Bull sperm morphology assessment varied by evaluator

Ashley Reeves,[a] Jessica Klabnik,[c] Lew Strickland,[b,c] Tulio Prado,[b,c]
Pablo Jarrin-Yepez,[b] Cheyenne Dingemans,[d] Liesel Schneider,[c] Brian Whitlock[b,c]
[a]Comparative and Experimental Medicine, [b]Department of Large Animal Clinical Sciences
[c]Department of Animal Science, University of Tennessee, Knoxville, TN
[d]Department of Field Services and Theriogenology, University of Georgia, Athens, GA

Abstract

Assessment of sperm morphology is an important part of bull breeding soundness evaluation (BBSE). Whereas the effects of evaluator experience and evaluation method on the sperm morphology estimation were assessed in other species, no such study was conducted for bulls. Our objectives were to assess the effects of evaluator experience and number of sperm assessed on BBSE outcomes. A single eosin-nigrosin sperm morphology slide from individual semen samples, collected from 35 yearling bulls was used. In Experiment I, 6 individuals (3 board-certified theriogenologists [DACT] and 3 fourth-year veterinary students [VS]) evaluated 100 sperm from 35 slides twice (at least 1 week between evaluations). In Experiment 2, 3 DACT evaluated 100, 200, and 400 sperm from the same 5 sperm morphology slides to determine if assessing a higher number of sperm would increase the agreement of morphologic characteristics. In Experiment 1, there was a difference ($p < 0.0001$) in the percent of sperm classified as morphologically normal between evaluator types (VS versus DACT). Furthermore, variation among evaluators affected sperm morphology assessments and bull breeding soundness evaluation classifications. Whereas the time needed to evaluate slides increased ($p = 0.96$) with increasing number of sperm assessed, there was no effect (Experiment 2) of number of sperm evaluated on percent normal sperm, indicating that evaluating more than 100 sperm may not be justifiable. Further investigation on slide preparation, microscope use, assessor experience, and continuing education/training is important to ensure the repeatability and validity of evaluating bovine sperm morphology.

Keywords: Bull, sperm, morphology, assessment

Introduction

Reproduction is economically important to production animal species and removing subfertile individuals is of high priority as 1 bull may breed many cows. The breeding potential of an individual bull can be determined by performing a bull breeding soundness evaluation (BBSE). It consist of a brief physical examination, assessment of reproductive organs, transrectal palpation of the accessory sex glands, collection of an ejaculate, and evaluation of sperm motility and morphology.[1,2] To be classified as a 'satisfactory potential breeder', a bull should be physically and reproductively sound and at least 70% of sperm evaluated must be morphologically normal.[2] Abnormal sperm morphology is the most common reason that bulls are classified as a category other than 'satisfactory' (> 75% of bulls were classified as 'unsatisfactory' due to morphology alone or in combination with other measurements[1]). Normal morphology is important for sperm motility within the reproductive tract and ability to fertilize oocytes.[3] Furthermore, the assessment of agreement among veterinarians performing BBSE becomes beneficial to develop future training standards.

Stallion and man sperm analysis studies[4-6] determined the effects of evaluator experience, sperm morphology slide preparation methods, and methodology by which sperm morphology slides were evaluated on the outcome of sperm morphology classification. Values reported in a subsequent human sperm morphology study[7] (by experienced laboratory technicians, biologists, and physicians) were comparable to reference values of andrology laboratories that had highly trained staff members. Authors concluded that following recommended methods coupled with experience are critical to achieve relevant and comparable results.[6]

Although substantial differences among evaluators and assessment methods were observed in stallion and man sperm morphology studies, to date, there have been no studies evaluating effects of these variables on the classifications of bull sperm morphology. Therefore, the objectives of this study were to compare: intra-evaluator variation for bull sperm morphology assessment; variation among veterinary students in their fourth year (clinical year) with strong interest in theriogenology; variation among theriogenologists; agreement between theriogenologists and fourth year veterinary students; effect of time to complete the analysis of bull sperm morphology slides on evaluator type and cell number assessed; effect of number of sperm assessed per morphological slide evaluation on the percent normal morphology.

