
An Article Submitted to

*Journal of Quantitative Analysis in
Sports*

Manuscript 1251

Scoring Variables and Judge Bias in
United States Dressage Competitions

Ana E. Diaz* Mary S. Johnston[†] Jennifer Lucitti[‡]
Wendi S. Neckameyer** Katy M. Moran^{††}

*Dupont, anadiaz@alumni.duke.edu

[†]marysjohnston@verizon.net

[‡]University of North Carolina-Chapel Hill, jennifer.lucitti@med.unc.edu

**Saint Louis University School of Medicine, neckamws@slu.edu

^{††}moranequinephoto@verizon.net

Scoring Variables and Judge Bias in United States Dressage Competitions*

Ana E. Diaz, Mary S. Johnston, Jennifer Lucitti, Wendi S. Neckameyer, and Katy M. Moran

Abstract

Dressage is an Olympic equestrian discipline with rules governed by The Fédération Equestre Internationale (FEI). After questions arose regarding dressage scoring during the 2008 Olympics, the FEI undertook studies of dressage scoring at international equestrian competitions in Europe. These studies included investigations of scoring scales and judging bias. Based on those results, the FEI recommended the use of statistics to improve the consistency of judge training in dressage competition. The authors performed a statistical analysis of dressage scores earned by riders in the United States. This work demonstrates the lack of consistency in judging in the United States. This is the first published study that looks at U.S. dressage results. It includes a very large data set considered representative of the population of riders in the U.S. Dressage is a subjectively judged sport. The final score, not the class placing, is of greatest importance when assessing rider results. The authors show how mainstream statistical tools can identify factors impacting the final dressage score. Analysis of Variance (ANOVA) was used to demonstrate the existence of key variables which impact scores. The analysis shows there are variations in scoring patterns by horse breed, as well as the existence of high-scoring and low-scoring judges. Finally, the analysis shows that there are increased standard deviations at the “extremes” of the judging scale. This would indicate there is variability in interpretation of scoring at the upper and lower ends of the scoring scale. Analysis of Means (ANOM) identified five judges whose scores were statistically different than the grand average scores.

KEYWORDS: equestrian, ANOVA, ANOM, dressage, Fédération Equestre Internationale

*The authors are scientists and amateur dressage riders interested in the use of statistics to improve dressage judging. Ana E. Diaz, PE is a registered Professional Engineer and Six Sigma Master Black Belt in Dupont’s Corporate Operations. Mary Stydnicki Johnston is retired from federal service where she led an Operations Research, Modeling & Simulation Office. Jennifer Lucitti, PhD is a Research Associate in the department of Cell and Molecular Physiology in the School of Medicine at University of North Carolina, Chapel Hill. Katy Moran, PhD has been self-employed as a freelance photographer, writer, and graphic designer since 1992. Wendi Neckameyer, PhD is a full professor in the Department of Pharmacological and Physiological Science at Saint Louis University School of Medicine.

1. Introduction

The sport of dressage is an Olympic equestrian discipline where a horse and rider are judged on the performance of specific movements which require high levels of athleticism, skill, and communication between the pair. The movements and patterns have their roots in the techniques of the mounted cavalry and equestrian traditions dating back for centuries. Modern dressage competition is an international sport that has been part of the Olympics since 1912.

Dressage is similar to the compulsory and freestyle routines in figure skating and gymnastic competition. Each compulsory test requires the execution of specified movements in a specified order. Each movement is compared against an accepted standard for that movement and is assigned a score from zero (movement not executed) to ten (excellent) by the officiating judge or judging panel. Competitions are organized on both national and international levels. International competitions are judged by five judges distributed throughout the competition court. U.S. national competitions are allowed to be held with one judge presiding.

General scores are given for the quality of the horse's gaits, the horse's desire to cooperate, the horse's desire to travel energetically, and the rider's performance during the test. After the test, scores are summed and converted to a percentage of the total number of points available for that test. The horse-rider pair with the highest percentage is the class winner.

While riders compete against one another for class ranking, they ultimately compete against a standard to earn a final score. In the U.S., the final score, not the class ranking, dictates eligibility to compete in regional championships, to earn year-end achievement and breed awards, and to accumulate life-long rider qualification awards. A first place won with a score of 56% does not have the same value as a first place won with a score of 66%.

Eligibility and achievement of awards are important for several reasons. First, the monetary value of a horse as a riding horse and as breeding stock can be directly correlated to scores, achievements, and awards. Progeny are also evaluated by the record of the parents. Depending on their competitive success, horses can be valued into the seven figures. A recent article in the British newspaper *The Telegraph* suggests that the stallion, international dressage champion, Moorlands Totilas may now be the most valuable competition horse of all time, with an estimated value of up to €25 million (Cuckson 2009).

Second, the business reputation of the professional trainer and rider can be linked to successful scores in the show ring. Lastly, amateur riders spend considerable money in purchasing, housing, training and showing horses with usually little or no financial gain. Thus, the numerical value of the final score is more important to the sport than the final class ranking.

The Fédération Equestre Internationale (FEI) is the international body governing competitions in equestrian sports. The FEI represents all equestrian sports to the International Olympic Committee (IOC) and is the coordinating body for all national federations.

The FEI is responsible for the design and administration of dressage tests used in international competition. The International level tests are the highest level of difficulty. These tests are the same worldwide. There are four levels: Prix St. Georges, Intermediare I & II and Grand Prix.

The national tests are designed by each nation's governing body. In the United States, the national governing body is the United States Equestrian Federation (USEF). This is a federation of associations representing different horse breeds and equestrian sports. The USEF represents the United States in the FEI. The United States Dressage Federation (USDF) represents the discipline of dressage to the USEF. The USEF devises the dressage tests used in the United States in consultation with the USDF. These dressage tests range in increasing length and difficulty from Training Level to Fourth Level. In total, there are five levels with three to four tests of increasing difficulty at each level.

Scoring policies are scrutinized by dressage sport professionals and fans alike. In 2008, scoring variability at the Beijing Olympics sparked considerable debate within the FEI and among dressage fans around the world. The President of the FEI dismissed the official Dressage Committee (Haya 2008). The FEI then appointed a formal Dressage Task Force to study scoring methodologies, training of judges and to review other programs at the FEI (FEI 2009b). These issues included an investigation of “fitness for purpose of the method of judging dressage competitions, including the judging process” (FEI 2009a). The FEI recognized that scores given by qualified dressage judges should be examined for standardization and consistency.

In the United States, riders must qualify to participate in international competition. However, currently there is no qualifying process for riding at progressively higher National levels. At about the same time that questions arose about Olympic dressage scoring, the USEF proposed a rule that would have required riders to qualify in order to ride at each of the national levels. The proposed rule required earning a certain number of points based on scores at each level. The proposed qualifying rule was ultimately withdrawn in large part due to strong objection from USEF and USDF membership.

