AN ANALYSIS OF THE VARIATION IN DRESSAGE JUDGE SCORING

Sarah Kreuz

A Thesis

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

August 2018

Committee:

James Albert, Advisor

John Chen

Christopher Rump

ABSTRACT

James Albert, Advisor

In any subjectively scored sport, there is always the possibility of judge bias. After events at the 2008 Olympics at Beijing caused the scoring methods for international dressage competitions to come under scrutiny, the Federation Equestre Internationale (FEI) responded to the need for additional research into the issue. Following the patterns of previous research, their studies relied heavily on techniques such as Analysis of Variance (ANOVA) to make conclusions about contributing factors to judge bias and pointed to factors such as location of the event, the breed of the horse, and the test level as indicators of bias.

While ANOVA is helpful for finding variation between groups, it does not take into account the individuality of competitors and different sample sizes. For that reason, in this study we focus on Bayesian multilevel models to examine the variation in dressage judge scoring. Not only do these models allow for individual skill levels to vary between competitors, but they also adjust for different sample sizes when some individuals provide more information than others.

In our models, we examined the fixed factors of region, test level, and judge rating, but also allowed varying intercepts for individuals within groups for judge, horse, and rider. Since our focus was judge bias, we used different models to see how outside factors affected variation in judge scores. While adding different factors did show some impact on the variations for the three groups, those effects did not necessarily indicate bias. Instead, using multilevel models implied that most of the variation in dressage scores is due to differences between riders while judges score fairly similarly. Furthermore, the percentage of the overall score that is due to judge variability is quite small compared to the percentage contributed by the horse and rider implying that skill level is the most important factor in dressage scores.

ACKNOWLEDGMENTS

I would like to thank Dr. James Albert for his insights and guidance throughout this entire process as well as Dr. Christopher Rump and Dr. John Chen for serving on my committee.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER 1 INTRODUCTION

1.1 Background

The earliest records of dressage come from the Greek military leader Xenophon in the aptly titled manuscript, *The Art of Horsemanship*. As a military leader whose success in battle depended on the ability of horses and riders to maneuver through a battlefield, he understood the importance of rigorous training to improve dexterity and forge a bond between horse and rider. A disobedient or untrained horse on the battlefield could mean the difference between glorious victory or crushing defeat. For this reason, Xenophon began a training program to improve the athletic ability of his cavalry and build the relationship between horse and rider that allowed their movements together to look effortless and made horses willing to submit to the ordeal of going through battle.

Although the training Xenophon started continued in Greek military culture and spread throughout other nations, it was during the Renaissance that dressage truly became prominent and appreciated for its own sake. As beauty and art were once again considered worthy pursuits, the art of horseback riding began to develop. Riding schools catering to the nobility emerged throughout Europe and in 1735 the Spanish Riding School in Vienna was established, making great strides in exploring what horse and rider could accomplish together.

As defined by the Oxford Dictionary, dressage is "the art of riding and training a horse in a manner that develops obedience, flexibility, and balance." When horses were still used as a primary form of transportation, it was especially vital for cavalry men to be adept in dressage. Civilians, however, did not always recognize the benefit of dressage training. In fact, dressage was considered a purely military sport to the point that when it was eventually added to the Olympic games in 1912, only military officers were allowed to compete. It was not until 1953 that civilians as well as military personnel participated in dressage competitions. In those first Olympic games, dressage looked more like a military exercise than an art form. Nevertheless, as riders pushed boundaries and explored possibilities, the level of intricacy in patterns and athletic ability required

continued to increase. Even as horseback riding decreased in its daily utility, dressage helped strengthen the understanding of appreciating its beauty.

In order to compete in dressage, riders perform what are called “tests” which are standardized patterns created to showcase certain levels of skill. Tests are conducted in a dressage arena which is designed as either a 20x40 meter (small) or 20x60 meter (standard) rectangle with letters at specific intervals around the perimeter. For example, on the short end of the rectangle on one side is the letter A where the rider enters the arena and on the opposite side of the rectangle is C where the judge is positioned. Sometimes there are judges at other letters such as B and E which are in the middle of the long sides, or at H, and M which are each six meters from the corner across a long diagonal. All of the letters that are found on the perimeter of the rectangle are marked with a sign, however, there are also unmarked letters on the centerline of the arena at specified intervals with the letter X exactly in the middle of the arena. When performing the test, riders use these letters as markers for the patterns they are executing. Each pattern is broken up into a sequence of movements which indicate what the horse and rider should be doing at that moment. Each movement is scored individually by the judge and the points are all added together and turned into a percentage of points possible to generate the final score for that test.

There are a few different types of tests riders can choose from based on where they are competing. Some tests, like those written by the Federation Equestre Internationale (FEI) are shown at the international level while others, like those written by the United States Equestrian Federation (USEF), are only ridden at the national level. In USEF dressage tests, the lowest level of difficulty is the Introductory Level which features basic maneuvers like a twenty-meter circle. The next level of difficulty is Training Level which includes more cantering and greater variety in the patterns. After Training Level, riders move through First to Fourth Level tests which require a much higher skill level and feature more intricate movements as the riders progress through the levels. Every four years these tests are updated by the USEF to remedy problems that were noted throughout their tenure or to add skills deemed important for that level. The patterns are first decided upon by a committee to make sure that all the necessary skills are showcased in each test from both

directions. Once the committee has agreed on a set of movements, they conduct trial runs with many different horses and riders at varying levels of experience. In that way, they can make sure that certain types of horses do not have an advantage over others and that all the movement combinations flow together smoothly. They can also get an idea of what issues riders might run into when performing that test. For instance, some movements may be easier for smaller horses to perform than larger horses. If size of the horse becomes an issue, it may require rewriting part of that movement in a way that ascertains no particular horses are given an advantage. Furthermore, since smaller horses may take longer to complete a test, the committee needs to make sure that the tests are not overly long. Most dressage tests are completed in under six minutes so if a new test is consistently taking competitors longer than that time the patterns are restructured to optimize the balance between content and length. Once all the tests have been tried out a satisfactory amount of times the tests go into the final editing stages. All necessary movement changes are made, the wording is edited, and the finished document is sent through many stages of proofreading. After all final changes are approved, the committee decides the tests are ready to publish.

Part of the entire process is deciding exactly how each movement should be scored. When the committee is revising the wording on some sections, they are making sure that the expectations are communicated clearly to the rider as well as to the judge so that both are held to the same standard. At the top of each test is a paragraph concerning the purpose of the test that gives both riders and judges an overview of the goals and expectations of that level. Furthermore, placed next to each movement is a box containing a directive idea which lets the rider know the essential qualities of the movement and also serves as a reminder to the judge of the important aspects to score.

Each movement is scored on a scale of 0-10 with a 0 meaning the movement was not performed at all, and 10 indicating that the execution was excellent. Half points are allowed, and on average most riders tend to receive scores of about 6 (satisfactory), or 7 (very good). Not all movements are considered of the same importance, however. If a movement is considered more vital to the performance, it is given a higher weight through use of a multiplicative coefficient such as multiplying the movement score by 2. Once the test is complete and all the movements have been scored, the

judge awards collective marks which rate the overall performance. These collective marks are also awarded on a scale of 0-10 and not only score the performance of the horse, but that of the rider as well. Once all of these points have been added up the last step before awarding the final percentage is to subtract any error penalties. If a rider performs a movement incorrectly, uses vocal cues, the horse leaves the ring, etc. the judge will ring a bell to indicate a penalty. The first error is a 2 point deduction, a second error is a 4 point deduction, and a third error is cause for elimination from the class although the rider may be allowed to finish the test at the discretion of the judge. After these points are deducted, the final overall score is assigned as a simple percentage created by taking the number of points awarded out of the total number of points possible.

1.2 Motivation

Even with extensive training programs, sports that rely on subjective judges are constantly searching for ways to minimize any possible judge bias and achieve the most objective assessments possible. Failures at attaining this goal, however, continue to arise as evidenced by the figure skating scandal at the 2002 winter Olympics, as well as the debate surrounding Paul Hamms gymnastics gold medal in the 2004 summer Olympics. Dressage is no exception to the conversation about judge bias. On the unofficial level, competitors complain about unfair judging based on superficial qualities. It is not uncommon for riders and trainers to make comments about how other competitors place higher based on gender rather than skill or about how a judge scored them lower because their horse was a certain breed.

After the 2008 Summer Olympics in Beijing, an official complaint on dressage judge scoring practices was issued by FEI President, Princess Haya (Eurodressage, 2008). In a letter published by *Eurodressage*, she listed several concerns about conflicts of interest for judges selected to preside at the Olympic Games. Many of these concerns related back to whether or not the judges chosen had sufficient training in order to accurately judge at the Olympic level.

Due to these concerns, the FEI commissioned a task force to look into the established judging practices and develop recommendations for improvements. After consulting scoring methods in other sports such as ice skating and gymnastics, and carefully reviewing current dressage protocols,

the task force published their proposal (Federation Equestre Internationale, 2009). This proposal contained several measures that would change the way internationally ranked judges were trained and selected for events as well as how those events were designed.

While dressage judge scoring only came under strict scrutiny in 2008, a few papers were written on the topic prior to that time. One of the earlier papers addressing the subject was by Deuel and Russek-Cohen (1995), which focused on analyzing scoring factors in the three-day events at the world championships from 1988 to 1992. Eventing shows include dressage as one of the events, however, that is not the sole purpose of the competition. Horses at these events have to be skilled in multiple disciplines, one of which is dressage, but they also need to excel in jumping and cross-country courses. With such a diverse skill set necessary to compete at these shows, it introduces some nuances to the scores. Therefore, although this paper did look at the dressage scores on their own to see how they contributed to the overall eventing score, the authors were looking through a slightly different lens than they would for a dressage competition.

In 2005, Tim Whitaker teamed up with Julian Hill to write two articles for the journal *Equine and Comparative Physiology* about judge bias in dressage. Similar to Deuels research, these two articles focused on judge scoring at British eventing competitions. In their first paper, Whitaker and Hill (2005b) concentrated on discussing judge scoring bias in the dressage portion of three-day events. Through the use of analysis of variance (ANOVA), they were able to discover differences between scores in the dressage ring based on the location of the event and the test performed.