Materials and methods

Animals and slide preparation

Thirty-five semen samples from 35 consigned Black Angus bulls (14 - 18 months of age; 1 sample per bull) were used in this study. Bulls were housed at the University of Tennessee bull test station (Spring Hill, TN). All bulls received ad libitum access to pelleted feed, hay, and water. Semen was collected (as 'convenience samples'; breeding soundness evaluation was conducted on each bull for reasons unrelated to this study) by electroejaculation using a Lane Pulsator V (Lane Manufacturing, Denver, CO). A single eosin-nigrosin (Morphology Stain, Society for Theriogenology, Mathews, AL) stained semen smear slide was prepared from the semen of each bull by 1 theriogenologist. After a cursory view of each slide to ensure each slide had an acceptable concentration and distribution of appropriately stained sperm, slides were coded (1 - 35) to eliminate any identification of the bull and date of collection.

Evaluators and experimental design

Evaluators reviewed bull sperm morphology classification criteria defined by the manual for breeding soundness examination of bulls (2018 edition).[2] Objective of this initiative was to ensure that evaluators were using the same criteria to classify sperm morphology to minimize the variation between and among individuals. Sperm morphology was evaluated under 1,000 x (oil immersion; Cargile nondrying immersion oil for microscopy, Type B code 1248, lot 110193) magnification and bright field microscopy using the same microscope (OMAX Lab Biological Binocular Compound Microscope 40 - 2,000 x W Halogen Light) in the same room.

Experiment 1

Six individuals evaluated all 35 slides. Three were diplomates of American College of Theriogenologists (DACT) and 3 were fourth-year veterinary students (VS) who completed an advanced veterinary reproductive elective and self-identified as interested in theriogenology. For each slide, 100 sperm were evaluated and classified according to the 2018 Society for Theriogenology, Manual for Breeding Soundness Examination of Bulls.[2] Each individual evaluated all 35 slides in the same order (1 - 35) in 1 'timed sitting'. After 1 week, second assessment of all slides was completed by each individual.

Experiment 2

Three DACT evaluated 5 morphology slides using similar methods described before. Sperm (100, 200, and 400) were evaluated in 1 'timed sitting' for each slide twice with at least 1 week between evaluations.

Data analyses

To calculate the intraclass correlation coefficient (ICC) for each evaluator, counts of normal sperm for each of the 35 slides were analyzed within the ICC_SAS macro.[8] The 2-way mixed effects ICC calculation was used to assess the consistency of 1 evaluator among multiple measures of the same slide.[9] The agreement among DACT of grading a slide as pass or fail was analyzed using Cohen's kappa agreement. Data on percent normal sperm were analyzed using separate mixed model ANOVA (Proc GLIMMIX, SAS 9.4); fixed effects of evaluator, evaluator type, slide number (to determine effect of order in which the slides were evaluated) or number of sperm assessed, and interactions (as appropriate) were tested. Simple linear regression (PROC REG) was utilized to assess the effect of the number of sperm read on the time required to complete the assessment, and separately, the effect of the number of sperm evaluated on the time to complete the evaluation of 5 slides.

Data from Experiment 2 were analyzed using logistic regression (PROC GLIMMIX) with a binary distribution. Effect of total number of sperm assessed impacted the probability for a bull to fail the morphology assessment was tested. Statistical significance was set at $\alpha = 0.05$.

Results

Experiment 1

Intraclass correlation coefficients were calculated for each evaluator based on 2 readings of normal sperm for 35 slides (Table 1). Each evaluator had varying degrees of consistency ($p < 0.001$; Table 1). There was an interaction ($p < 0.0001$) of evaluator type (DACT versus VS) and slide number on the percentage or proportion of sperm classified as morphologically normal, indicating that the effect of evaluator type depended on the slides that were evaluated. Among slides that differed between reviewer type, VS had a higher ($p = 0.0001$) percent normal sperm recorded compared to DACT (Table 2). Mean percent normal sperm recorded for VS was numerically higher for most slides, changing the outcome of a potential BBSE from 'deferred' or 'unsatisfactory' to 'satisfactory' for 7 of the 35 (20%) slides. Furthermore, there was an effect