Given the importance of the final score in the sport of dressage at the national, international and Olympic levels, it is surprising that so very little formal attention has been given to the analysis of consistency in dressage scoring. A search of the literature reveals that there have been three peer-reviewed statistical analyses done related to dressage and dressage scoring. Whittaker (2005a and 2005b) and Deuel (1995) focused on the impact of the dressage segment in British

eventing competitions. The sport of eventing is a multi-phase equestrian sport which includes a dressage test along with cross-country and jumping tests. Different type horses and training programs are required for this sport. No analyses of scoring solely in dressage competitions were found in the peer-reviewed literature.

There have been a few “informal” studies written up on dressage sport enthusiast websites (PVDA, Stickland 2009a, and Stickland 2009b). This paper is the first to present a rigorous, quantitative study of judging in United States dressage competitions. The authors show how mainstream statistical tools can be applied to identify factors impacting the final dressage score. The work also illustrates the application of industrial quality management techniques which can be used to improve any subjectively-judged sport.

Because of the importance of the final score, as opposed to a simple class ranking, it is imperative that dressage judges are judging to the same standard and award scores that are consistent with other judges. To date, an analysis of U.S. inter- and intra-judging trends has not been performed.

This paper is important because it is the first comprehensive analysis of dressage scores earned in the United States. It is based on 2009 Open Dressage competitions and is the first paper to include the lower national levels along with international levels. Finally, this paper encompasses the largest data set analyzed to date by any investigator.

Since dressage is a subjectively judged sport, our purpose is to identify key factors which influence the final score. Specifically this paper explores whether there is possible judge bias. Analysis of variance (ANOVA) and analysis of means (ANOM) were used to identify variables and to test whether judging is similar at each show and for each region in the United States. We explore the theory that some judges score statistically higher or lower than their peers. We present novel findings that can improve judging and at all levels.

This paper is the first to rigorously examine factors contributing to dressage scores, with specific attention to judge bias throughout the United States. The tools used can be applied to all subjectively judged sports. The results present novel findings that can improve judging and competition participation in the sport at all levels.

2. Methods

The software package, Minitab[®]-15 (Minitab, Inc) was used for all statistical analyses. Descriptive statistics were run on all the data. All score data sets were normally distributed. The principal statistical tests used to detect differences between factors were analysis of variance (ANOVA) and analysis of means (ANOM). The data were acquired from publicly available sources.

ANOVA is a statistical tool to test for differences in the means. ANOVA was used to test for differences in scores when grouped by judge, breed, region or level. ANOM is a graphical analog to ANOVA, which tests the equality of population means and displays the confidence intervals for the means. Minitab[®] displays a graph showing how the mean for each factor compares to the overall mean (also called the grand mean). ANOM can identify when the level means differ and what those differences are. As with analysis of variance, ANOM can be used if the response approximately follows a normal distribution. All data in this study met this criterion.

U.S. Dressage scores are available on the USDF website (<http://www.usdf.org>). Scores from all shows held between March - August 2009, were used for this analysis. The posted scores were retrieved between September 28 and October 6, 2009. The scores are posted in the format of Show Name, Show Date, Class Level, Test Number, Composite Score, Rider, Horse, Breed, Owner, Judge(s). Where multiple judges officiated, only the final composite score is reported. In the analysis for variability by judge, a data subset was used. This data included only scores provided by one judge.

The scoring scale for each movement in a dressage test is an integer on an ordinal scale defined by: 0 – Not Executed; 1 – Very Bad; 2 – Bad; 3 – Fairly Bad; 4 – Insufficient; 5 – Marginal; 6 – Satisfactory; 7 – Fairly Good; 8 – Good; 9 – Very Good; 10 – Excellent. The scores of some particularly difficult and/or essential movements are weighted by coefficients to illustrate their importance. The same scoring scale is applied to the Collective Marks at the end of the test. These marks assess the horse's gaits, impulsion, submission, and the rider's position and seat. They are usually given a coefficient of 2 or 3 to emphasize the importance of the rider's skill, the horse's training, and the horse's natural talent.

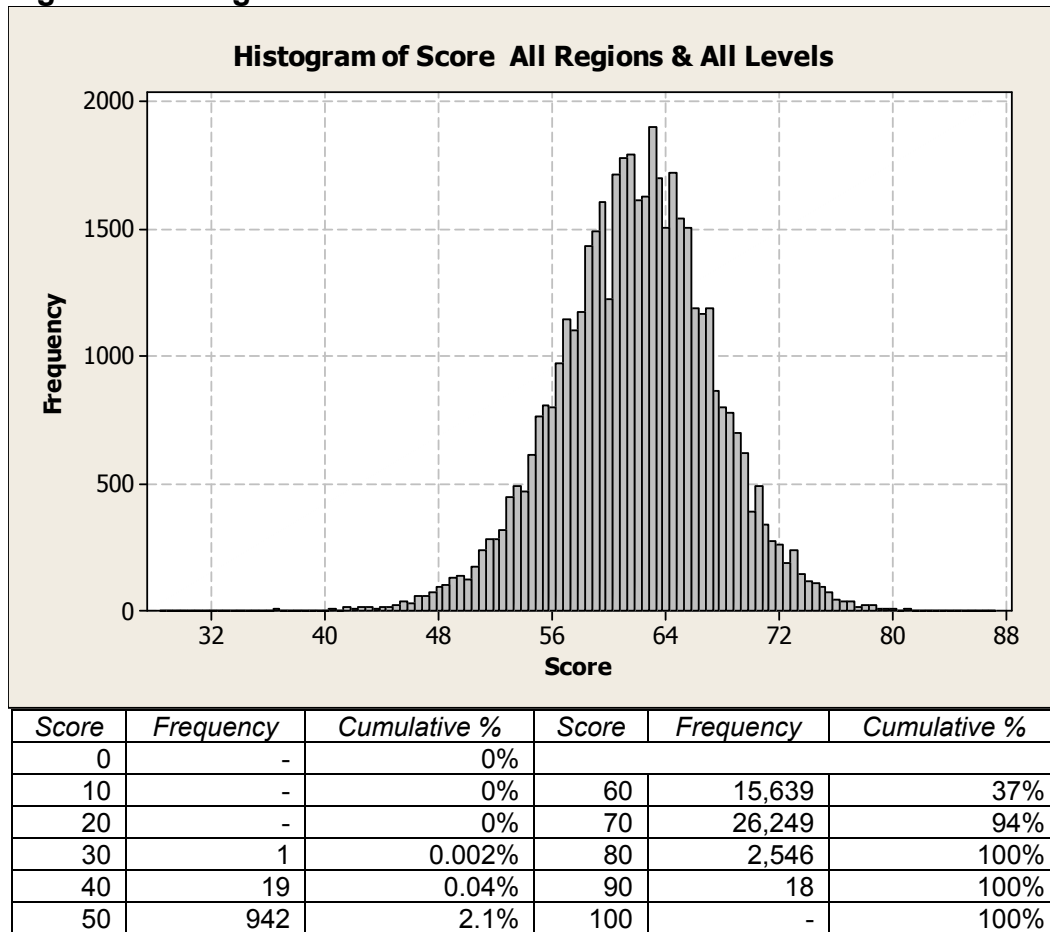
For this analysis, we examined 45,413 rides in Open Performance Dressage Classes. Scores from all classes with restriction descriptors such as: Breed, Height, Materiale, Restricted, Freestyle, Pony, Junior, Young Rider, Sport Horse, or Qualifying, were excluded from the analysis. The data includes Adult Amateur (AA) and Professional riders. There is no information available to allow separation of AA's from professionals.