In their second paper for *Equine and Comparative Physiology*, Whitaker and Hill (2005a) still examined judge scoring at British eventing competitions, however, they moved their focus from analyzing differences between dressage scores toward using linear regression to discern how much each of the three events contributed to the final overall rider score. The final eventing score is tallied by combining the scores from each of the three events. The paper states that originally 75% of the final score was supposed to be explained by the cross-country score while 18.75% came from dressage and 6.25% from show jumping. In more recent years, these requirements have become less strict with the only guidelines indicating that cross-country should carry the most weight,

but not stipulating how much weight. Using multivariate analysis, Whitaker and Hill found that dressage was actually being given the most weight in the final score with over 50% of the eventing score allocated from dressage events at all test levels. While this result does not directly indicate whether judge bias is involved in that score or not, it does point out that judge bias at dressage competitions is a significant factor to consider not only at dressage events, but at any type of show which includes dressage as part of the competition.

More recently, the FEIs examination of the issues cited during the 2008 Olympics in Beijing prompted others to reexamine dressage judge scoring practices as well. In January of 2010, an article was published in the *Journal of Quantitative Analysis in Sports* titled "Scoring Variables and Judge Bias in United States Dressage Competitions" by Diaz, Johnston, Lucitti, Neckameyer, and M. Moran (2010). While the FEI took an international approach to the problem, this study engaged in a more focused view on the issue by narrowing their data to competitions within the United States. This was the first paper to look solely as U.S. dressage competitions from a statistical viewpoint. In their analysis, they primarily used ANOVA and some analysis of means (ANOM) to look for unusual deviations from typical judge rulings. After completing all their analysis, these methods showed significant variations in judge scores by region, test level, and horse breed.

1.3 Taking a Bayesian Perspective

Many of these research papers relied heavily on ANOVA to conduct their analyses. The purpose of ANOVA is to recognize significant differences between group means. For example, in Diaz et al., one of the groups they were interested in was region. While it is not reasonable to expect each region to have the exact same mean, theoretically there should not be a lot of difference between the average dressage score for an East Coast region and a Midwest region. ANOVA assumes that all group means are the same and finds the probability of our observed group means occurring under that assumption. In this example, after analyzing variances between and within the different regions, ANOVA is able to determine if one regions mean score is higher or lower than would be considered normal under our assumption of equal means.

While an ANOVA is a useful tool for showing which factors contribute significant amounts to

overall variation, it cannot account for the variation that should exist based off different athletic abilities. While previous research in judge scoring indicated that region is a significant factor in variation, the authors could not determine if the judges in that region were scoring higher because of bias or if the riders in that region were better athletes. Similarly, perhaps ANOVA shows that test level is a significant factor in scoring. That tells us nothing about whether the judges are scoring one test more harshly than another or if riders who choose to compete with that test are able to perform better than riders who choose another test.

In any sport, we expect to see variation. When an athlete is announced as the winner of a competition it is because they performed better than any of the other competitors. In dressage, we expect one rider to receive higher scores than another. If one person has been riding for twenty years and on the show circuit for ten years and they perform the same test as someone that started riding a year ago and is competing at their first show, the more experienced rider will most likely receive the higher score. In this scenario, the difference in dressage scores was not the result of judge bias but rather a difference in athletic ability.

The goal of this paper is to quantify how much of the variation in dressage scores is due to differences between judges and how much is due to differences between riders. We want to see whether factors like region and test level affect how judges score rides or if it is possible that confounding factors such as riders in certain areas of the United States having better training resources are the true source of variation.

When looking for an appropriate model, we have three options. The first option would be to make one model with one estimate for the average judge score. This model, however, would not give us enough information as it would treat every judge the same and ignore possibly relevant grouping characteristics. Another choice for a model would be to consider every person individually. This model would give us way too much information as it would create a separate model for each individual. This would completely ignore relevant similarities that we could use to help us create better estimates for individuals with smaller sample sizes.

For that reason, we wanted to take a hybrid approach and use a Bayesian hierarchical model

to exhibit the variation. Multilevel models allow us to look at individuality within groups. For instance, suppose we separate our data by the groups judge, horse, and rider. Within each of those groups are many individuals with their own scoring records and experiences. Using random effects for judge, horse, and rider can then tell us how far the average dressage score for each individual in that group deviates from the population mean. To take it a step further, we can also add fixed factors like region and test level to see how they affect the variation within each group. In that way, we not only allow for individuality but can capture information about different groups as well.

Since these models create estimates for each individual in our groups, we can also estimate overall scores for each judge or rider. One issue with comparing judges or riders is that the individual means are based off different sample sizes. With Bayesian multilevel models, this is no longer an issue. When estimating judge scores, we will get both a population average as well as estimated average score for each judge. If one judge only scores four or five rides, they do not have a lot of information about their judging trends while another judge that scores over twenty rides yields a lot more information. The model will then pull each of these estimates towards the population average, however, the judge with less information about them will be pulled closer to the population average than the judge with more observations. This allows us to use the information we have from other individuals in that group to create more accurate results even with limited access to data.

Although Bayesian techniques have not been used to assess equestrian judging methods, it has been used for analysis of bias in other fields. One of the more recent publications by Zupanc and Strumbelj (2018) focused on rater effects in essay scoring. In that paper, they were looking for rater bias on essays written by students graduating secondary school in Slovenia. These essays were each assessed by multiple graders and the variances were analyzed to see how much was due to bias. Although they were using a Bayesian model to assess essay grades, they also pointed out that their model "can be applied in any setting where a set of performances is rated by a set of raters and where at least some performances are assessed by two or more different raters" (Zupanc and Strumbelj, 2018).

In our scenario, our performances happen each time a test is ridden. At times, there will be multiple judges positioned around the ring, however, these positions all have different perspectives and are not fulfilled at every competition. Therefore, we chose to focus only on the judge at C to eliminate that additional factor of perspective and make much more data available for analysis. With the limitation of only using one judge score per ride, we have to ignore any possible time differences in one show season. Throughout a show season, a rider will compete in the same test level multiple times, therefore they will have multiple judge scores per test within a year which we can treat as one performance assessed by multiple raters. Once we make that generalization, we can create a model very similar to the one discussed in Zupanc and Strumbelj.

CHAPTER 2 EXPLORATORY DATA ANALYSIS

2.1 Data Collection

The data for this project was collected from publicly available files on Horse Show Office's website – a horse show management company based out of Dexter, Michigan. In order to focus on the most current information, we selected rides from the year 2017 for analysis. The variables provided by Horse Show Office were Show Name, Location, Dates (Show Weekend), Class Number, Test Name, Arena, Date (Individual Day), Judge(s), Time of Ride, Rider Number, Horse, Owner, Rider, Score(s), Total Score, Status, Place, and Division which we then used to create our final dataset.

The first change we made to the provided variables was to convert the time of ride into a categorical variable. While we still wanted to know the general times of each ride, there should not be a significant difference from one hour to the next. There may, however, be a difference between morning and afternoon rides since that allows time for weather changes or judge fatigue. By indicating whether the ride occurred in the AM or PM instead of the exact minute of the ride, we simplified the variable for ease of analysis while still preserving all pertinent information.

After creating the categorical variable for time, we augmented the data set with information we thought may be influential to judge bias. One of those additional variables was the competition region. The USEF has separated the United States into nine competition regions. To create the variable, we simply looked at the state indicated in the show location variable and noted which region included that state. Seven of the nine regions were represented in our data, however, Region 6 (northwest corner of the U.S) and Region 9 (southwest) did not appear in our data set at all.

We also wanted to look at the level of training received by each judge. There are five different types of judge classifications in the United States. The lowest level is the "L" graduate that has passed the initial coursework towards becoming a registered judge. While "L" program graduates cannot preside at recognized shows, they are allowed to judge at schooling shows or other unrec-

ognized shows that do not require judge credentials. Completing the "L" program is the first step in achieving higher judge ratings, however, only those candidates that graduate with distinction having scored above an 80 on the written exam and above a 70 on the practical exam are allowed to apply for the next higher level of "r" judge. For that reason, we have separated "L" graduates and "L" graduates with distinction (notated here as "LD") since they do represent slightly different levels of qualifications.

Once applicants pass through the "L" program with distinction, they still need to achieve high scores in their own riding, apprentice judge for already recognized judges, secure recommendations from several USEF licensed officials, and complete many more hours of training and exams before receiving their "r" license. Once someone receives their "r" license they become a "recorded" judge and are allowed to judge through Second Level tests at recognized shows. If they want to judge at higher levels they need to gain more experience and complete even higher levels of the training they went through to get their "r" license. If they do meet those training requirements, they can receive their "R" license to become a "registered" judge who is permitted to judge through Fourth Level at recognized shows. Repeating a similar process will allow someone to reach the highest level of USEF judge which is the "S" license or "senior" judge who is allowed to judge through Grand Prix at the national level.

2.2 Exploration

Since the judge at C is the only position that is consistent throughout every test and every show, we focused solely on those scores for this analysis. We started exploring the data by looking at the overall distribution of the scores. In Figure 2.1 we see that the scores look fairly normally distributed with a mean of 63.25 and a standard deviation of 4.995. There are some pretty significant outliers, however, with the lowest score at 38.75 and the highest at 80.96.

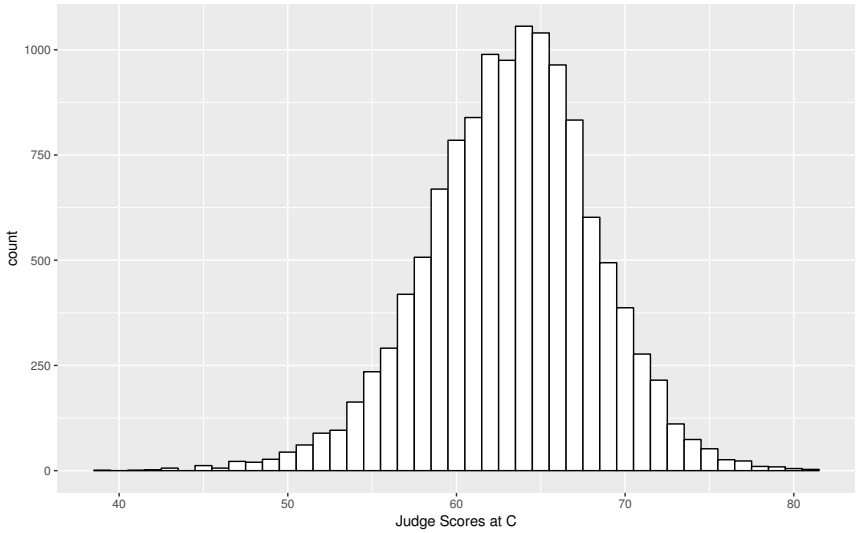

Figure 2.1: Judge Scores

Since the scores exhibit a symmetric distribution, we do not need to attempt any transformations that would complicate our interpretations. From this distribution, we would expect typical scores to fall between 53 to 73%. That raises the question of what caused those unusually high and low scores to occur.

