliability when comparing GPS with IPS could be due to the order in which heifers were evaluated. Based on their own choice, the three observers watched one heifer, assigned their score, and then moved to the next. By evaluating the heifers in different orders, they may have observed different expressions of behavior leading to slightly different scores being assigned to the same animal.

Average GPS on the first day of observation (1.69 ± 0.06) indicated the heifers were docile when placed in a group, generally. Responsiveness to fear-eliciting stimulation when in the presence of peers was less. It could be argued that the presence of peers masked the actual temperament of excitable animals. Heifers did, however, acclimate to repeated handling in an individual pen setting. Frequently handled heifers decreased in IPS more significantly from $d_{1,1}$ to $d_{3,1}$; such was not the case with the infrequently handled heifers. When assessing change in average IPS across days in F, the largest decrease in temperament occurred during the first event. However, IPS was lowest on $d_{3,3}$, and significantly lower than on $d_{1,1}$, $d_{1,2}$, and $d_{1,3}$. This observation is consistent with change in CS within F as reported by Parham et al. (2019b).

As discussed by Parham et al. (2019b), a potential explanation for the acclimation of heifers to individual pen restraint may be due to what is deemed personality, instead of temperament. Personality is defined as inherited, early appearing tendencies (Finkemeier et al., 2018) that must be consistent and repeatable (Mackay and Haskell, 2015). Temperament is used more broadly to describe how an animal reacts to a situation (Mackay and Haskell, 2015). In each subjective measurement of behavior, heifers in this experiment had expressed both their underlying personality and their current behavioral responses, or temperament. Based on the definitions of Mackay and Haskell (2015), it could be argued that an animal's initial reaction to handling was the most reliable estimate of their temperament, while multiple observations would allow for an estimate of their personality. Selection decisions could be based on either situation depending on a producer's priorities. If an animal's temperament was completely unmanageable on the first day of handling, or was a primary trait of interest, then selection decisions based on that initial assessment would result in a more docile herd over time. However, allowing for acclimation to handling, or expression of personality, may be of value when a specific animal had borderline acceptable temperament.

Although it takes more time, evaluating heifers individually rather than in a group setting may be more useful simply because greater variations in behavior were expressed. The ability of an animal to display their full repertoire of behaviors

in response to stress impacts the effectiveness of an ethogram. Grandin (2014) warns that the utility of CS depends on how tightly the animal was restrained. Catching the head of an animal in the head gate, and/or narrowing the width of side panels, restrict movement thereby reducing variation in behavioral response (Vetters et al., 2013). Conversely, in being a non-restrained test, IPS allows the animal freedom to move and behave as they choose within the confines of the pen. As an example, cattle exposed to people daily had smaller flight zones than cattle raised on pasture (Grandin and Deesing, 2014). Differences in flight zone will impact how stressed an animal appears when secluded in a pen with a human, and the behaviors they express in trying to escape that threat. Removing restraint therefore allows for a more comprehensive evaluation of the temperament of an animal. Heifers that acclimated to handling in this study also acclimated to handling in the chute (Parham et al., 2019b), but to a larger degree. It could be hypothesized that the larger decrease in IPS as compared with CS with repeated handling reflected the opportunity for expression of a larger range of behaviors when an animal's movement was not restrained.

When comparing the usefulness of CS and IPS to measure temperament of an animal, an important consideration was which was safer and easier to implement in a production setting. The initial use of pen score was proposed by Le Neindre et al. (1995). Using this method, an animal was isolated individually in a pen with a handler who had 2 min to direct it into a corner, hold it there for 30 s, and then stroke it, after which a subjective score was assigned from 1 (calm) to 5 (very excited) based on an animal's response. Concern arose about handler safety while attempting to stroke the animal (Kilgour et al., 2006). The method was curtailed to the handler simply standing in the middle of the pen for 30 s, with no attempt made to restrain the animal (Turner et al., 2011). This method was instead referred to as an "isolation score" and rated on a scale of 1 to 6. This revised IPS still introduced a concern for handler safety avoided when animals were restrained in a chute. As an illustration, the human in this study did not feel safe enough to enter the pen with a heifer 10% of the time.

Secondly, most, if not all, producers will likely place their cattle in a chute during their first year of life. Therefore, CS would provide an easy method of quantifying temperament that requires little extra time or effort to utilize. It would be less likely that producers would be individually secluded in a pen with their cattle as part of normal handling practices. Recording an IPS on each animal therefore would require more time, resources, and effort. Although there was more variation in response when using a non-restrained test such as pen score, safety, and ease of use also should be considered when producers choose between methodologies.