($p < 0.0001$) of evaluator on coefficient of variation among individuals, but no effect ($p = 0.78$) of evaluator type. The agreements among DACT when examining 35 of the same morphology slides, differed among individuals (Table 3). Number of bulls that were classified as 'unsatisfactory'/'deferred' based on morphological assessment differed numerically among evaluators. There was no effect ($p = 0.53$) of evaluator type on the time needed to complete evaluations of slides; the average time to evaluate 100 sperm for 35 slides was 134.6 minutes for DACT and 145.5 minutes for VS with a SEM of 11.43 minutes. However, there was an effect of individual within evaluator type ($p = 0.04$; Figure 1), such that some evaluators took 30 minutes longer to complete their assessments compared to the fastest evaluators within an evaluator type.

Table 1. Intraclass correlation coefficients (ICC) for each evaluator (35 slides in 2 separate duplicate sessions)

Evaluator	ICC	95% CI
$DACT_1$	0.94	0.89, 0.97
$DACT_2$	0.91	0.83, 0.95
$DACT_3$	0.95	0.90, 0.97
VS_1	0.95	0.90, 0.97
VS_2	0.75	0.56, 0.86
VS_3	0.86	0.74, 0.93

Table 2. Mean percent of morphologically normal sperm as determined by theriogenologists (DACT) and veterinary students (VS) in Experiment 1. Grey highlighted numbers represent slides for which the DACT and VS results would change the BSE outcome from 'unsatisfactory' to 'satisfactory'. Asterisks (*) denote significant differences between DACT and VS.

Slide Number	DACT	VS	SEM
1	41.33	41.50	5.52
2	66.83	69.33	5.52
3	62.16	72.00	5.52
4	38.16	48.00	5.52
5	82.00	91.16	5.52
6	78.66	85.33	5.52
7	67.83	68.83	5.52
8	80.00	73.16	5.52
9	39.66	46.66	5.52
10	67.00	81.83	5.52
11	76.00	77.16	5.52
12	72.33	80.16	5.52
13	84.16	80.83	5.52
14*	26.16*	45.66*	5.52
15	62.66	74.00	5.52
16	78.83	81.00	5.52
17	82.83	84.50	5.52
18	81.00	88.33	5.52
19	74.33	81.16	5.52
20	78.33	79.33	5.52
21	64.50	71.00	5.52
22	82.83	78.66	5.52
23	66.33	72.50	5.52
24	81.00	81.33	5.52
25	78.66	78.33	5.52
26	61.16	75.66	5.52
27	86.66	83.50	5.52
28	72.66	83.83	5.52
29	67.50	68.83	5.52
30*	43.16*	62.16*	5.52
31	4.33	10.33	5.52
32	73.00	78.83	5.52
33*	58.33*	83.16*	5.52
34*	8.00*	30.83*	5.52
35	58.33	65.50	5.52

Table 3. Agreement among theriogenologists (DACT) using the Fleiss agreement guidelines with multiple evaluators examining the same slides.

Evaluator 1	Evalutor 2	Kappa Coefficient	Agreement
$DACT_1$	$DACT_2$	0.3158	Poor
$DACT_2$	$DACT_3$	0.4944	Fair
$DACT_1$	$DACT_3$	0.6441	Good