Ideally, we would want the unusual scores to occur because of exceptionally good or terrible rides. If instead these scores were the result of some type of bias, we would expect to see unusual variation in scores within one variable rather than spread out across categories. In most other papers on the subject of judge bias, they discovered unusual patterns in variation by region. Since that variable has been a consistent issue in previous research, it is a good starting point for our exploration.

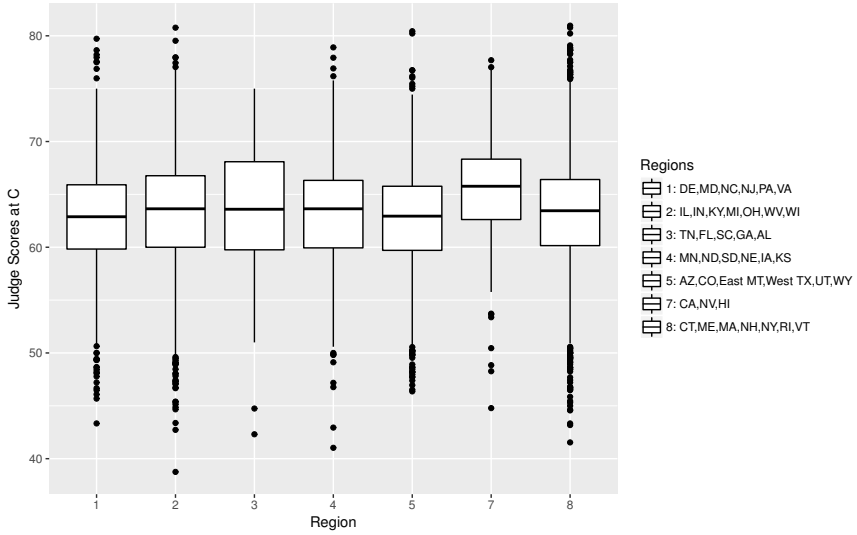

Figure 2.2: Judge Scores by Region

In Figure 2.2, we see pretty equal spreads for all regions with the exception of Region 3. Since Region 3 has the smallest number of rides out of all regions, however, we would expect to see a larger standard deviation. We also notice quite a few outliers both above and below the median throughout all regions. The number of outliers increases with the number of rides in each region as Region 8 seems to have the highest number of outliers but they also have the highest number of rides, while Region 3 seems to have the fewest outliers but that is expected since they also have the fewest rides. With a wide range of possible scores going from 0 to 100% and the typical range of scores only falling between 53 to 73%, its not really concerning to have this many outliers unless we find that all the outliers are coming from the same judge.

One of the most interesting things about this graph is that although most of the median scores look pretty close to the overall average of 63%, Region 7 has a median score close to 66%. This mirrors the results found by Diaz et al. when they also noted that Region 7 (which includes the states of California, Nevada, and Hawaii) held the highest average scores of any region. It may be interesting to take a closer look at what is happening in that particular area and explore the cause of that distinction. Perhaps the judges in that region do have a tendency to rate rides higher than the rest of the United States judges, but perhaps they have a higher concentration of skilled horse and rider pairs in that area.

The next variable of interest was the test level. With the difference in skill level for each type of test, it is possible that some levels are judged more harshly than others. Although there are three tests in every level, rather than examining eighteen individual tests we took a look at the six different levels.

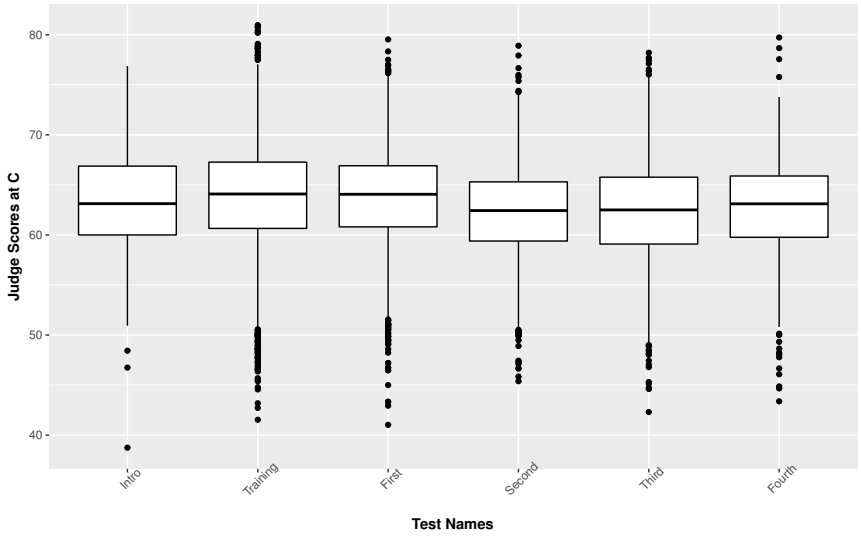

Figure 2.3: Judge Scores by Test Level

In Figure 2.3 we see hardly any variations in scores for the different test levels. The spreads for each tests scores stay pretty consistent and the means only differ through a range of about 1.5 percentage points. One reason this might be happening is that when a rider selects a test to ride, they are not restricted to choosing tests sequentially. That means riders are free to select tests that fit their skill level as well as the abilities of their horse. Some riders will also skip levels, such as Second Level, which are more difficult in order to reach the higher levels without receiving low scores on tests where they might perform poorly. With all the self-selection of tests, riders will tend to elect to compete where they will fall in the average range so that they feel like they can succeed but also have room to improve their scores.

The variable that proved the most interesting was the certification level of the judge. We decided to look at each individual test and see how the different types of judges' scores differed. Since each test has a different focus, even within the same level, we thought it would be important to look at each of those tests individually.

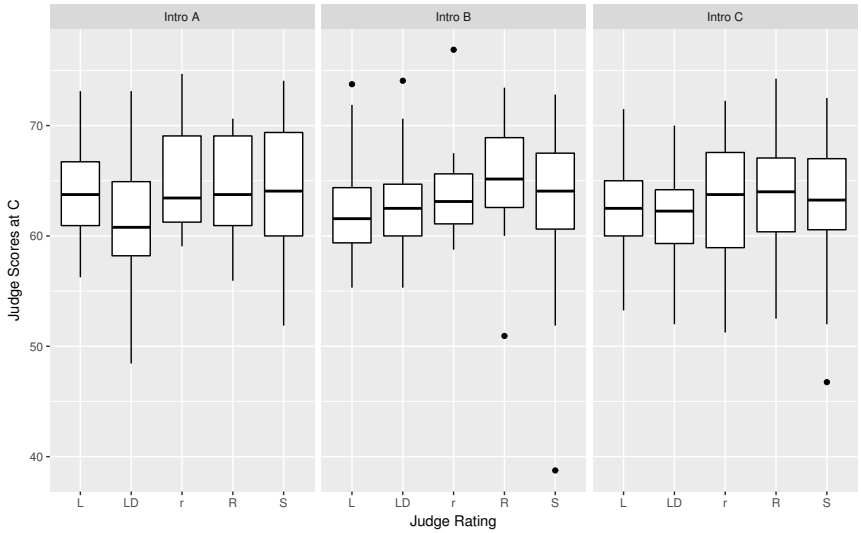

Figure 2.4: Introductory Level Judge Scores

Intro A appears to have pretty constant median scores for the riders except for the scores from the "L" program graduates with distinction. Intro B experiences a slight rise in scores as the judge training increases and Intro C appears pretty constant throughout. We do see a few outliers in Intro B tests, but those are likely from either exceptionally good rides for the high outliers, or riders that were having a rough day with their horse for the low ones. The general trend here is that the very experienced judges appear more lenient in their scoring of beginning riders than judges that have recently gone through the initial judge training program.

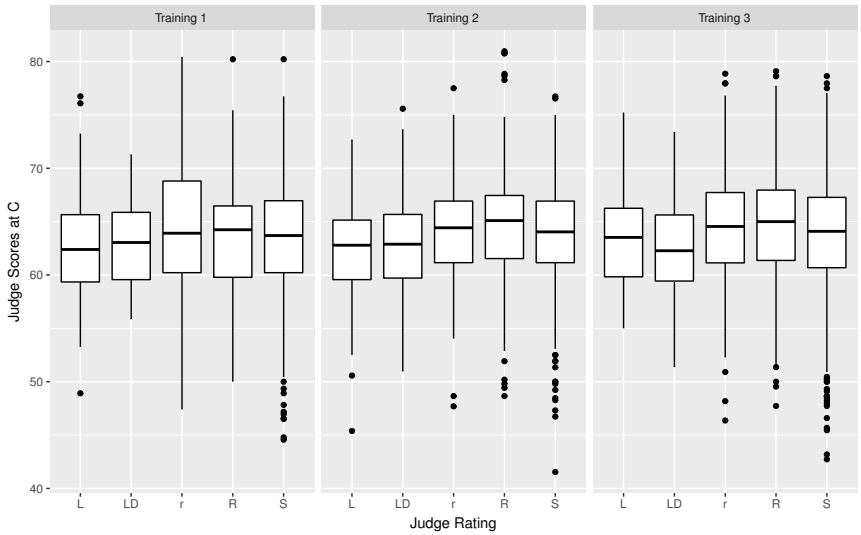

Figure 2.5: Training Level Judge Scores

The Training Level scores look a little more consistent than the Intro test scores. We do see a lot more outliers, but with the large amounts of rides in this level that is not unexpected. Overall, it once again looks like “L” and “LD” graduates score slighter lower than the rated judge scores with the “LD” percentages in Training Level 3 appearing especially low. Training Level tests are still considered to require lower skill levels as it is the level that riders usually begin with when they start showing. With that in mind, we have a similar situation to the Introductory tests where the highly experienced judges may be exhibiting more leniency in their scoring for beginning competitors.