3. Results

OVERALL ANALYSIS: The 45,413 rides in Open Class rides were performed by 8,341 individual riders at 485 separate shows across the United States. This data set was approximately normal. (Anderson-Darling statistic $p=0.005$). ANOVA and ANOM examination of this data set showed that average scores varied by region, horse breed, level and judge.

Figure 1 shows a histogram and table of scores in this large data set. There was only one score in the “Bad (10-30%) range. There were 961 scores in the “Insufficient/Marginal” range (2.5%). There were 15,639 (34%) scores in the “Satisfactory” range. There were 26,249 (58%) scores in “Fairly Good” range; and 2,546 (5.65%) scores in the “Good/Fairly Good” range. There were no scores above 90%. The highest score was 87.2%.

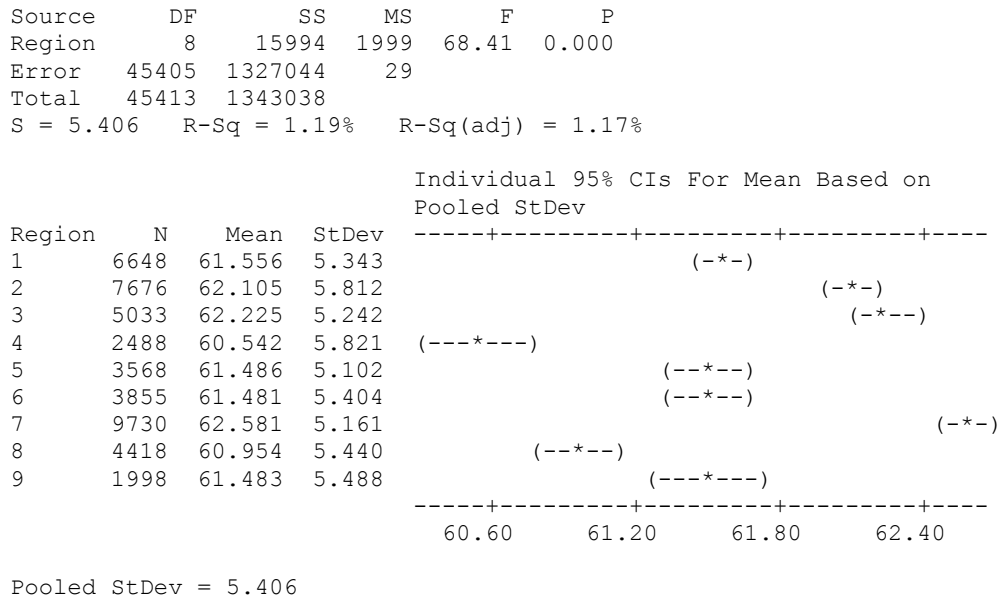
Figure 1: Histogram and Table of Scores



ANALYSIS OF SCORE BY REGION: In the United States there are nine geographical regions for national competition in dressage. ANOVA showed that there were differences in average scores by region. The average score in the lowest scoring regions, Regions 4 and 8, were approximately 2 points lower than the average scores in the highest scoring region, Region 7. The difference between the highest and lowest scoring regions is a 3% difference.

Figure 2 shows the ANOVA analysis by region. The ANOVA results indicate that there is a statistically significant difference in mean score by region. The highest scoring region is Region 7 with 9,730 rides and a mean score of 62.581. The lowest scoring region is Region 4 with 2,488 rides and a mean score of 60.542. The p-value = 0.000.

Figure 2: One-way ANOVA: Score versus Region



The Minitab® ANOVA results indicate the average score by an asterisk. The 95% confidence interval (CI) for the mean is indicated by the dashes between parentheses. The confidence interval can be considered a tolerance band for concluding with 95% confidence that the mean would occur inside the brackets. The p-value allows one to draw a conclusion from an ANOVA test. The authors use the Minitab® output format indicating p=0.000, when p is very small (much less than 0.01).

Figure 3: ANOM by Region

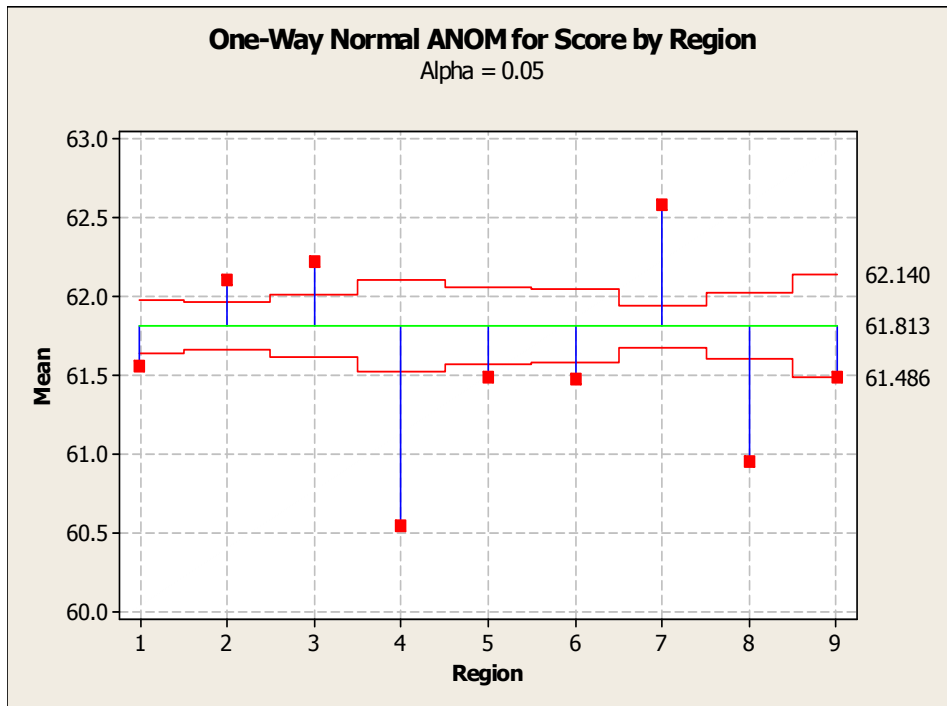


Figure 3 shows the ANOM by region. This analysis identifies graphically the six regions with average scores lower than the grand average and three regions with average scores above the grand average.

ANALYSIS OF SCORES BY LEVEL: There are five national levels and three international test levels. ANOVA showed that there are differences in scores by level. Training, First, and Intermediate Levels had higher average scores than the other levels.

Figure 4 shows the ANOVA by level. This analysis shows the differences in the average score by dressage test level.

Figure 5 shows the ANOM by level. This analysis identifies graphically the five levels with average scores lower than the grand average and three levels with average scores above the grand average.