In conclusion, pen scores collected on heifers either individually or as a group were reliably assigned and were indicative of an animal's response to stress during normal handling practices. Due to cattle being a herd species, responses to stress were muted when temperament was analyzed in a group setting. Therefore, IPS may prove more useful than GPS to categorize behavior. Safety and ease of use should, however, be taken into consideration when choosing among methodologies. Lastly, cattle did acclimate to repeated exposure to a human stressor in an individual pen setting. Therefore, regardless of method, with cattle more excitable during their first handling experience, more than one observation of temperament may be beneficial before assessing temperament.

Acknowledgments

This project was based on research supported by the U.S. Department of Agriculture (USDA), Agricultural Research Service (ARS), Award Number 1932-21630-003-06. We thank Dr. Jim Neel (USDA-ARS, Forage and Livestock Production Research, El Reno, OK) for his advice on experimental protocols and equipment. The assistance of technical staff, and graduate and undergraduate students, particularly Roberto Franco, Napoleón Vargas Jurado, Katharine Barkley, Lyla Pullen, Ashley Kopanko, and Kathryn Slaughter, is sincerely appreciated.

Conflict of interest statement

The authors declare that there is no conflict of interest regarding the publication of this article.

LITERATURE CITED

BIF Guidelines. 2002. *Guidelines for uniform beef improvement programs*. 8th ed. Athens (GA): University of Georgia.

Boissy, A., and P. Le Neindre. 1990. Social influences on the reactivity of heifers: implications for learning abilities in operant conditioning. *Appl. Anim. Behav. Sci.* 25:149–165. doi:10.1016/0168-1591(90)90077-Q

Burrow, H. M., G. W. Seifert, and N. J. Corbet. 1988. A new technique for measuring temperament in cattle. *Proc. Aust. Soc. Anim. Prod.* 17:154–157.

Cooke, R. F., J. D. Arthington, D. B. Araujo, and G. C. Lamb. 2009. Effects of acclimation to human interaction on performance, temperament, physiological responses, and pregnancy rates of brahman-crossbred cows. *J. Anim. Sci.* 87:4125–4132. doi:10.2527/jas.2009-2021

Finkemeier, M. A., J. Langbein, and B. Puppe. 2018. Personality research in mammaliam farm animals: concepts, measures, and relationship to welfare. *Front. Vet. Sci.* 5:131–146. doi:10.3389/fvets.2018.00131

Fisher, R. A. 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* 1:3–32.

Fordyce, G., M. E. Goddard, and G. W. Seifert. 1982. The measurement of temperament in cattle and the effect of experience and genotype. *Proc. Aust. Soc. Anim. Prod.* 14:329–332.

Gamer, M., J. Lemon, and I. F. P. Singh. 2012. irr: various coefficients of interrater reliability and agreement. *R package version 0.84*. https://CRAN.R-project.org/package=irr.

Grandin, T. 2014. Handling facilities and restraint of extensively raised range cattle. In: Grandin, T., editors. *Livestock handling and transport*. 4th ed. Cambridge (MA): CABI International; p. 94–115.

Grandin, T., and M. J. Deesing. 2014. Genetics and behavior during handling, restraint, and herding. In: Grandin, T. and M. J. Deesing, editors. *Genetics and the behavior of domestic animals*. 2nd ed. London (UK): Publisher is Academic Press; p. 115–158.

Gringard, L., A. Boissy, X. Boivin, and J. P. Garel. 2000. The social environment influences the behavioural responses of beef cattle to handling. *Appl. Anim. Behav. Sci.* 68:1–11. doi:10.1016/S0168-1591(00)00085-X

Hammond, A. C., T. A. Olson, C. C. Chase, Jr., E. J. Bowers, R. D. Randel, C. N. Murphy, D. W. Vogt, and A. Tewolde. 1996. Heat tolerance in two tropically adapted *Bos taurus* breeds, Senepol and Romosinuano, compared with Brahman, Angus, and Hereford cattle in Florida. *J. Anim. Sci.* 74:295–303. doi:10.2527/1996.742295x

Hoppe, S., H. R. Brandt, S. König, G. Erhardt, and M. Gauly. 2010. Temperament traits of beef calves measured under field conditions and their relationships to performance. *J. Anim. Sci.* 88:1982–1989. doi:10.2527/2008-1557