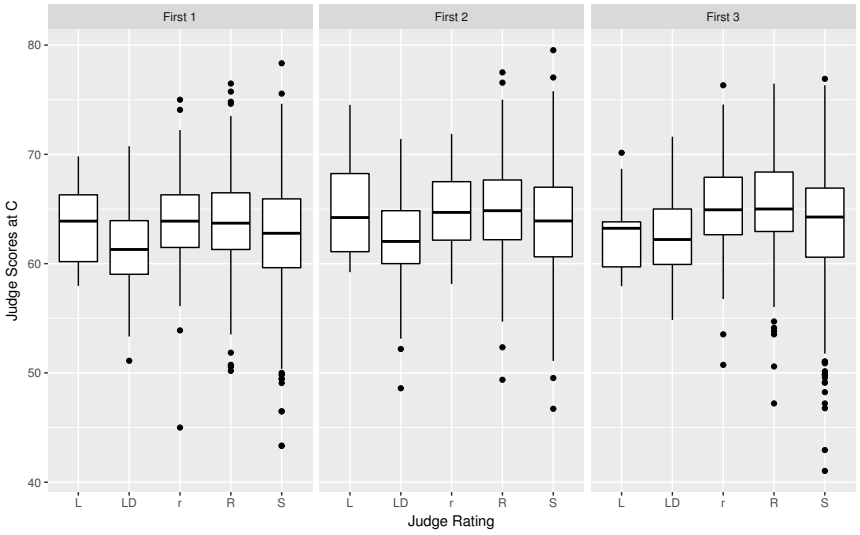

Figure 2.6: First Level Judge Scores

In the First Level test scores, it looks like the “LD” category is consistently much lower than all the rest. “L” looks a little lower than the official judges for First Level 3, but nothing like the dips that “LD” scores are taking. This behavior is a little unusual since “L” and “LD” judges have the exact same training and the only difference between the two categories is how well they performed on their training assessments. The most obvious difference between “LD” and all the other judge ratings for this test level is the number of observations in each group. Out of all the categories, “LD” judges scored the most First Level rides which gives us more information on how they would score those rides than we have for any other type of judge. Thus, it is possible that the other categories of judges on average would score lower except that we do not have enough information on them to see that behavior with the raw data. This is another instance where Bayesian multilevel

models will help us alleviate this issue through shrinking observations with smaller amounts of data towards the overall mean so that these differences in amount of information do not mask general trends.

Figure 2.7 looks at the scores for Second Level. Since “L” and “LD” judges cannot score these rides at rated shows, they had very low frequencies for these rides. Thus, we chose to omit the “L” and “LD” scores in this category.

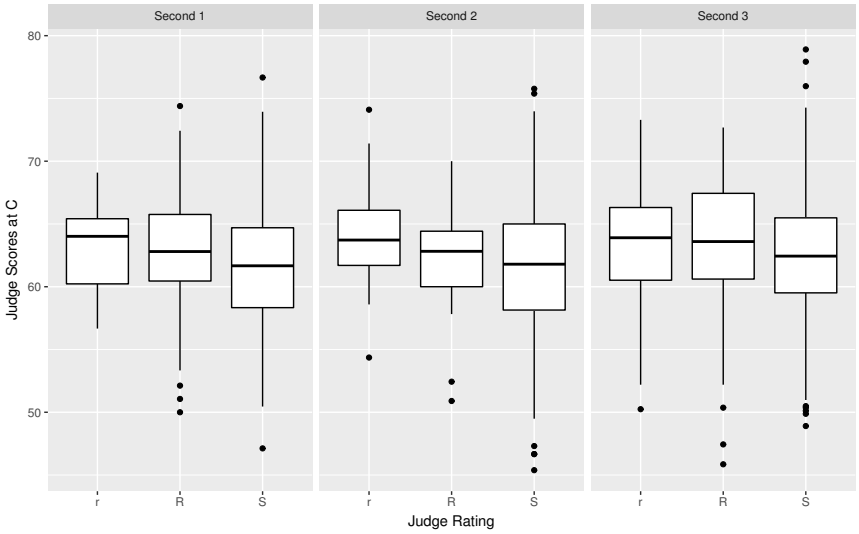

Figure 2.7: Second Level Judge Scores

For Second Level, it looks like the higher the judge rating, the lower the judge tends to score the rides. If we look back at Figures 2.4, 2.5, and 2.6 we see similar behavior happening in every level. While other levels sometimes switched whether “r” or “R” judges scored higher, “S” judges consistently scored the lowest of all three types of rated judges.

This behavior could possibly reflect some of the perspective gained through experience. While “r” and “R” judges did not show large differences in judgement for Training and First Level tests, Second Level is the highest level that “r” rated judges can officiate at rated shows which may mean that they have a less experienced eye for newer and more intricate moves than the more seasoned “R” judges. “S” judges then have by far the most experience of these judges and may start looking more critically at riders who elect to try their skills at the higher-level tests.

When looking at Third Level tests, we only focused on “R” and “S” judges since there were

no tests for “r” judges and the number of tests judged by “L” and “LD” graduates were extremely small.

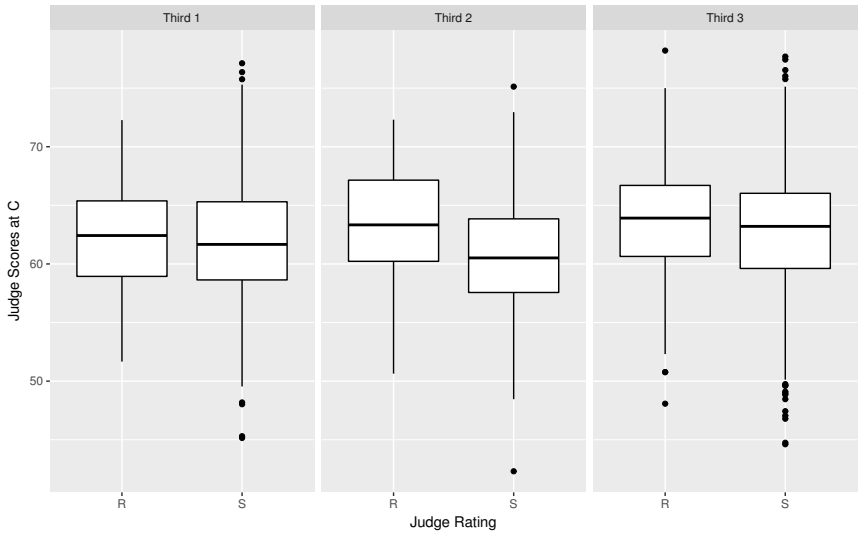

Figure 2.8: Third Level Judge Scores

In Figure 2.8 we see that the “S” judges are scoring slightly lower than the “R” judges. This is most obvious in Third Level Test 2, however, as Third 1 and Third 3 only exhibit miniscule differences. This further indicates the possibility that higher trained judges score rides lower than a newer judge would.

Lastly, we looked at the Fourth Level tests. Similar to the Third Level tests, we only plotted “R” and “S” rated judges.

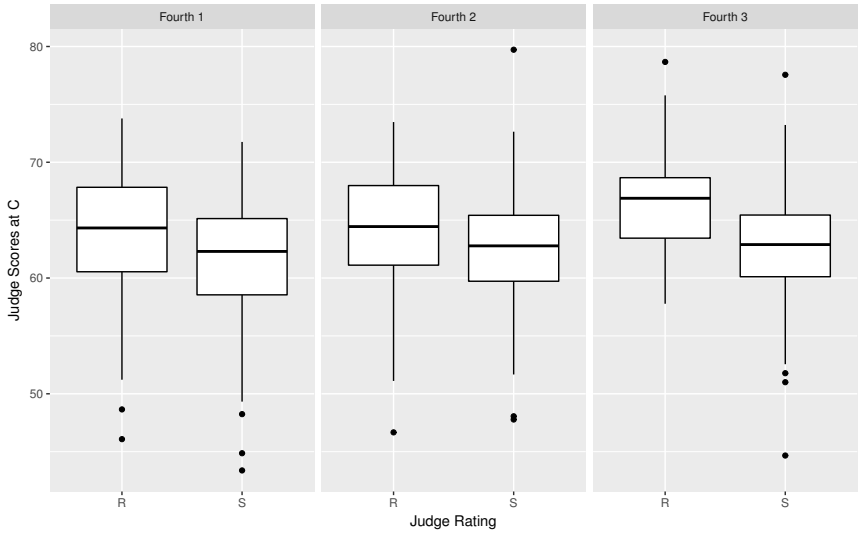

Figure 2.9: Fourth Level Judge Scores

Fourth Level has a more obvious difference between the scores from the two types of judges. Similar to how we saw a distinct difference between "r" and "R" judges once we reached the threshold of "r" judge expertise, Fourth Level is the top level that "R" judges are permitted to score. Although lower levels show a trend of "S" judges scoring lower than "R" judges, Fourth Level takes that difference and increases it substantially. Since Fourth Level is the threshold of "R" judge experience, it is possible that the difference in judge skill level is starting to become more defined and the difference between scores shows up more significantly than it did in the lower levels.

2.3 EDA Conclusions

From our initial analysis, it does not look like region and test level have as large of an impact as previous research has suggested. While many previous papers cite region as one of the important factors in dressage judge bias, our data does not indicate such bias at first glance. Although we do see a higher average score in California and Nevada, we do not yet have sufficient evidence to claim that judge bias is the cause of that peak. Many other factors, such as available training resources, may contribute to that behavior. All other regions show reasonably consistent median scores in these plots.

Test level also does not indicate the large amounts of judge bias suggested by previous research. In fact, our exploration of the data shows reasonably consistent median scores as well as spreads throughout all test levels. Rather than judge bias, this behavior indicates educated self-selection by riders. Since the averages are pretty close, it suggests that riders are choosing to compete in levels that are appropriate for their skill level so that an average performance at any level will receive about the same score.

The variable that actually shows the most variation in scores is the level of judge training. In the first few levels, the more experienced judges appeared to score much more leniently than the newer judges. Nevertheless, for these levels, it is possible that this is not a case of bias. "L" and "LD" judges are usually contracted for smaller schooling shows which attract amateur riders. If a show can afford an upper level judge like an "S" judge, they are most likely attracting riders who frequently show and are trying to make riding into more of a profession than a hobby. Therefore, "L" and "LD" judges are more likely to see the riders that are earning lower scores at those levels than "S" or even "R" judges. Once we reach the higher test levels, there is more consistency in the quality of rider. When we look at Second Level and above, we see that the more experienced the judge is, the lower they tend to score rides. Whether that is caused by the training program itself or is a byproduct of gaining experience would be areas for further study if this trend is established as significant.

CHAPTER 3 BAYESIAN MULTILEVEL MODELING

3.1 Varying Intercepts Model

When looking at scores in any sport, we expect to see some variation. In fact, diversity in scores is necessary for any type of sporting competition. The variation we expect, however, is one that allows us to see that a particular athlete is superior to another in order to determine the winner of a competition. We want to see the differences between performances and resolve that one was distinctly better than another. In scoring competitors, the ideal situation would be to have all sources of variation tied up in differences between athletes skill levels and eliminate any variation caused by the individual judging the event.

For that reason, we want to separate these possible sources of variation in our model to see where the variability in scores is truly coming from. Bayesian multilevel models allow us to make these distinctions through varying intercepts or slopes for different groups while also giving the freedom to include fixed effects like those used in regression models. For our scenario, we can use judge, rider, and horse as our varying intercepts to see how each group contributes to the overall variation. Then, we can designate region, test name, and judge rating as fixed factors to see if they significantly affect the variation associated with the different judges or not.