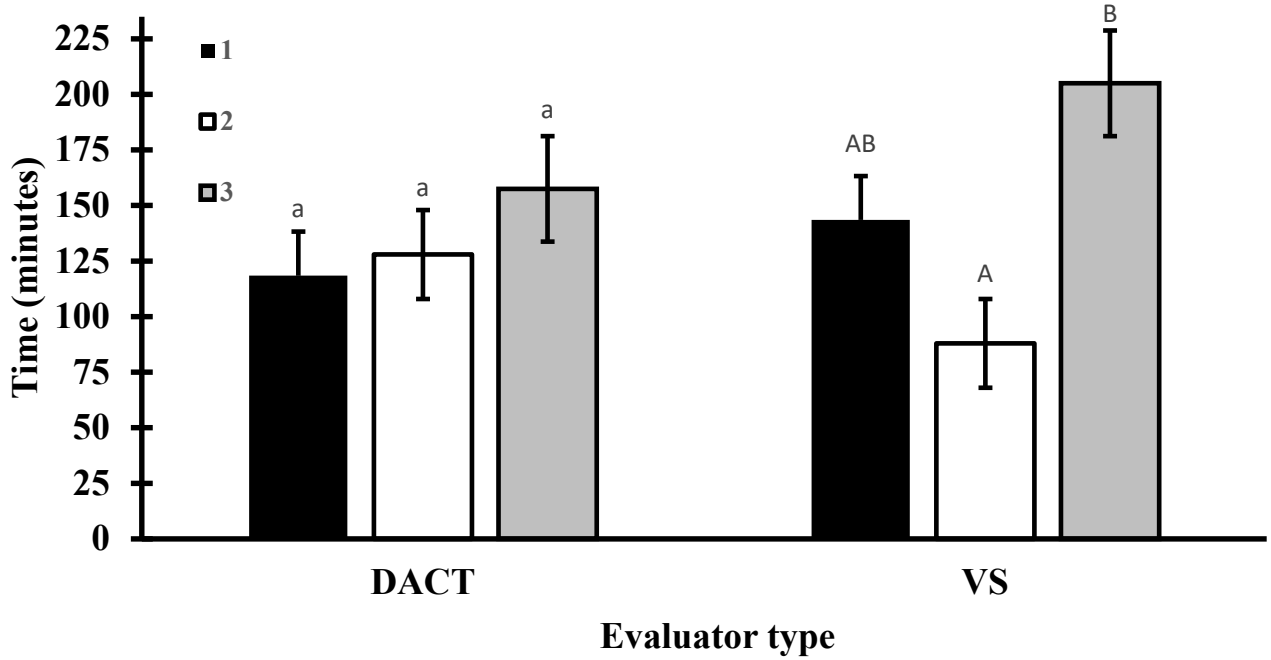

Figure 1. Time (in minutes) to evaluate 100 sperm for 35 slides consecutively per individual. Least square means and standard error bars are represented. Each colored bar denotes individual evaluator for each evaluator type. Lower case letters above bars denote differences ($p < 0.05$) among DACT evaluators; capital letters denote differences among VS evaluators; and similar letters within evaluator type did not differ.

Experiment 2

For DACT evaluators, there was no effect ($p = 0.96$) of increasing the number of sperm evaluated on percent of sperm classified as morphologically normal, and the total number of sperm assessed did not impact the probability of that bull to fail the morphology assessment ($p = 0.95$), indicating that evaluating > 100 sperm is unlikely to have an effect on the BBSE classification for a particular bull (Figure 2). There was an effect of number of sperm assessed on time needed to complete the evaluation ($p = < 0.0001$) such that the variation in the number of sperm read explained 73% of the variation in time taken for the assessments ($R^2 = 0.73$). Additionally, there was an interaction ($p < 0.0001$) between evaluator and number of sperm assessed on time to completion; for some evaluators, as the total number of sperm evaluated per slide increased, time taken for completion also increased (Figure 3).