Figure 4: One-way ANOVA: Score versus Level

Source	DF	SS	MS	F	P
Level	8	21946	2743	94.28	0.000
Error	45405	1321092	29		
Total	45413	1343038			

S = 5.394 R-Sq = 1.63% R-Sq(adj) = 1.62%

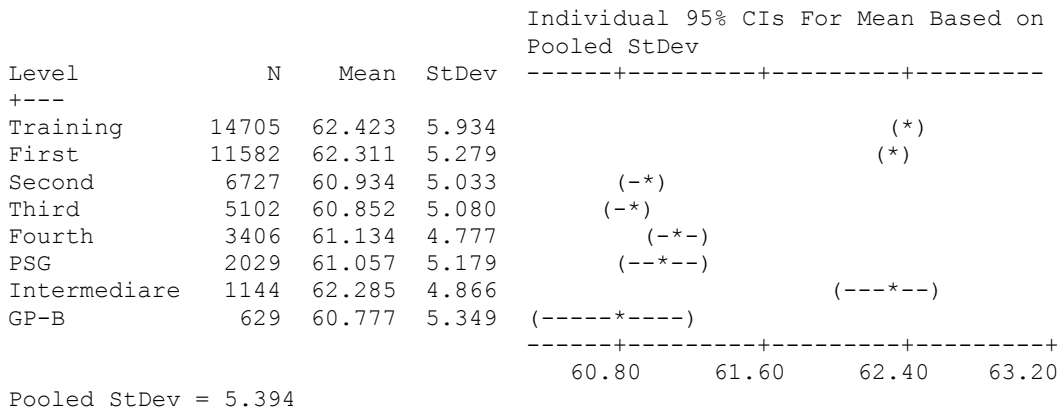
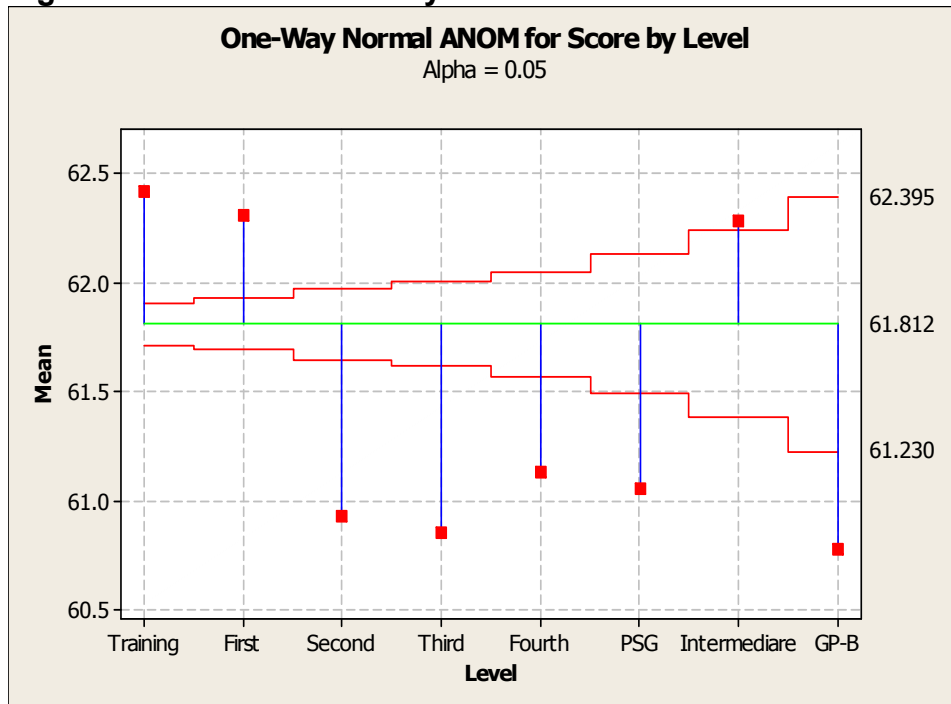


Figure 5: ANOM of Score by Level



ANALYSIS OF SCORE BY BREED: In order to understand the impact of horse breed on final score, the data was segmented by horse breed. This analysis was done to test the hypothesis (and common perception) that certain horse breeds are more favorably evaluated than others. A simple count of the number of rides by breed was done. There were 177 specific breeds or breed crosses contributing the 45,413 scores. Of these, 44 breeds contributed 40,991 (90%) of the scores. There were 2,515 rides (5%) excluded from the analysis where the breed was not listed or listed as “unknown.”

Horse breeds were then grouped according to type as indicated by conformation and way of moving, which are factors that are scored on the tests. The Figure 6 ANOVA and Figure 7 ANOM results indicate statistically significant differences in mean score by breed. European Warmblood breeds, which are generally bred specifically for dressage, scored approximately 2.5 points higher than the lowest scoring breeds, which are bred primarily for other sports (Arabian, Quarter Horses, Thoroughbreds, Appaloosa and Paint). This difference represents approximately a 4% difference between highest and lowest scoring breeds.

Figure 6: One-way ANOVA: Score versus Breed Group

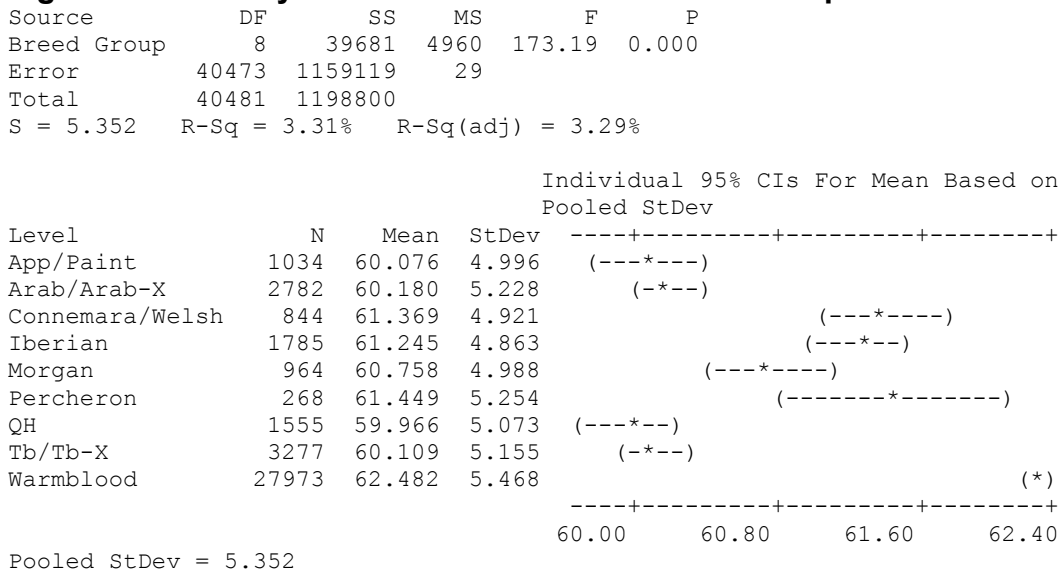
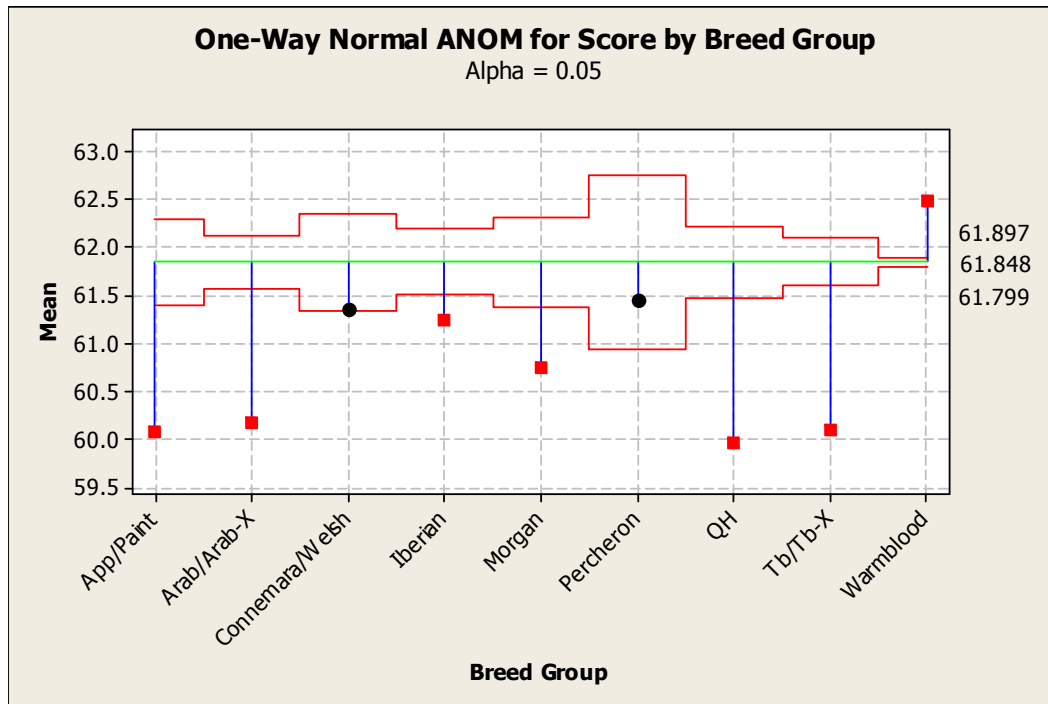