The simplest model we can use to see the effect of judge bias is a varying intercept model with judge as the group. While it does not look at any other contributing factors, it does give us an idea of how much of the overall variation in scores is due to judge subjectivity. We set up the overall model for the scores y_i as a normal distribution, but we also add a layer by including a distribution for our judge group as well.

When adding the distribution for judges, we need to assign what is called a prior distribution. One of the benefits of Bayesian analysis is that it allows us to use previous information to create more accurate models. If we know something about the distribution of a variable, we are able to include that information within the model and reduce the levels of uncertainty in our results. How-

ever, if we do not want to influence the results of our model we can also assign an uninformative prior which assumes we do not have any information about the distribution. For all of our models in this analysis we will use uninformative prior distributions centered around a mean of 0. Using this distribution gives us a varying intercept model of the form

$$y_i \sim N(\mu + \alpha^{judge}_{j[i]}, \sigma^2_\epsilon), \text{for ride } i = 1, ..., n$$

$$\alpha^{judge}_j \sim N(0, \sigma^2_{judge}), \text{for judge } j = 1, ..., J$$

Which yields parameter estimates of $\hat{\sigma}^2_{judge} = 2.502$ for the judge effect with an estimated residual variance of $\hat{\sigma}^2_\epsilon = 22.617$ representing an error for the variation not explained by the judge effect. The total variance in our model is the sum of these two sources of variation giving us $\hat{\sigma}^2_y = 25.119$. To find the typical range of judge scores, we can add and subtract the standard deviation $\hat{\sigma}_y = 5.012$ to our estimated average judge score of 63.25. Thus, on average we expect the typical judge score to fall between 58.2 to 68.3%.

If we focus on the size of the standard deviation, the model says that a judges average final score will typically be $63.25 \pm 5.012\%$. Whether a judge score is biased depends on where those 5.012 percentage points are coming from. If they say that a more talented rider generally scores 5 points higher than an average one, there are no issues with judge bias. However, if those percentage points say that some judges tend to score all riders 5 percentage points higher than other judges do, then we may have some judge subjectivity issues involved.

The varying intercepts model tells us that out of the total variance of 25.119, 2.502 is due to the judge effect while 22.617 is from the residual variation which is unexplained by the individual judge. That means that approximately 10% of the variation is coming from the subjectivity of the judge while 90% is coming from other factors we have not yet explored. While this does not tell us why judges are scoring the way they do, it does tell us that one tenth of the variability in a riders score is determined by which judge is scoring their ride. That means that one tenth of the variability in their score is taken out of a riders hands and is instead decided upon subjectively.

Now that we know how much judge scores tend to vary, we can also discern which judges tend to score higher and which tend to score lower. Bayesian multilevel models allow us to extract random effects for each individual in our designated groups which we can then add to the estimated average score from the overall model to find estimated average scores for each individual judge. Figure 3.1 illustrates the estimated values for each of the 130 judges and compares them with their observed average scores.

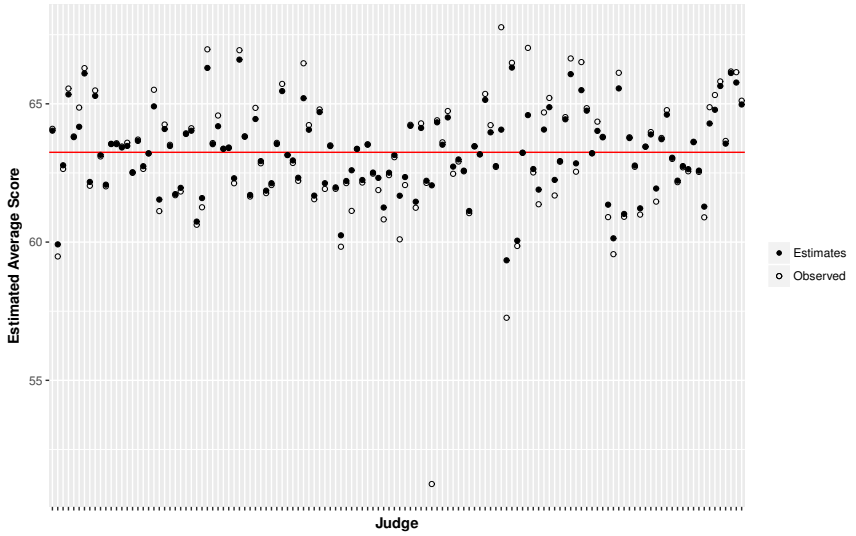

Figure 3.1: Observed Average Judge Scores vs. Estimated Average Judge Scores

In the comparison, we see that the random effects create a shrinking effect towards the overall mean in order to get better estimates even when we do not have a lot of data points. For instance, we observe a distinct outlier where Leonie Fernandes average score is 51.25%. Upon further investigation, however, Leonie Fernandes only scored one ride in our dataset from the position at C. This leaves us with very little information about how Fernandes typically scores tests since we only have one observation. Using the random effects provided by our multilevel model, we estimate that Fernandes will actually produce an average score of 62.05%. In this case, we see a lot of shrinkage towards the population mean since we did not have a lot of information about that judge.

If, however, we do have a lot of information about how a judge tends to score tests, we will not see as much shrinkage in our estimates. An example is Willette Brown who scored 469 rides.

Using those rides, we found her average score as 66.17%. Our model, however, estimates that her average score will be about 66.11%. So, we see that her estimate moved down towards our overall mean slightly, but not significantly since we already had so much information about her judging behavior that the model did not need to make a lot of changes.

While Figure 3.1 shows us a picture of how the model changed our observed averages into better predictions of behavior trends, it does not give us a lot of insight into these predictions. Figure 3.2 shows a clearer distribution of our estimated judge scores.

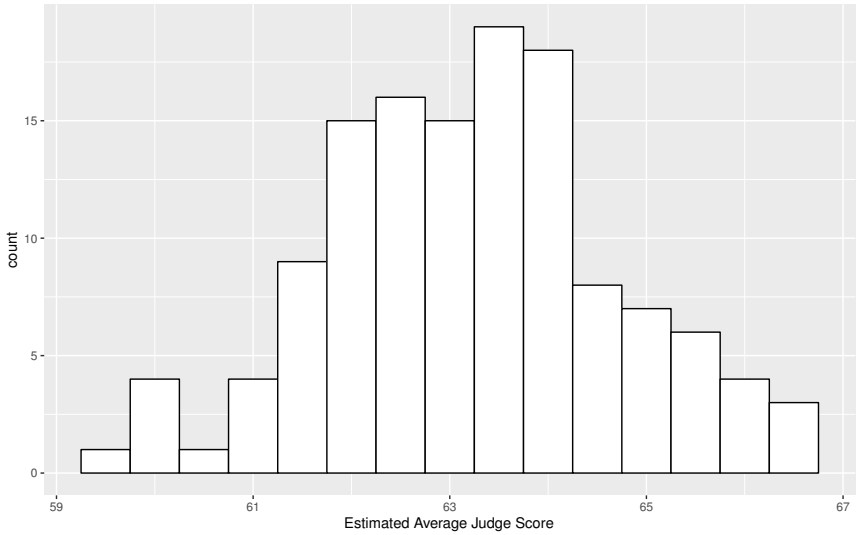

Figure 3.2: Estimated Judge Scores with One Varying Intercept

We see that most judges are predicted to award final scores around the average of 63.25%, however, there are some judges that tend to score outside the average range. Out of these unusual scores, it appears more typical for a judge to score less harshly than their peers. This suggests that if a judge is going to deviate from typical scoring methods, they prefer to give the benefit of the doubt when assigning points. While judges that score very harshly on average are much fewer in number we are still interested in why judges are making choices on their scores that remove them from the average.

3.2 Two Varying Intercepts

Since the scoring system is designed to quantify competitors skill levels, the rider should be one of the biggest factors in score variability. Ideally, when a judge watches a rider they are only focusing on that riders performance. This is not always the case, however, as gender bias, personal interest in a competitors success, or a preconceived idea of a competitors abilities based on past experience always have the possibility of arising.

With multilevel modeling, we can add a group for the individual riders in order to take into account how each rider generally performs. If one rider is a stronger competitor than another, they should exhibit higher dressage scores throughout their career and it would not indicate any sort of bias. If, however, a judge tends to score some riders differently than most other judges it might indicate some bias. Looking at the general trends here will help since a judge might score one rider differently than their average scores if they have an unusually good or poor ride. In order to visualize this behavior, we fit the model

$$y_i \sim N(\mu + \alpha^{judge}_{j[i]} + \alpha^{rider}_{k[i]}, \sigma^2_\epsilon), \text{for ride } i = 1, ..., n$$

$$\alpha^{judge}_j \sim N(0, \sigma^2_{judge}), \text{for judge } j = 1, ..., J$$

$$\alpha^{rider}_k \sim N(0, \sigma^2_{rider}), \text{for rider } k = 1, ..., K$$

Once again, we used uninformative priors for judge and rider variation. This model in turn gives us the estimated variances $\hat{\sigma}^2_{rider} = 12.668$ for the rider effect, $\hat{\sigma}^2_{judge} = 1.801$ for the judge effect and a residual variance of $\hat{\sigma}^2_\epsilon = 10.356$ which we add together for an overall model estimated variance of $\hat{\sigma}^2_y = 24.825$.

If we look at the percentages each component represents out of the total variance of 24.825 we see that 51% of the variation is due to the riders skill while 7.3% of the variation is due to judge subjectivity. This shows us that if we take rider individuality into account, the judges are not differing in their assessments as much as they were when we treated each rider the same. This

is encouraging since it means that most judges are able to recognize competitors skills and score them appropriately. Figure 3.3 illustrates the distribution of estimated average judge scores under this model.

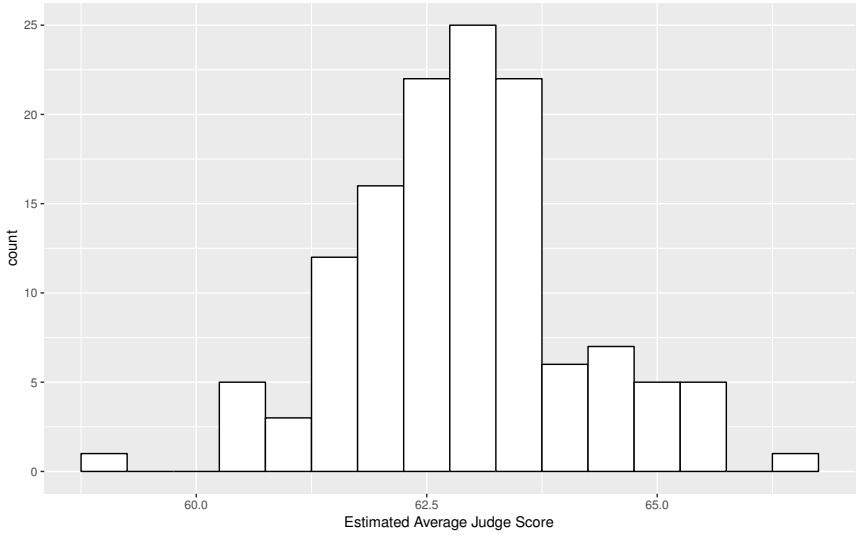

Figure 3.3: Estimated Judge Scores with Two Varying Intercepts

First, we see that the overall average has decreased to 62.9% and that our distribution looks slightly more symmetric than it did before. The most interesting behavior in this distribution, however, is that it points out very distinct outliers both above and below the mean.

