Measuring Judge Agreement for a High School Diving Competition

Monnie McGee* Jing Cao†

Abstract

In a high school championship diving competition, athletes are asked to perform 11 dives. Each of 5 judges gives each dive a rating from 0, for a failed dive, to 10 for a perfect dive. Round scores are determined by dropping the lowest and highest judge's rating, adding the remaining three ratings, and multiplying that sum by the degree of difficulty of the dive. The total score is determined by the sum of the round scores. Agreement in the context of multiple raters has been well explored in the literature. However, measuring agreement in the context of multiple raters, with multiple scores per rater, and with covariates influencing the ratings, has been much less explored. Data from a regional high school diving competition has all of these components. We examine various measures of interrater agreement from the literature, and discuss the shortcomings of each in the context of this data set.

Key Words: Interrater agreement, multiple raters, Fleiss's kappa, Kraemer's kappa, Ordinal data

1. Introduction

Agreement among judges in athletic competitions has been discussed extensively in the literature (Pajek, *et. al.*, 2013) and in the media (Easterbrook, 2004, Clarke, 2014 and Associated Press, 2002). Most of the attention has been on gymnastics and figure skating. Diving is another sport where awards are given based on ratings from a panel of judges. National bias in Olympic diving events has been investigated in Emerson, *et. al.*, 2000.

The data for this study consist of scores from a regional high-school diving competition. Each dive is rated from 0, which means a failed dive, to 10, which is a perfect dive, by 5 judges, all of whom are high school coaches for the teams represented in the competition. In reality, it is rare for a dive to be rated below a 3.0 or above a 7.0 by any judge. Scores typically are given in increments of half points, for example, A score of 5.0 or 5.5 is a valid score, but a score of 5.3 is not. Effectively, the scoring by half points results in 21 possible ratings that can be given to a diver for a given dive. Even there are are many possible scores, the mechanism is ordinal, because a difference of a point (or a half point) does not necessarily indicate the same difference in dive quality for all points on the scale.

Thirteen girls and 12 boys participated in the competition from which the data for this study come. In high school championship competitions, each diver performs 11 dives. Divers must perform one "voluntary" dive from each of the five categories (forward, backward, inward, reverse, and twist), five optional dives (one from each category), and a sixth dive that can be from any category (Franklin, 2018). The diver has a choice of which dives to perform and the order in which to perform them; however, 11 dives in the prescribed categories must be performed by each diver. In addition, each dive has a degree of difficulty associated with it, ranging from 1.3 to 2.8 for the dives performed in this competition. These data will be used to illustrate the calculation and interpretation of common measures of interrater agreement.

*Department of Statistical Science, Southern Methodist University, 3225 Daniel Ave Room 144 Heroy, Dallas, TX, 75275-0332

†Department of Statistical Science, Southern Methodist University, 3225 Daniel Ave Room 144 Heroy, Dallas, TX, 75275-0332

Figure 1: Ratings for female divers

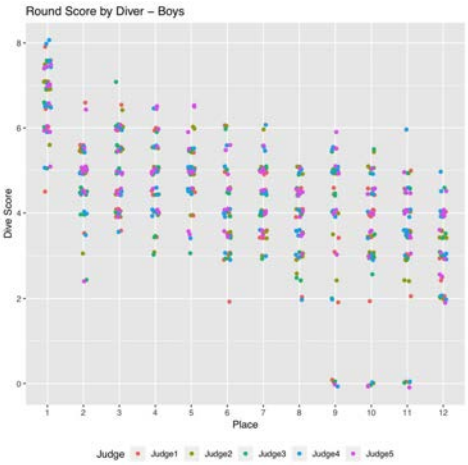

Figure 2: Ratings for male divers

Before examining numerical measures of any kind, it is a good idea to plot the data. Figures 1 and 2 show the judges' scores for each dive on the y-axis by the finishing place on the x-axis. Each dot represents a dive, each column of dots is a diver, and the color of a dot indicates a judge. The points are jittered so that individual points are easier to see. The chart for the girls is on the left, and the chart for the boys is on the right. For both groups, we see that Judge 1 (orange dot) tends to award lower scores than the other judges. We also see more spread in the girls' scores than in the boys' scores. On average, the boys tend to have higher scores than the girls; however, the plot indicates generally good agreement among the judges for each diver.

Now that we have an idea of what to expect from the data, we can examine numerical ratings to see whether their values support what was seen in Figures 1 and 2.