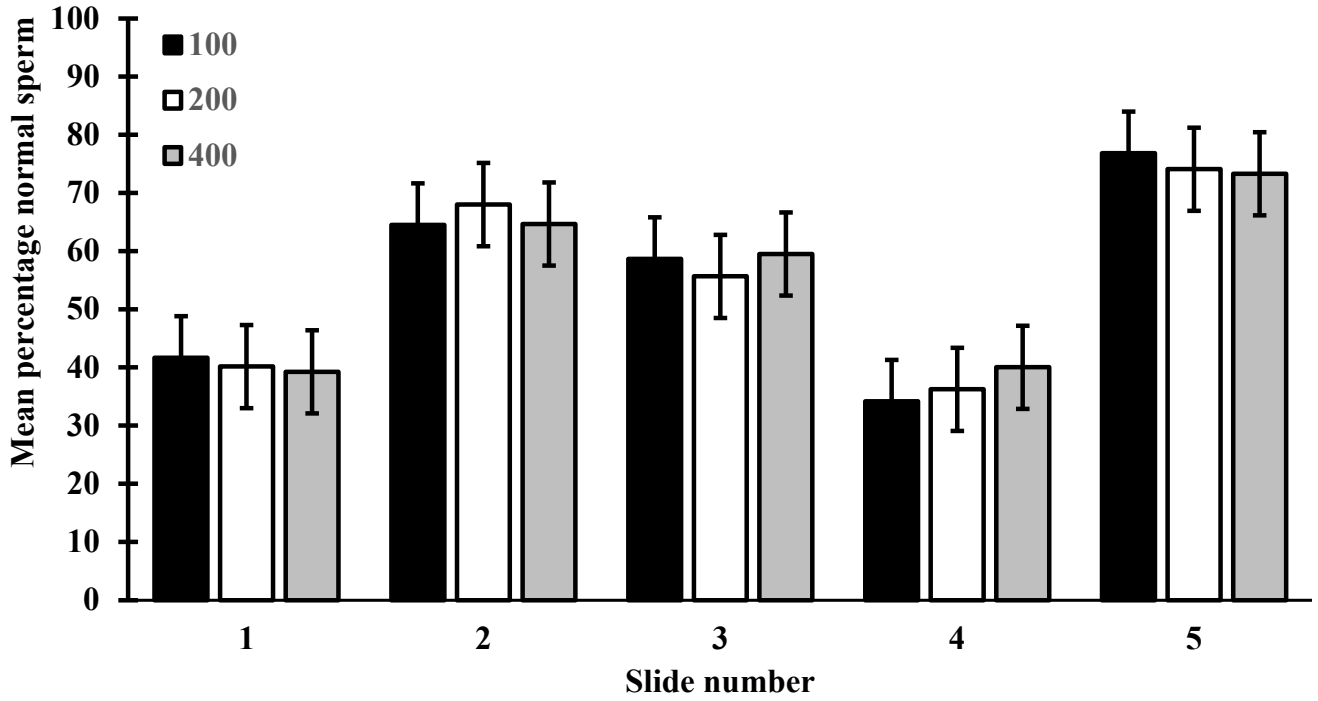

Figure 2. Percentage normal sperm by increasing cell count: performed by DACT. Least square means and standard error bars are represented. Note: no change in 'satisfactory' versus 'unsatisfactory' outcome.