Figure 7: ANOM by Breed Group



ANALYSIS OF JUDGING VARIABILITY: The ANOVA analysis of the large data set indicated that there is variability in score by judge. Because the factors of region, breed and level had been found to be statistically significant, the larger data set was subset to eliminate these confounding factors. The analysis for variability in judging was done on only one level, in two regions, for riders riding one “type” of horse. Results from the two largest regions (Region 2 and Region 7) contributed 38% of the data and were chosen for the analysis of variability in judging. This selection offered the largest population of scores from which to do the analysis.

Region 1: DE, MD, NJ, NC, PA, VA, Eastern WV, DC

Region 2: IL, IN, KY, MI, OH, WV, WI

Region 3: AL, FL, GA, SC, TN

Region 4: IA, KS, MN, MO, NE, ND, SD

Region 5: AS, CO, Eastern MT, NM, Western TX, UT, WY

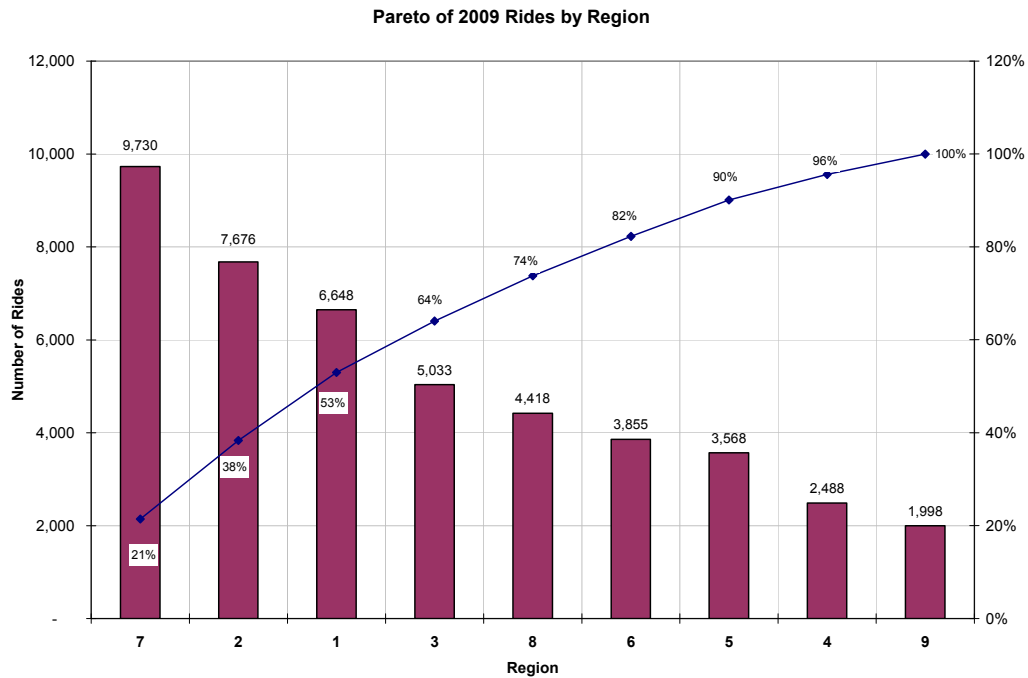
Region 6: AK, ID, Western MT, OR, WA

Region 7: CA, NV, HI

Region 8: CT, ME, MA, NH, NY, RI, VT

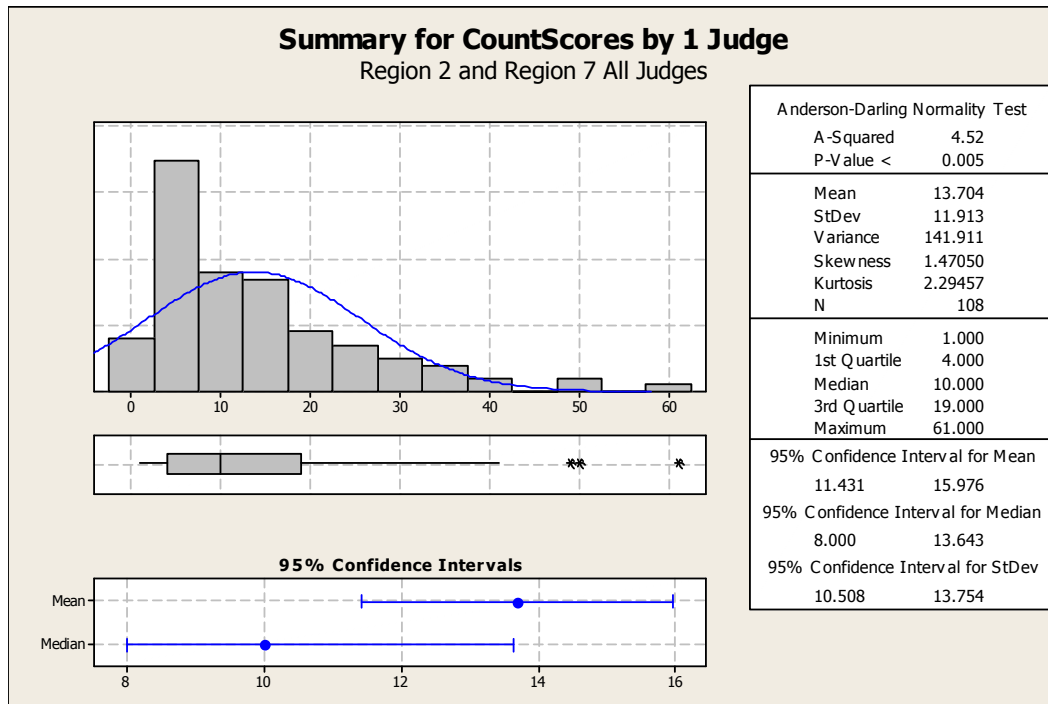
Region 9: AK, LA, MS, OK, TX

Figure 8: Pareto of Number of Rides by Region



To reduce bias from covariants, we examined a subset of the data from two regions having similar average scores at one level. A very basic performance level with a large population was chosen (Training Level). These scores came from tests ridden on horse breeds of a similar genetic and geographical origin (Dutch Warmblood, Hanoverian and Oldenburg). These breeds are the three most popular. There were 688 rides from USDF Region 2 (Illinois, Indiana, Kentucky, Michigan, Ohio, West Virginia, Wisconsin) and 792 rides from USDF Region 7 (California, Nevada, Hawaii) used in the judging variability analysis. The data set of scores was approximately normal for both regions according to the Anderson-Darling statistic ($p=0.049$). There were 108 judges identified judging in the two regions. Two judges provided only one score and were excluded from the analysis. The average judge provided 13 scores, though individual judges provided from two to sixty scores. The mean score was tabulated by judge. The results were order-ranked by quintile and decile for each region.