If we look at the judge with the low outlier, we see that the model estimates her average score will be 59.18%. This estimate is almost 2 percentage points higher than her observed average score which leads us to ask why she is scoring so low. Examining the 17 rides she judged in our dataset, all the rides represented are Third Level tests at one particular show. We also notice that her scores have one of the largest standard deviations in our dataset as her scores have a standard deviation of $\sigma = 6.73$ while the average standard deviation in our observed data is $\sigma = 4.64$. This is because while many of those 17 rides were given fairly average scores she also awarded some very low scores that day. Whether that was due to riders having a bad day or if she was judging particularly harshly that day is unknown.

The judge with the higher than average score is a little more concerning. While all her scores were obtained at the same show, they spanned a two-day time period and contained 89 rides rep-

resenting all possible test levels. Her observed standard deviation of $\sigma = 3.98$ was also slightly below the average standard deviation in our dataset, indicating that she was pretty consistent in giving higher than average scores.

3.3 Three Varying Intercepts

Dressage as a competition is about the partnership between horse and rider. Therefore, we would be remiss to exclude the horses contribution to the equation. Just like in any other partnership, some personalities work better together than others even though exceptional skill is able to make almost any professional relationship work. Previous research has also suggested that horse breed is a significant factor in dressage judge bias. For instance, breeds that are traditionally built for dressage, such as Oldenburgs, are thought to generally score higher than a Thoroughbred or Arabian. While some would argue that an Arabian can perform just as well as an Oldenburg on a test and still score lower, it is also possible that the Oldenburg simply performed better. To test this theory, we add the horse effect to the equation giving us a model of the form

$$y_i \sim N(\mu + \alpha^{judge}_{j[i]} + \alpha^{rider}_{k[i]} + \alpha^{horse}_{l[i]}, \sigma^2_\epsilon), \text{for ride } i = 1, ..., n$$

$$\alpha^{judge}_j \sim N(0, \sigma^2_{judge}), \text{for judge } j = 1, ..., J$$

$$\alpha^{rider}_k \sim N(0, \sigma^2_{rider}), \text{for rider } k = 1, ..., K$$

$$\alpha^{horse}_l \sim N(0, \sigma^2_{horse}), \text{for horse } l = 1, ..., L$$

This is turn gives us parameter estimates of $\hat{\sigma}^2_{horse} = 5.110$ for the horse effect, $\hat{\sigma}^2_{rider} = 8.782$ for the rider effect, $\hat{\sigma}^2_{judge} = 1.799$ for the judge effect and a residual variance of $\hat{\sigma}^2_\epsilon = 9.301$ for an overall estimated variance of $\hat{\sigma}^2_y = 24.992$.

We first notice that the variation due to the judge effect did not change at all when we added the horse to the equation. This is encouraging since it suggests that the source of judge bias is not coming from the horse they are watching. Instead, it suggests that different scores for different types of horses is due to the horses ability to execute movements rather than a judge predetermining

that a particular breed will score higher or lower than another. The variation due to the rider, however, did decrease significantly from 51% of the overall variation to 35%. Therefore, while the horse does not affect judge decisions it does affect rider performance. Because the variation between riders decreased when we added the horse to the equation we see that the horse can act as a stabilizing factor. Two riders with vastly different skill levels can ride the same horse and their scores will not reveal all the disparity between their skill levels simply because the horse is able to mask some of those differences.

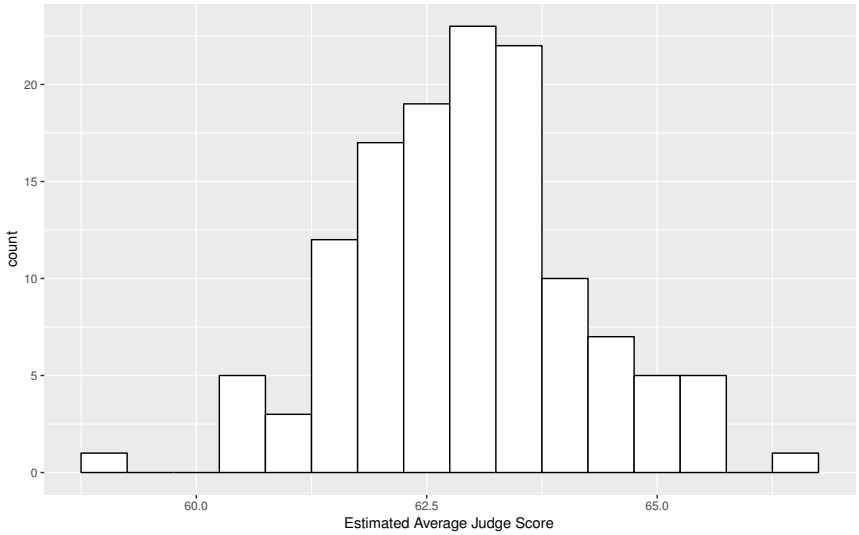

Figure 3.4: Estimated Judge Scores with Three Varying Intercepts

In Figure 3.4, both the overall average and the general shape of the distribution for judge scores stayed almost exactly where they were with the model that only included judge and rider effects. Since adding the horse did not have an impact on judge variability we would not expect to see a large change in our distribution of average judge scores. Thus, we still see the same outliers that we did in our previous model regardless of the horse the competitor rode.

3.4 Adding a Fixed Factor – Region

While the varying intercept factors in the previous sections revolved around the subjectivity and skill levels associated with individuals sometimes external factors have an impact on competitors as well. In particular, we have looked at region, test level, and judge rating as possibly influential

external factors in this analysis. While our initial exploration of the data was able to give some insight into how these factors affect the final score, we are more interested in how these factors affect the judges decisions.

First, we want to look at how region affects judge variability. While we could go back to the first model with only judge as the varying intercept, it would be more informative to include rider and horse in the model as well. When we were exploring the data, we saw that Region 7 typically produced higher scores than other regions but were not able to conclude if that difference was due to judge bias or rider skill in that area. If we include all three varying intercepts, we can get a better idea of where that difference falls.

This model will take on the form

$$y_i \sim N(\mu + \beta^{region} \cdot region_i + \alpha^{judge}_{j[i]} + \alpha^{rider}_{k[i]} + \alpha^{horse}_{l[i]}, \sigma^2_\epsilon), \text{for ride } i = 1, ..., n$$

$$\alpha^{judge}_j \sim N(0, \sigma^2_{judge}), \text{for judge } j = 1, ..., J$$

$$\alpha^{rider}_k \sim N(0, \sigma^2_{rider}), \text{for rider } k = 1, ..., K$$

$$\alpha^{horse}_l \sim N(0, \sigma^2_{horse}), \text{for horse } l = 1, ..., L$$

and yields the parameter estimates $\hat{\sigma}^2_{horse} = 5.037$ for the horse effect, $\hat{\sigma}^2_{rider} = 8.610$ for the rider effect, $\hat{\sigma}^2_{judge} = 1.858$ for the judge effect and a residual variance of $\hat{\sigma}^2_\epsilon = 9.297$ which we add together for an overall model estimated variance of $\hat{\sigma}^2_y = 24.802$.

If we look back at our base model with three varying intercepts, we see very little change in variation. This implies that the region has little to no impact on judge bias, on the quality of the rider, or on the quality of the horse. In fact, it also had very little impact on our residual standard deviation meaning that the region offers us hardly any information on the final score at all. Although the parameter estimates for our random effects do not show much impact from the regional effects, there are a few distinctions which show up in the parameter estimates for the fixed effects.

Table 3.1: Estimates for Fixed Factor – Region

Fixed Effect	Estimate	Std. Error	t value
Intercept	62.77	0.2448	256.45
Region 2	0.1942	0.2596	0.75
Region 3	0.8539	0.5706	1.50
Region 4	0.9219	0.3901	2.36
Region 5	0.1897	0.3760	0.50
Region 7	2.6327	0.4650	5.66
Region 8	-0.2593	0.2479	-1.05

In Table 3.1, the estimates for β^{region} reveal how much each region is expected to deviate from the model intercept. The intercept estimate is the expected average score for the omitted region, Region 1, and all the other estimates reveal how much higher or lower each regions score is expected to fall compared with Region 1. These estimates reveal some slight grouping in the regions as Region 2, Region 5, and Region 8 do not differ that much from the estimate for Region 1, Regions 3 and 4 have very similar values, and Region 7 exhibits an estimated average score much higher than any other region. When conducting the exploratory analysis, Figure 2.2 showed a strong distinction in the Region 7 scores. Looking at the t-values for each region, Region 7 does show the most significance with the largest t-value at 5.66. Since adding regions to the model did not affect our parameter estimates for group variations, however, this model does not give us much insight as to why that particular region is showing higher scores.

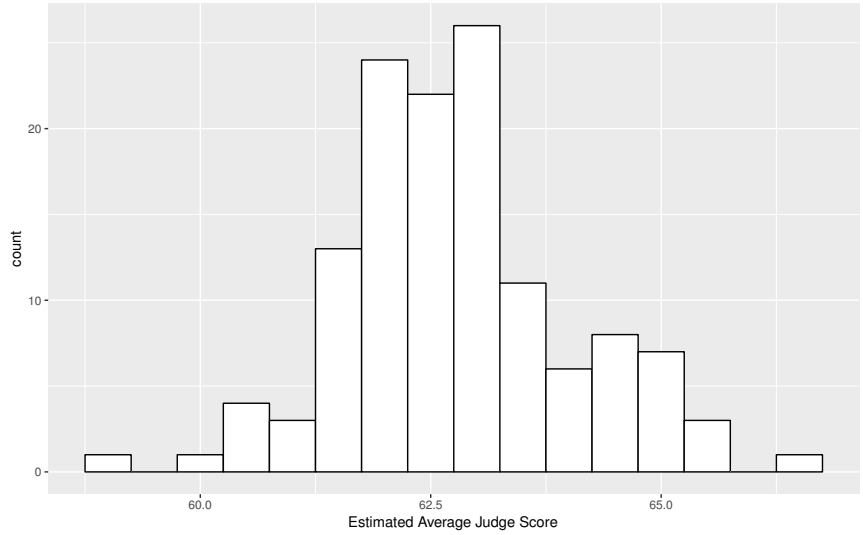

Figure 3.5: Estimated Judge Scores After Adding Region

The shape of the distribution for the estimated average judge scores does show slight changes. We now see a stronger concentration of scores from 62 to 63% and more diversity in the less frequent average score values. Those two changes cancel each other out when recalculating the variance which is why we do not see any significant changes to judge score variation upon adding region to the model. So, Figure 3.5 implies that region does have an impact on our estimated average judge scores since adding any variable to our model will change our estimates slightly, however, that impact does not affect how much the scores are varying overall.