Jarque, C. M., and A. K. Bera. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Econ. Letters* 6:255–259. doi:10.1016/0165-1765(80)90024-5

Kilgour, R. J., G. L. Melville, and P. L. Greenwood. 2006. Individual differences in the reaction of beef cattle to situations involving social isolation, close proximity of humans, restraint and novelty. *Appl. Anim. Behav. Sci.* 99:21–40. doi:10.1016/j.applanim.2005.09.012

King, D. A., C. E. Schuehle Pfeiffer, R. D. Randel, T. H. Welsh, Jr., R. A. Oliphint, B. E. Baird, K. O. Curley, Jr., R. C. Vann, D. S. Hale, and J. W. Savell. 2006. Influence of animal temperament and stress responsiveness on the carcass quality and beef tenderness of feedlot cattle. *Meat Sci.* 74:546–556. doi:10.1016/jmeatsci.2006.05.004

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*. 33:159–174. doi:10.2307/2529310

Le Neindre, P., G. Trillat, J. Sapa, F. Ménissier, J. N. Bonnet, and J. M. Chupin. 1995. Individual differences in docility in limousin cattle. *J. Anim. Sci.* 73:2249–2253. doi:10.2527/1995.7382249x

Mackay, J. R. D., and M. J. Haskell. 2015. Consistent individual behavioral variation: the difference between temperament, personality and behavioral syndromes. *Animals*. 5:455–478. doi:10.3390/ani5030366

Martin, P., and P. Bateson. 1993. *Measuring behaviour: an introductory guide*. Cambridge (UK): Cambridge University Press.

Parham, J. T., A. E. Tanner, M. L. Wahlberg, T. Grandin, and R. M. Lewis. 2019a. Subjective methods to quantify temperament in beef cattle are insensitive to the number and biases of observers. *Appl. Anim. Behav. Sci.* 212:30–35. doi:10.1016/j.applanim.2019.01.005

Parham, J. T., A. E. Tanner, K. Barkely, L. Pullen, M. L. Wahlberg, W. S. Swecker, Jr., and R. M. Lewis. 2019b. Temperamental cattle acclimate more substantially to repeated handling. *Appl. Anim. Behav. Sci.* 212:36–43. doi:10.1016/j.applanim.2019.01.001

Parham, J. T., A. E. Tanner, M. L. Wahlberg, W. S. Swecker, Jr., and R. M. Lewis. 2021. Subjective methods of quantifying temperament in heifers are indicative of physiological stress. *Appl. Anim. Behav. Sci.* 234:105197. doi:10.1016/j.applanim.2020.105197

R Core Team. 2013. *R: a language and environment for statistical computing*. Vienna (Austria): R Foundation for statistical Computing. http://www.R-project.org/.

Sporer, K. R. B., P. S. D. Weber, J. L. Burton, B. Earley, and M. A. Crowe. 2008. Transportation of young beef bulls alters circulating physiological parameters that may be effective biomarkers of stress. *J. Anim. Sci.* 86:1325–1334. doi:10.2527/jas.2007-0762

Tulloh, N. M. 1961. Behavior of cattle in yards. II. A study of temperament. *Anim. Behav.* 9:25–30. doi:10.1016/0003-3472(61)90046-X

Turner, S. P., E. A. Navajas, J. J. Jyslop, D. W. Ross, R. I. Richardson, N. Prieto, M. Bell, M. C. Jack, and R. Roehe. 2011 Associations between response to handling and growth and meat quality in frequently handled *Bos taurus* beef cattle. *J. Anim. Sci.* 89:4329–4248. doi:10.2527/jas.2010.3790

Van de Water, G., F. Verjans, and R. Geers. 2003. The effect of short distance transport under commercial conditions on the physiology of slaughter calves; pH and colour profiles of veal. *Livest. Prod. Sci.* 82:171–179. doi:10.1016/S0301-6226(03)00010-1

Vargas Jurado, N., A. E. Tanner, S. R. Blevins, J. Rich, R. W. Mayes, D. Fiske, W. S. Swecker, Jr., and R. M. Lewis. 2015. Feed intake and diet selection in angus-cross heifers of two frame sizes at two stages of growth. *J. Anim. Sci.* 93:1565–1572. doi:10.25274/jas.2014-8453

Vetters, M. D. D., T. E. Engle, J. K. Ahola, and T. Grandin. 2013. Comparison of flight speed and exit score as measurements of temperament in beef cattle. *J. Anim. Sci.* 91:374–381. doi:10.2527/jsa.2012-5122