Figure 9: Histogram for number of scores awarded by one judge



Because of regional differences, the judging analysis looks at the two regions in total, and again by individual region. A 2-Sample t-Test was done on the scores by region (Figure 10) which showed that the difference in average score between regions was 0.63 points. The p-value=0.049. One could say there is no practical significance between the scores in the two regions.

Figure 10: Two-Sample T-Test and CI: Score, Region

Region	N	Mean	StDev	SE Mean
2	688	65.48	6.16	0.24
7	792	64.85	6.13	0.22

Difference = mu (2) - mu (7)
 Estimate for difference: 0.632
 95% CI for difference: (0.004, 1.261)
 T-Test of difference = 0 (vs not =): T-Value=1.97 P-Value=0.049 DF=1446

In Region 2, (Figure 11) ANOVA shows the average score awarded by judges of the lowest decile was 13 points lower than the average score of the judges in the upper decile. When grouped by quintile, the average score in the lowest quintile was almost 10 points lower than the upper quintile (Figure 12). The standard deviation of the upper and lower deciles was 8 to 10 times greater than the middle deciles. The standard deviation of the upper and lower quintiles was 4 to 5 times greater than the middle quintiles.

Figure 11: Region 2 One-way ANOVA: Mean versus decile

Source	DF	SS	MS	F	P
Decile	9	781.17	86.80	83.97	0.000
Error	56	57.88	1.03		
Total	65	839.05			

S = 1.017 R-Sq = 93.10% R-Sq(adj) = 91.99%

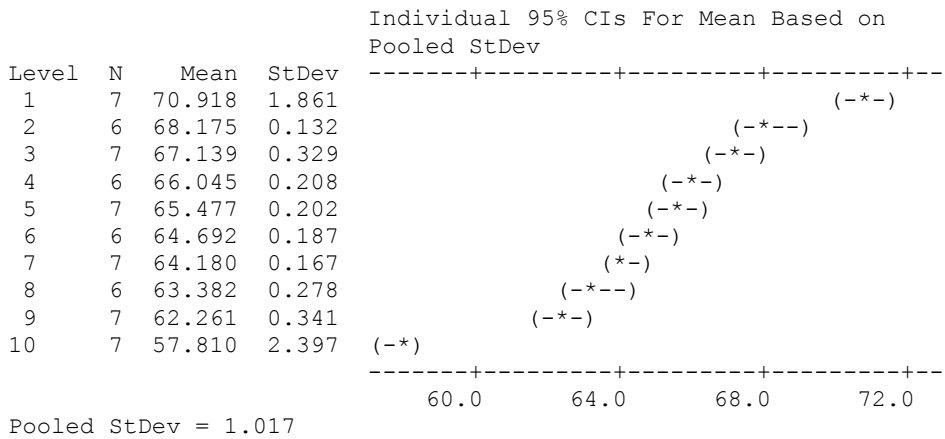
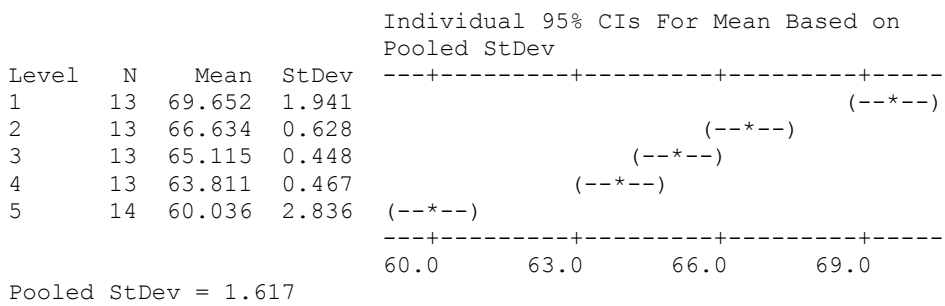


Figure 12: Region 2 One-way ANOVA: Mean versus quintile

Source	DF	SS	MS	F	P
Quintile	4	679.57	169.89	64.98	0.000
Error	61	159.48	2.61		
Total	65	839.05			

S = 1.617 R-Sq = 80.99% R-Sq(adj) = 79.75%



In Region 7 ANOVA shows the average score awarded by judges of the lowest decile was almost 10 points lower than the average score of judges in the upper decile (Figure 13). When grouped by quintile the lowest quintile was about 7 points lower than the upper quintile (Figure 14). The standard deviation of the upper and lower deciles was 4 to 5 times greater than the middle deciles. The standard deviation of the upper and lower quintile was 4 times greater than the middle quintiles.