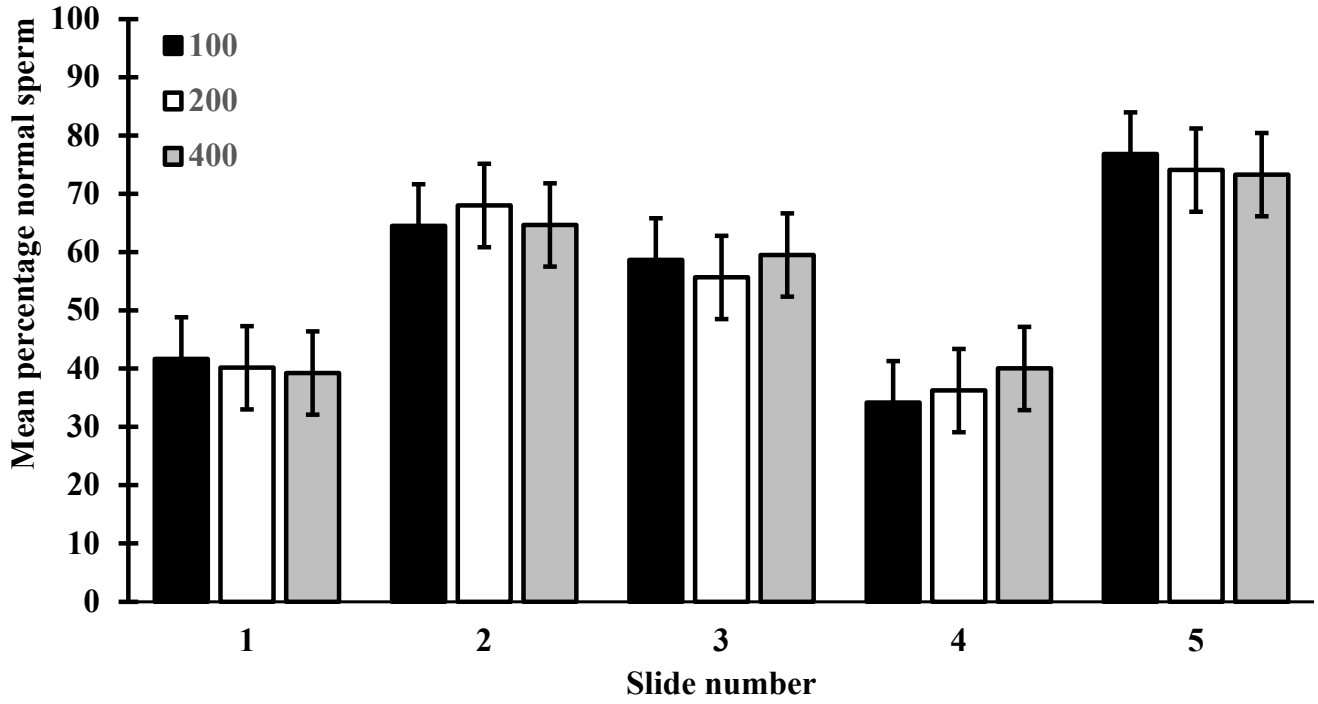

Figure 3. Time (in minutes) taken for DACT to assess 100, 200, and 400 sperm in 5 morphology slides. Least square means and standard error bars are represented. Similar letters did not differ ($p < 0.05$).

Discussion

Effect of evaluator experience on the morphologic classification of bull sperm was determined for the first time. Increasing the number of sperm evaluated per slide during a BBSE, from 100 (the current accepted standard) up to 400, is unlikely to alter the percent of sperm classified as morphologically normal and therefore unlikely to affect the outcome of a BBSE. Therefore, for bulls presented for a BBSE that are just below the acceptable percent of normal sperm morphology for a 'satisfactory' classification, continuing to assess more sperm is unlikely to yield a different outcome.

Although standard cutoffs for adequate normal sperm morphology have been used for decades, most bulls not classified as 'satisfactory potential breeders' are a result of too few morphologically normal sperm[10-13] and results reported here indicate that variation among evaluators may affect the proportion of sperm classified as morphologically normal and ultimately the outcome of the BBSE. An interesting finding was the effect of individual evaluator on intra-evaluator variation of sperm morphology, indicating that an evaluator is not necessarily consistent with their own classification criteria. This should be considered when performing multiple breeding soundness examinations in 1 sitting. This effect was observed when the evaluator assessed 35 slides in 1 sitting, but would it occur for an occasional BBSE not at a high-volume testing station or veterinary hospital?

Another interesting finding was that for some slides there were differences between DACT and VS in the percent normal sperm, and in this case, VS always reported a higher percentage of sperm as morphologically normal. We concluded that occasionally VS are less discriminating than DACT in morphological evaluation. Other studies involving the assessment of stallion sperm morphology reported variation among veterinarians for all morphology classification categories.[4] Similarly, a second investigation attributed the differences of sperm abnormalities to the education level of the evaluator, years of experience evaluating sperm morphology slides, and the differences among evaluators on classifications of abnormalities.[6] Our results indicate a variation among evaluators for morphology classifications (normal versus abnormal). However, the classification of specific morphological abnormalities was not recorded in our experiments to further investigate where those differences may lie.