3.5 Adding a Fixed Factor – Test Level

Another factor that has often been pointed to as a source of judge bias is the test level. Previous research argues that judges tend to score particular tests differently. While we did not see much evidence supporting this claim in our exploratory data analysis, with all the previous research indicating that this behavior exists it is still worthwhile to model the relationship. Adding test name as a fixed factor and maintaining uninformative priors for the grouping factors would create a model on the form

$$y_i \sim N(\mu + \beta^{test} \cdot test_i + \alpha^{judge}_{j[i]} + \alpha^{rider}_{k[i]} + \alpha^{horse}_{l[i]}, \sigma^2_\epsilon), \text{for ride } i = 1, ..., n$$

$$\alpha_j^{judge} \sim N(0, \sigma^2_{judge}), \text{for judge } j = 1, ..., J$$

$$\alpha_k^{rider} \sim N(0, \sigma^2_{rider}), \text{for rider } k = 1, ..., K$$

$$\alpha_l^{horse} \sim N(0, \sigma^2_{horse}), \text{for horse } l = 1, ..., L$$

Creating this model gives us variance estimates of $\hat{\sigma}^2_{horse} = 4.450$ for the horse effect, $\hat{\sigma}^2_{rider} = 10.185$ for the rider effect, $\hat{\sigma}^2_{judge} = 1.851$ for the judge effect and a residual variance of $\hat{\sigma}^2_{\epsilon} = 8.775$ which we add together for an overall model estimated variance of $\hat{\sigma}^2_y = 25.261$.

Once again, the variation associated with the judge effect hardly changed at all. The variation for horse and rider, however, do change when we include test level in the model. The change implies that if we see differences in average scores for different test levels it is not due to judges altering their assessment practices for separate levels but is more likely due to changes in competitors abilities. It is also interesting to note that a lot of the responsibility for variation in score is shifted to the rider when we consider the test level. Often, a horse is capable at competing at higher levels of competition, but because their rider is inexperienced the horse is not able to perform to their full potential. Unless the rider knows how to ask their horse to perform certain movements the horse will not execute patterns as well as they could in the hands of another rider. Thus, when there are significant differences in scoring patterns for the test levels, it is more likely because certain types of riders are choosing to compete at those levels than that judges are basing their decisions off preconceived ideas about that test.

The baseline test in this model is First Level 1. Unsurprisingly, the other First Level tests' β^{test} parameter estimates and t-values indicate that their average scores do not vary that much from test 1. More interesting is the fact that the higher levels, Second Level, Third Level, and Fourth Level, all show average scores significantly lower than the average First Level scores. Conversely, the lower levels, Intro and Training Level, both indicate significantly higher average scores than the First Level tests.

Comparing these results with the variance estimates from the random effects suggests that riders who elect to compete at lower levels tend to score higher than riders choosing to undertake

Table 3.2: Estimates for Fixed Factor – Test Level

Fixed Effect	Estimate	Std. Error	t value
Intercept	62.78	0.1948	322.3
First 2	0.4945	0.1643	3.0
First 3	0.1194	0.1527	0.8
Fourth 1	-1.7647	0.2606	-6.8
Fourth 2	-1.5595	0.3215	-4.9
Fourth 3	-2.4932	0.3265	-7.6
Intro A	2.8300	0.4023	7.0
Intro B	2.3969	0.3539	6.8
Intro C	1.6238	0.3502	4.6
Second 1	-1.1801	0.2142	-5.5
Second 2	-1.7822	0.2625	-6.8
Second 3	-1.4234	0.2066	-6.9
Third 1	-0.9408	0.2255	-4.2
Third 2	-1.5291	0.3179	-4.8
Third 3	-0.8656	0.2137	-4.0
Training 1	1.7133	0.1986	8.6
Training 2	1.9035	0.1814	10.5
Training 3	1.0127	0.1551	6.5

more difficult tests. This behavior appears consistent across all judge scores, however, indicating that the disparity between scores based on test level is due to test difficulty rather than judge leniency. Lower level tests were designed as introductions to the art of dressage and therefore simply do not provide as much room for critical judging as more complex tests.

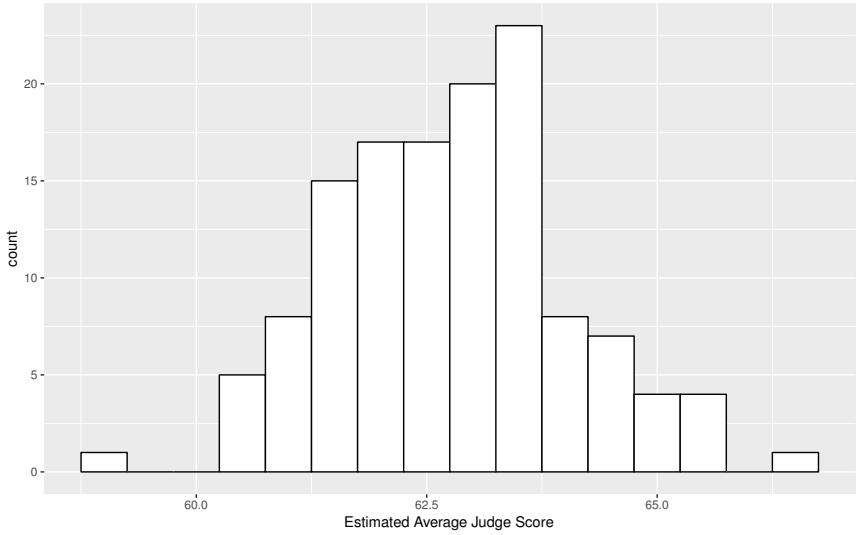

Figure 3.6: Estimated Judge Scores After Adding Test Level

Figure 3.6 shows the distribution of estimated judge score averages after adding test level to the model and our estimates actually look more uniform than they had previously. It is not a huge change, but it does appear that adding the test level as a factor in our model allows slightly more consistency in judge scoring. This further indicates that although tests typically receive different scores, these scores are coming from all judges and not just a subset of biased individuals.

3.6 Adding a Fixed Factor – Judge Rating

Out of all the fixed factors we considered in our analysis, the judges level of training emerged with the most potential for impact on a judges scoring tendencies. We saw distinct differences in how types of judges were scoring the tests based on their certification level so this variable offers the most possibility of introducing bias into judge scores. We use the same process as before to create the model

$$y_i \sim N(\mu + \beta^{rating} \cdot rating_i + \alpha^{judge}_{j[i]} + \alpha^{rider}_{k[i]} + \alpha^{horse}_{l[i]}, \sigma^2_\epsilon), \text{for ride } i = 1, ..., n$$

$$\alpha^{judge}_j \sim N(0, \sigma^2_{judge}), \text{for judge } j = 1, ..., J$$

$$\alpha^{rider}_k \sim N(0, \sigma^2_{rider}), \text{for rider } k = 1, ..., K$$

$$\alpha_l^{horse} \sim N(0, \sigma^2_{horse}), \text{for horse } l = 1, ..., L$$

which yields variance estimates of $\hat{\sigma}^2_{horse} = 5.103$ for the horse effect, $\hat{\sigma}^2_{rider} = 8.818$ for the rider effect, $\hat{\sigma}^2_{judge} = 1.634$ for the judge effect and a residual variance of $\hat{\sigma}^2_{\epsilon} = 9.298$ which we add together for an overall model estimated variance of $\hat{\sigma}^2_y = 24.853$.

Adding the judge certification does not affect the variations associated with horse and rider. Since the level of judge training has no relationship with any of the athletes skills this is exactly what we expected to see. The more interesting aspect of this model is the decrease in judge score variability. If we include judge rating in the model, the judge effect accounts for 6.6% of the overall variation a decrease of 0.6% from our original three varying intercepts model.

The decrease in variation implies that more experienced judges are going to score rides differently than a newer judge. When we group the judges by their experience level, we see less variability within the groups than we saw when we treated every judge as if they had the same level of training.

Table 3.3: Estimates for Fixed Factor – Judge Rating

Fixed Effect	Estimate	Std. Error	t value
Intercept	63.19	0.4761	132.71
LD	-0.1335	0.6183	-0.22
r	0.7241	0.6377	1.14
R	0.3349	0.5744	0.58
S	-0.5946	0.5015	-1.19

While adding judge experience to the model reduced the overall variation in judge scores, there is a reasonable amount of variation within each group. The smallest standard error is found in the group for "S" judges at 0.5015 which is on the higher end of the standard deviations compared with those for region and test level. We also notice that none of the groups have significantly different parameter estimates. Thus, even though there is some variation in our models random effects for judge scores and some variation within each group of judge type, these differences still do not emerge as significant.

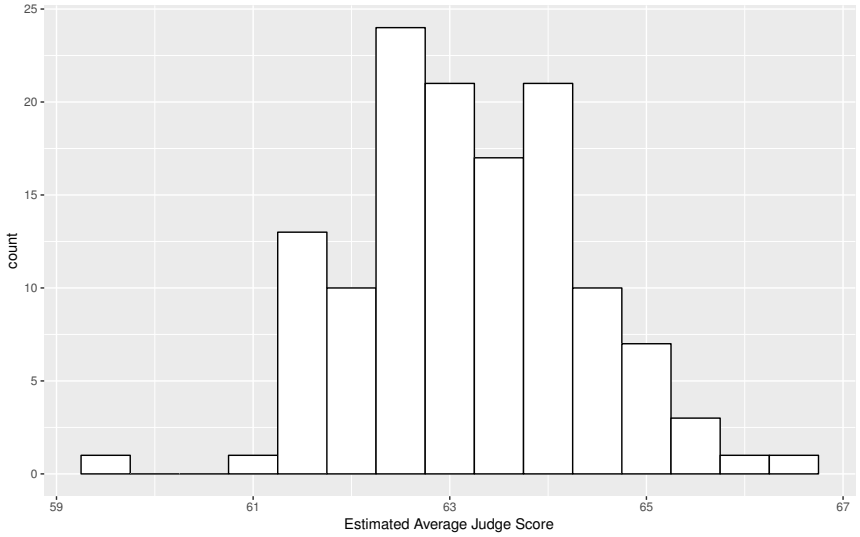

Figure 3.7: Estimated Judge Scores After Adding Judge Rating

Along with the smaller spread, the upper outlier of 66.4% no longer looks as unusual as it did before. While that judges estimated score is still quite a bit higher than the average estimate, there are now other judges that have estimated average scores almost as high. Although no other judges cross the 66% threshold, three of them do have average score estimates over 65.5%. With less uncharacteristic behavior in the distribution, it generates more confidence in our estimated results.