Figure 13: Region 7 One-way ANOVA: Mean versus decile

Source	DF	SS	MS	F	P
Decile	9	466.763	51.863	133.47	0.000
Error	56	21.759	0.389		
Total	65	488.522			

S = 0.6233 R-Sq = 95.55% R-Sq(adj) = 94.83%

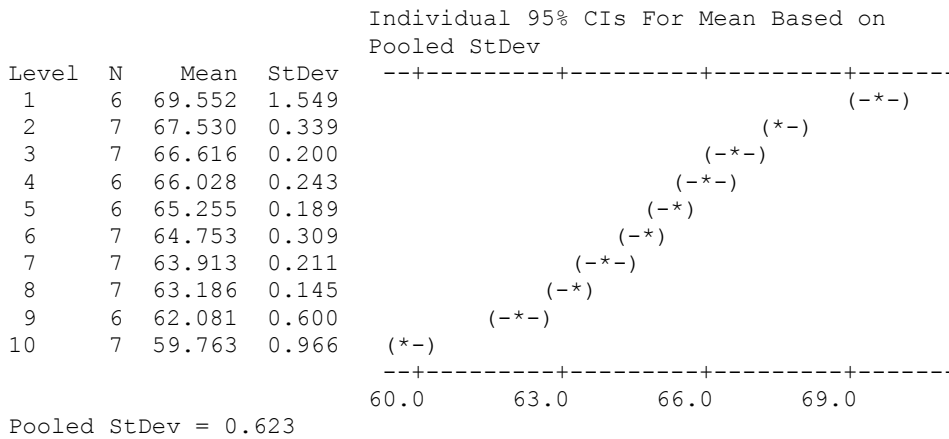


Figure 14: Region 7 One-way ANOVA: Mean versus quintile

Source	DF	SS	MS	F	P
Quintile	4	424.82	106.21	101.70	0.000
Error	61	63.70	1.04		
Total	65	488.52			

S = 1.022 R-Sq = 86.96% R-Sq(adj) = 86.11%

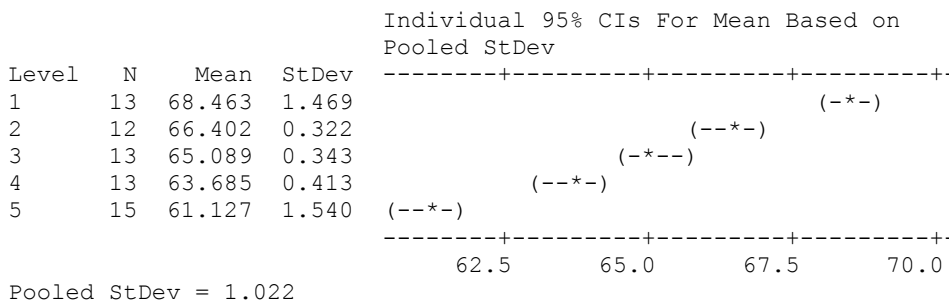


Figure 15 and Figure 16 show the ANOM results by region. These figures show the average score and 95% confidence interval for the mean by judge. Figure 15 shows the results for competitions held in Region 2. Figure 16 shows the results for competitions held in Region 7.

Figure 15: ANOM for Region 2. Three judges scored differently than grand average

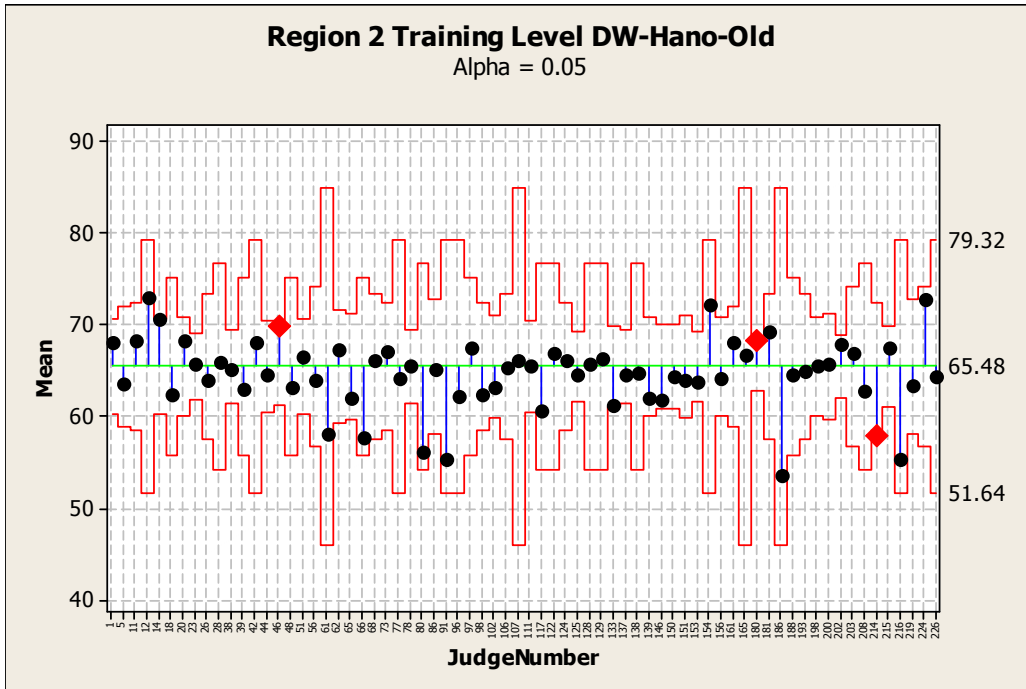


Figure 16: ANOM for Region 7. Two judges scored higher than grand average

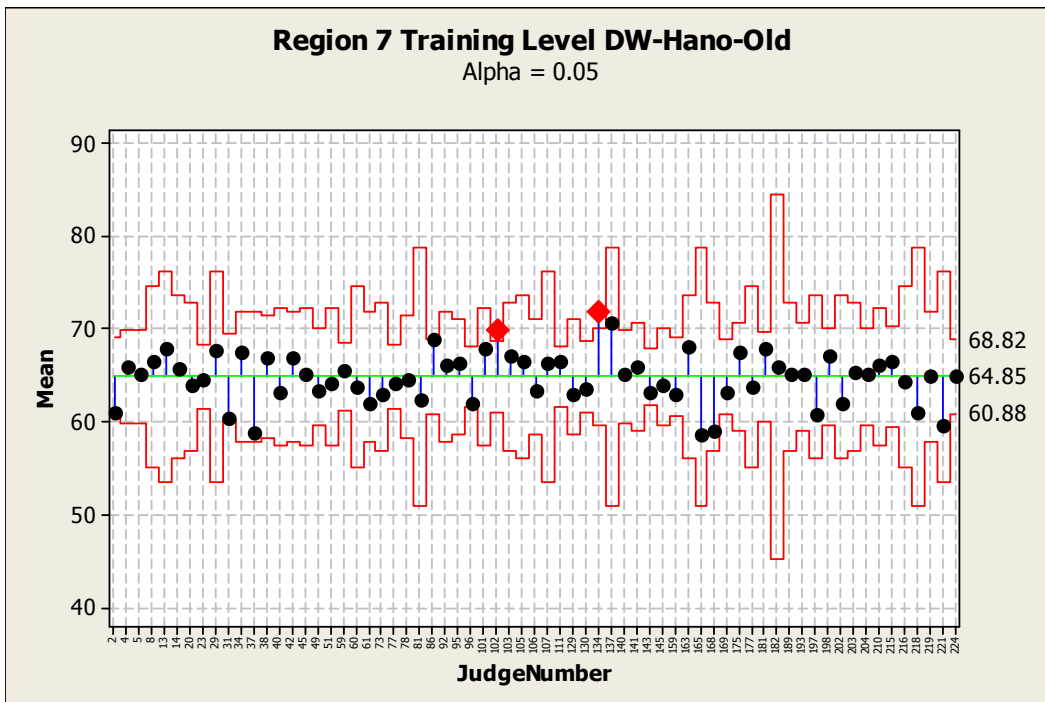


Figure 17 shows tabular results for the Analysis of Means (ANOM) which identified five judges in the two regions who scored statistically differently than the grand average. Only one judge (Judge #102) judged in both Regions. In Region 2, Judge #102's average scores were 63.24 ± 6.03 .