Furthermore, our results indicate a difference between evaluator types (DACT versus VS) that could be due to a difference in quality or quantity of training or experience. The individuals within the DACT evaluators had received more extensive/exhaustive training in the assessment of sperm morphology of multiple species and they had performed more BBSE than VS evaluators, meaning that they had more experience evaluating bull spermiograms. Whereas the DACT evaluators differed from the VS evaluators, agreement among DACT evaluators was not consistent suggesting variation within the same evaluator type should be considered. The studies conducted in man sperm morphology assessment included laboratory technicians, physicians, and biologists.[5,7] Their results indicate that participants who do not routinely use the recommended procedures were less strict and gave varied results compared to participants who followed the guidelines and reference values established by andrologists.[5] Their various training methods/guidelines included attending training sessions in semen analysis at a central standardizing/validating laboratory with proficiency assessment biannually using blind coded sperm suspensions and videotapes, following criteria set forth in the World Health Organization laboratory manual for the examination of man semen and/or standardization workshops on semen morphology assessment prior to the study.[7,14,15] This study demonstrated the imperative need for standardization, training, and quality control for man sperm morphology assessment.[5] Similarly, a manual is available for the recommended criteria for BBSE;[2] however, recommended training and quality controls have not been established. This information should encourage discussions directed toward improvement of continuing education opportunities and quality controls for veterinarians to remain current in their sperm morphology examination skills.

The differences between the DACT and VS evaluators on percent of normal sperm would have changed the classification of the BBSE ('satisfactory,' 'deferred,' or 'unsatisfactory') for multiple evaluators (7 out of 35 evaluators (20%) associated with the various sperm morphology slides. For example, using only sperm morphology as a criterion, the bull associated with microscope slide 10 would have been classified as 'satisfactory' by the VS evaluators (81% of sperm were classified as morphologically normal) but 'unsatisfactory' or deferred by the DACT evaluators (67% of sperm were classified as morphologically normal). Since current standards were determined from a study by an unknown number of evaluators with undescribed experience, the true fertility status of 'borderline' bull in a natural service scenario may or may not be congruent with their classification.[16] Further investigation into the natural breeding success of the bulls in this study can determine if the differences in the evaluations of the DACT and VS had clinically relevant effects on fertility. Although studies have reported an effect of evaluators and their experience on sperm morphology evaluation, those studies did not determine evaluators with minimal experience or training in morphology assessment for a comparison. Further investigation would include more evaluators of varying experience and training with a larger sample size of evaluators to confirm if this difference is repeatable.

For decades, the standard number of sperm assessed for morphology examination has been 100. This was most likely chosen for ease of calculation conversion to percent, but to our knowledge there are no reported data to validate this as the ideal number of sperm to be evaluated in bulls or other species. It is logical to expect that assessing a higher number of sperm would mean the estimated sperm morphology would more accurately represent the true sperm morphology. Therefore, evaluating a higher number of sperm during a morphological assessment would more accurately estimate the true percentage of normal sperm, thereby improving the accuracy of BBSE classifications ('satisfactory', 'unsatisfactory', or 'deferred') since sperm morphology has the single greatest effect on the BBSE classification.[1] However, results reported here indicate that increasing the number of sperm evaluated (up to 400), did not improve or change the result of classifying a bull as 'satisfactory based on morphology assessment. Based

on results reported, additional time needed to evaluate more than 100 sperm during a routine BBSE would be difficult to justify, in part, because evaluating additional sperm did not result in higher agreement among assessments. Further investigation with increased slide numbers would be beneficial to validate our findings.

The final suggestions and results of the previously mentioned studies were centered on investments in training, continuing education, proficiency testing, and other quality control measures to minimize this variation in evaluation.[4,5] Another aspect of the BBSE that may affect the ability of the evaluator to assess sperm morphology is the equipment used to complete the evaluation. The same microscope, room, and oil were used for this study to help eliminate bias and most veterinarians will use a traditional compound light microscope similar to the one used here. However, if a veterinarian performs a high volume of BBSE, they may invest in a microscope with fluorescent capabilities, phase contrast, or the ability to increase magnification. A fluorescent stain with subsequent microscopy would allow evaluation of sperm membrane integrity, acrosome status, DNA integrity and more.[17] Additionally, using phase contrast allows the sperm to appear on a darker background to better highlight the sperm and eliminates the need for staining. Increasing magnification would increase the resolution of individual sperm allowing for more precise evaluation. These alternate microscopic techniques could prove useful for breeding programs in further assessing sperm function. In this study, the evaluators utilized the same room for the entirety of the investigation, whereas in practice, facilities may vary from indoor to outdoor, lighting, and temperatures possibly affecting the results.