CHAPTER 4 CONCLUSIONS

4.1 Interpreting the Results

While previous research indicated that region, test level, and horse breed were significant factors in assessing judge bias, this analysis did not find that those particular factors affected judge decisions. In fact, introducing the individual rider into the equation decreased judge variability so that riders were consistently earning scores appropriate for their skill level. Furthermore, adding the horse to the equation had almost no effect on judge variability offering no evidence that the breed of a horse will have any impact on judging decisions.

Since so much research emphasized the significance of fixed factors like region and test level in dressage scoring, it is impossible to deny that outside factors can influence a riders final score. The question we wanted to address was whether the impact of these factors affected how the judge viewed a ride or affected the quality of the ride itself. The first fixed variable we explored, region, not only did not appear as a significant influence on judge decisions but did not seem to affect horse or rider variability either. While the parameter estimates showed significant differences between regions and Region 7 produced significantly higher mean scores than any other region, the multilevel models did not indicate that this difference was due to effects from judge, horse, or rider. That leaves very little information known about why this abnormality appears to occur. However, the fact that the model does not indicate the differences have any relationship with judge subjectivity is encouraging for the dressage judging system.

Test level also did not influence judge score variability in these models. Dressage judges tended to score in the same ranges whether they were watching beginning riders or advanced ones. Including test level in the model did reveal, however, that it affects horse and rider performance. Thus, we do see a difference in scores due to test level, but the evidence suggests that this is due to competitors skill levels rather than any subjectivity of the judge. Test level also places more emphasis on the riders skill instead of the horses ability as a more experienced rider is able to execute

a better performance regardless of the horse they are riding.

In our analysis, we introduced the fixed variable of judge certification to the conversation. Throughout all our models, the judges experience level was the only fixed variable we found to influence judge decision variability. If we group judges based on the certification they hold we are able to minimize the variability in scores due to judge subjectivity. While the ideal situation would have all judges scoring riders the same, this result does indicate that the USEF judge training program is producing results. As a judge receives more training in the field of dressage, their perceptions on how to score rides is continually evolving. With this information, program coordinators can become more aware of how certain types of judges will tend to score rides and perhaps even refine their training programs to better educate newer judges and minimize the discrepancies between types of judges mindsets.

When researchers analyze the differences in test scores based off these fixed effects, they want to answer the question of whether or not judge bias exists. Once they find variation in scores based off a fixed variable they start to claim judge bias. The purpose of this analysis was to look behind the initial glimpse of variation in scores and ascertain a better idea of the true cause. While it may be true that dressage scores are significantly higher in Region 7, it would be presumptuous to claim that judges in Region 7 award higher scores than judges in other regions. Not accounting for the variability in individual riders abilities leaves an incomplete picture of how scores are earned and places more power in judges hands than actually exists. This study used multilevel modeling in order to include individuality of both judges and riders in the model. By including individuality, we were able to estimate how each judge would typically score a ride and identify judges that scored unusually high or low on average. Other than one or two judges, however, this analysis did not reveal a large amount of variation in scores due to a judge effect. In fact, the largest portion of variation in dressage scores was due to the rider effect. That leaves the responsibility for earning high scores exactly where it should be in the hands of the rider.

4.2 Future Research

Even with the variation we were able to explain in this modeling process, there were certain aspects that further research could develop. One of these would be collecting more data on judge characteristics. Variables like gender, age, years of experience, and number of shows judged could all add insight into fixed effects that may influence judge subjectivity. Do male and female judges score the same? Does a judge who recently received their "S" certification score the same as someone who has held the same certification for twenty years? How do judges scores change as they gain show experience? While collecting this data could take more time, the results would introduce much greater insight into judging decisions.

Although this study focused on judging patterns in the overall score, every final score is compiled from many movement scores. Horse Show Office provides the movement scores from every test on their website. Using that information could highlight if certain movements have higher variability in how judges are scoring which could lead to more concentrated training on those passages when judges are going through certification programs.

There will always be some variability in judge scores no matter how rigorous the training program is that the judges complete. The ultimate goal is to reduce variation due to judge subjectivity so that the riders skill is the most important aspect of the competition. With more analysis exploring sources of variation, new training programs can be developed to combat undesired scoring behaviors until variability from the judge effect is minimized.

BIBLIOGRAPHY

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Deuel, N. and R. Russek-Cohen (1995, 11). Scoring Analysis of Three World Championship Three-Day Events. *Journal of Equine Veterinary Science 15*(11), 479–486.

Diaz, A., M. Johnston, J. Lucitti, W. Neckameyer, and K. M. Moran (2010, 01). Scoring Variables and Judge Bias in United States Dressage Competitions. *Journal of Quantitative Analysis in Sports 6*, 13–13.

Eurodressage (2008, 11). FEI President Princess Haya Asks for Resignation of FEI Dressage Committee. *Eurodressage*.

Federation Equestre Internationale (2009, 10). Report of the FEI Dressage Task Force.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Whitaker, T. and J. Hill (2005a). A Study of Scoring Patterns at National Level Eventing Competitions in the UK. *Equine and Comparative Exercise Physiology 2*(3), 171183.

Whitaker, T. and J. Hill (2005b). Dressage Scoring Patterns at Selected British Eventing Novice Events. *Equine and Comparative Exercise Physiology 2*(2), 97104.

Zupanc, K. and E. Strumbelj (2018, 04). A Bayesian Hierarchical Latent Trait Model for Estimating Rater Bias and Reliability in Large-Scale Performance Assessment. *PLoS ONE 13*(4).

APPENDIX A SELECTED R PROGRAMS

All models in this analysis were created using the statistical software R (R Core Team, 2017)). In order to create the multilevel models, we used the lmer function found in the package lme4 created by Bates, Mächler, Bolker, and Walker (2015)). The lmer function fits a linear mixed effects model meaning that it includes both fixed and random effects. In this section, we have included the code and output for some of the models explored in this paper.

.1 One Varying Intercept Model

The first model we created was the simplest model. It only includes the random effect for judges which it treats as a varying intercept and generates estimated average scores for each individual judge.

```
Model1 <- lmer(C~(1|JudgeC))
```

A summary of the output for this model yields variances and standard deviations for the random effects. Since we did not add any fixed effects, the only fixed effect in our output is the intercept which is the parameter estimate for μ.

```
Linear mixed model fit by REML ['lmerMod']
Formula: C ~ (1 | JudgeC)

REML criterion at convergence: 74379.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.8331 -0.6261  0.0206  0.6548  3.9900

Random effects:
```

```
 Groups   Name        Variance Std.Dev.
 JudgeC   (Intercept)  2.502   1.582
 Residual             22.617   4.756
Number of obs: 12440, groups:  JudgeC, 130

Fixed effects:
            Estimate Std. Error t value
(Intercept)  63.2463     0.1518   416.8
```

.2 Two Varying Intercepts Model

The second model we created included both judge and rider as varying intercepts.

```
Model2 <- lmer(C~(1|JudgeC) + (1|Rider))
```

The output from this model is also fairly straightforward. We have simply added another random effect and grouping which have their own parameter estimates. Therefore, we can now extract estimated mean values for every judge and for every rider.

```
Linear mixed model fit by REML ['lmerMod']
Formula: C ~ (1 | JudgeC) + (1 | Rider)

REML criterion at convergence: 69363.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.0985 -0.5370  0.0237  0.5650  4.1323

Random effects:
```

```
 Groups   Name        Variance Std.Dev.
 Rider    (Intercept) 12.668   3.559
 JudgeC   (Intercept)  1.801   1.342
 Residual             10.356   3.218
Number of obs: 12440, groups:  Rider, 2842; JudgeC, 130

Fixed effects:
            Estimate Std. Error t value
(Intercept)   62.897      0.145   433.7
```

.3 Adding a Fixed Effect

Some of the later models also included fixed effects. This will increase the amount of terms in the second table and will also generate a correlation matrix. As an example, we will show the code and output for the very last model we created.

```
M6 = lmer(C ~ factor(JudgeCRating) + (1|JudgeC)+(1|Horse)+(1|Rider))
```

When writing this model, we need to note that the fixed factor is a categorical variable. The function “factor causes the model to treat judge rating as if we had created columns of coded variables. For numeric variables, simply including the variable name in the model is sufficient.

```
Linear mixed model fit by REML ['lmerMod']
Formula: C ~ factor(JudgeCRating) + (1 | JudgeC) + (1 | Horse) + (1 |      R

REML criterion at convergence: 68878.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.4674 -0.5265  0.0236  0.5597  4.2424
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 Horse    (Intercept) 5.103    2.259
 Rider    (Intercept) 8.818    2.969
 JudgeC   (Intercept) 1.634    1.278
 Residual             9.298    3.049
Number of obs: 12440, groups:  Horse, 3102; Rider, 2842; JudgeC, 130

Fixed effects:
                       Estimate Std. Error t value
(Intercept)             63.1889     0.4761  132.71
factor(JudgeCRating)LD  -0.1335     0.6183   -0.22
factor(JudgeCRating)r    0.7241     0.6377    1.14
factor(JudgeCRating)R    0.3349     0.5744    0.58
factor(JudgeCRating)S   -0.5946     0.5015   -1.19

Correlation of Fixed Effects:
            (Intr) f(JCR)L fc(JCR) f(JCR)R
fctr(JCR)LD -0.704
fctr(JdgCR) -0.736  0.525
fctr(JdCR)R -0.816  0.585   0.615
fctr(JdCR)S -0.938  0.668   0.705   0.782
```

Including a fixed factor yields a lot more output in our model summary. In addition to the parameter estimates for the variance components of our random effects, we also have estimates for each factor of the fixed effect categorical variable. As fixed effects, they also provide t-values to specify how significant each level is for the model. The summary also includes a correlation

matrix for the fixed effects which can indicate any multicollinearity problems.