Figure 17: ANOM Results for judges identified different than mean

Region	Judge #	Mean	Std. Deviation	Rides Scored
2	46	69.9	6.24	21
2	180	68.3	6.23	50
2	214	58.0	3.24	8
7	102	69.84	7.96	25
7	134	71.94	6.65	14

4. Discussion

This paper describes patterns observed in recent dressage scores in the United States. It examines larger trends and explores the question of consistency in judging. The work limits the judging analysis to one elementary level, in the two largest regions with riders riding similar horses. The basis for the analysis of judging variability is the assumption that independent and identically-distributed random variables (riders) have the same probability of riding for any judge. In other words, the judges should have seen the same distribution of talent and that any difference in the mean score must be attributed to judge bias. Both tests used in the analysis require approximately “normal” data. All data met this criterion.

An elementary level was chosen for the analysis as this provides the greatest number of scores and explores the question of whether there is consistency in judging the basic movements of dressage.

The ANOVA results show significant differences in the average score awarded by the highest and lowest scoring groups of judges. The results also show unexpectedly large standard deviations at the “extremes” of the groupings (highest and lowest decile or quintile) of judges. These results show that there is increased variability in how to interpret scoring at the upper and especially at the lower ends of the scale. The standard deviations of the highest quintiles are approximately 4 times greater the standard deviation of the center quintiles for both regions. The standard deviations of the *lowest* quintile are 4 to 6 times greater than the average of the three middle quintiles for both regions.

When the data is more finely segmented by decile, regional differences appear. In Region 2 the standard deviations of the *highest* decile are approximately 8 times greater the standard deviation of the center deciles. The *lowest* decile showed a standard deviation that was 10 times greater than the

middle deciles. For Region 7, the standard deviation of the *highest* decile was 5.5 times greater than the standard deviation of the middle deciles. The standard deviation of the *lowest* decile was 3.5 times greater than that of the middle deciles. These findings raise questions on judges' consistency in interpreting judging standards away from the center of the scale.

Understanding the current state of scoring in dressage equestrian competition is a needed first step before implementing any change. Quality engineering recommends that one develop a baseline with knowledge about what is broken before implementing improvement efforts. Based on the uncertainties raised about dressage competition judging, there is probable agreement that something can be improved. However, just what should be done is not universally agreed upon. In problem solving, the point of departure should be a dialog to identify points of agreement, disagreement and areas needing further study. This paper hopes to further the dialog with a statistical analysis of dressage scores in the United States.

Dressage movements are judged against standards described by the FEI for the international tests and the USDF for the national levels. This model of judging was created to eliminate individual bias so that a given ride would be judged similarly by all judges. Licensed judges must initially undergo an extensive classroom and field training program and must pass a final examination. In order to maintain their license, judges are required to participate in supplemental training programs. To judge at higher levels, individuals must undergo progressively further training and testing programs. The goal of these programs is to maintain consistency for all judges in all parts of the country.

The results presented in this paper show there are statistically significant differences in scores between groups of judges. The same horse and rider pair can potentially receive different scores for the same performance when ridden in front of different judges. Competitors have been known to choose specific shows featuring specific judges that favor their riding or their horse. In such cases, competitors are not obtaining truly accurate scores and horse/rider combinations are not being scored to the same standard. In order to address this variability, judge education programs should increase emphasis on how to evaluate exceptionally well-performed and exceptionally poorly-performed movements.

The FEI Dressage Task Force Final Report (FEI 2009b) recommended the use of Consistency Statistics as aids to improve the processes of judge training. As this paper demonstrates, the same lack of consistency in judging and the existence of judge biases also occur at the U.S. national level. Therefore, we recommend the USDF apply statistical reviews to evaluate the efficacy of judge training programs. Additionally, statistical analyses of judge scoring patterns can be performed to achieve better consistency with the stated judging standards.

5. Conclusions

Analysis of Variance was used to demonstrate that key variables impact dressage scores. The major part of this analysis was a study of variability in judging. ANOVA showed there are populations of high scoring and low scoring judges. The difference between the highest and lowest scores can be significant. The study shows increased standard deviations at the “extremes” of the sub-groups (high and low deciles or quintiles). These findings indicate that there is variability in how to interpret the scoring criteria at the upper and lower ends of the scale. Analysis of Means identified five judges whose scores were statistically different than the grand average score in two Regions.

References

- Cuckson, P, (2009) “Dutch dressage sensation Moorlands Totilas tops rich list at £22.5 million” *The Telegraph*, December 28, 2009, <http://www.telegraph.co.uk/sport/othersports/equestrianism/6900840/Dutch-dressage-sensation-Moorlands-Totilas-tops-rich-list-at-22.5-million.html>
- Deuel, NR and Russek-Cohen, E; (1995) “Scoring analysis of three world championship three-day events, *Journal of Equine Veterinary Science*, v.15 479-486
- Eurodressage (2008) “FEI President Princess Haya Asks for Resignation FEI Dressage Committee” November 3, 2008 <http://www.eurodressage.com/news/dressage/fei/2008/resignation.html>
- Fédération Equestre Internationale (FEI 2009a) “FEI Dressage Task Force - Summary Report for FEI Stakeholder Groups,” http://www.fei.org/Disciplines/Dressage/News/Info_Dressage/Pages/summ.aspx?newsName=FEITaskForce.aspx&inc=0
- Fédération Equestre Internationale (FEI 2009b) “Report of the FEI Dressage Task Force” http://www.fei.org/Disciplines/Dressage/News/Documents/FEI%20DTF%20report%2016%20October_final.pdf
- Haya (2008) “Princess Haya's letter to the FEI dressage committee” <http://www.horseandhound.co.uk/competitionnews/388/271337.html>
- PVDA “Updated Analysis of Dressage Scores” [http://www.pvda.org/Documents/Updated Analysis of Dressage Scores.aspx](http://www.pvda.org/Documents/Updated%20Analysis%20of%20Dressage%20Scores.aspx)

- Stickland, D (2009a) “Judging Myths Facts and Solutions; EC 2009” Presentation to the Global Dressage Forum October 2009 <http://www.dressage-analysis.com/GDF.pdf>
- Stickland, D (2009b) “Initial Report on the Judging Systems Trials held at Aachen, September 2009” <http://www.dressage-analysis.com/Aachen.pdf>
- United States Dressage Federation (USDF) <http://www.usdf.org>
- Whitaker, TC and Hill J (2005a) Dressage scoring patterns at selected British Eventing novice events, *Equine & Comparative Physiology*, 2(2); 97-104
- Whitaker, TC and Hill J (2005b) “A study of scoring patterns at national level eventing competitions in the UK,” *Equine & Comparative Physiology*, 2(3); 171-183