Sample preparation methods and setting recommendations must be strictly followed to obtain credible results and the species, extender, hardware/software, and operator can all affect the output values.[18,19] Although CASA was not used, visual sperm evaluation in men sperm is comparable to CASA to parameters assessed (including morphology).[18] Despite our smaller sample size, findings emphasized the importance of further investigation into slide preparation, equipment used, evaluator experience, and continued training and practice.

Conflict of Interest

None to declare.

References

1. Carson RL, Wenzel JGW: Observations using the new bull-breeding soundness evaluation forms in adult and young bulls. Vet Clin North Am Food Anim Pract 1997;13:305-311.

2. Koziol JH, Armstrong CL: Manual for Breeding Soundness Examination of bulls, 2nd edition, Mathews, Society for Theriogenology, 2018.

3. Barth AD, Oko RJ: Abnormal Morphology of Bovine Spermatozoa. Iowa State University Press. 1st edition, Ames, Iowa 1989:3-7;19-89.

4. Brito LFC, Greene LM, Kelleman A, et al: Effect of method and clinician on stallion sperm morphology evaluation. Theriogenology 2011;76:745-750.

5. Eustache F, Auger J: Interindividual variability in the morphological assessment of human sperm: effect of the level of experience and the use of standard methods. Hum Reprod 2003;18:1018-1022.

6. Murcia-Robayo RY, Jouanisson E, Beauchamp G, et al: Effects of staining method and clinician experience on the evaluation of stallion sperm morphology. Anim Reprod Sci 2018;188:165-169.

7. Guzick DS, Overstreet JW, Factor-Litvak P, et al: Sperm morphology, motility, and Concentration in fertile and infertile men. N Engl J Med 2001;345:1388-1393.

8. Lu L: Reliability analysis: calculate and compare intra-class correlation coefficients (ICC) in SAS. Nesug Statistics and Data Analysis 2007;1-4.
https://lexjansen.com/nesug/nesug07/sa/sa13.pdf.

9. Koo WK and MY Li: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15:155-163.

10. Monday JD, Larson RL, Theurer ME, et al: Factors associated with yearling bulls passing subsequent breeding soundness evaluations after failing an initial evaluation. J Am Vet Med Assoc 2018;253:1617-1622.

11. Conkey N, Okafor C, Strickland L, et al: Factors associated with bulls not classified as satisfactory potential breeders. Clinical Theriogenology 2019;11:447.

12. Bruner KA, McCraw RL, Whitacre MD, et al: Breeding soundness examination of 1,952 yearling beef bulls in North Carolina. Theriogenology 1995;44:129-145.

13. Roberts JN, Grooms DL, Thompson ER, et al: Evaluation of bull breeding soundness examination. Clinical Theriogenology 2018;10:397-408.

14. Kruger TF, Menkveld R, Stander FS, et al: Sperm morphologic features as a prognostic factor in in vitro fertilization. Fertil Steril 1986;46:1118-1123.

15. WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction. 4th edition, Cambridge, England: Cambridge University Press 1999.

16. Wiltbank JN, Parish NR: Pregnancy rate in cows and heifers bred to bulls selected for semen quality. Theriogenology 1986;25:779-783.

17. Farah OI, Cuiling L, Jiaojiao W, et al: Use of fluorescent dyes for readily recognizing sperm damage. J Reprod Infertil 2013;14:120-125.

18. Talarczyk-Desole J, Berger A, Taszarek-Hauke G, et al: Manual vs. computer-assisted sperm analysis: can CASA replace manual assessment of human sperm in clinical practice? Ginekologia Polska 2017;88:56-60.

19. Amann RP, Waberski W: Computer-assisted sperm analysis (CASA): Capabilities and potential developments. Theriogenology 2014;81:5-17.