2. Measures of Agreement

There are many ways to measure agreement in the literature. The first statistic for measuring agreement that most researchers consider is Cohen's kappa. It measures agreement between two raters where the rating scale is nominal. For the diving data, we have five raters who are rated on an ordinal scale. Some have used Cohen's kappa for multiple raters by computing agreement between all pairwise combinations of raters; however, this often leads to nonsensical results (Fleiss, 1971). Although Cohen's kappa is inappropriate for the diving data, it's formulation is given in (1) for purposes of completeness.

Suppose that there are 2 judges who rate n subjects into m mutually exclusive and exhaustive categories. Let $p_{i\cdot} = \sum_{j=1}^{m} p_{ij}$ be the proportion of subjects in the i^{th} row and $p_{\cdot j} = \sum_{i=1}^{m} p_{ij}$ denote the proportion of subjects in the j^{th} column. Then Cohen's kappa is given by

$$\hat{\kappa}_c = \frac{p_0 - p_c}{1 - p_c} \tag{1}$$

where $p_o = \sum_{i=1}^{m} p_{ii}$ and $p_c = \sum_{i=1}^{m} p_{i\cdot} p_{\cdot i}$. Cohen's kappa ranges from zero to one, where $\hat{\kappa} = 0$ indicates no detectable agreement and $\hat{\kappa} = 1$ indicates perfect agreement.

Fleiss (1971) developed a kappa-like statistic for use when a constant number of m raters is randomly sampled from a larger population of possible raters. Let K_{ij} be the number of raters who assigned the i^{th} subject to the j^{th} category, $i = 1, \ldots, n, j = 1, \ldots, m$,

and define $p_j = \frac{1}{K_n} \sum_{i=1}^{n} K_{ij}$. Fleiss's Kappa is given by

$$\hat{\kappa}_f = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} K_{ij}^2 - K_n[1 + (K+1)] \sum_{j=1}^{m} p_j^2}{nK(K-1)(1 - \sum_{j=1}^{m} p_j^2)} \tag{2}$$

This version of the formula comes from Banerjee, *et. al.*, 1999. Just as with Cohen's kappa, Fleiss's kappa is meant for measuring agreement between raters who rate objects into nominal categories. Furthermore, it ranges from 0 to 1, with 0 indicating no agreement beyond chance and 1 indicating perfect agreement.

Another measure of agreement is the intraclass correlation (ICC). There are many different formulations of the ICC, depending on whether subject and/or rater effects are fixed or random and the number of raters (Shrout and Fleiss, 1979). The ICC chosen for use on these data is ICC_3, which is intended for multiple raters where the raters and subjects are fixed effects. Typically, the ICC is also used when the rating scale is nominal.

Let k be a fixed number of raters who judge n subjects. Then, ICC_3 is given by

$$\frac{MSB - MSE}{MSB + (nk-1)MSE} \tag{3}$$

Fleiss's kappa and the ICC are a good place to start when measuring agreement. In general, the closer these values are to 1, the more agreement there is among judges. Both of these measures are intended for the situation where the number of raters is greater than 2, the scale is nominal, and each subject is rated only once. These measures have been used for ordinal data with reasonable results (Agresti, 1988; Maclure and Willett, 1987). The diving data are not only ordinal, but also each participant performs 11 dives. Therefore, each subject is rated by all of the judges 11 times.

One way to examine agreement using Fleiss's kappa and the ICC is to treat the different rounds as if they were simply different ratings. In other words, there would be $13 \times 11 = 143$ ratings or each female diver and $12 \times 11 = 132$ measures for each male diver. Essentially, this method treats each round as if it were a separate diver, and not the same diver performing a different dive.

Another way to examine the data, that preserves the fact that each diver performs 11 dives, would be to compute measures of agreement for each round. Doing so would allow the analyst to see patterns that may not be visible were all of the ratings treated simultaneously. Figures 3 and 4 shows Fleiss's kappa (left) and ICC3 (right) computed for each round of the diving competition. For each score, the agreement of the 5 judges in each of the 11 rounds was calculated, and the round is plotted on the x-axis.

One thing to notice about these measures is that the scales are quite different. Fleiss's kappa scores range from 0.1 to 0.4, and ICC scores range from 0.9 to 0.99. Part of the discrepancy can be explained by the underlying assumptions of the measurements. The ICC is calculated under the assumption that the ratings for each subject are uninfluenced by rater bias. Fleiss's kappa, on the other hand, is not influenced by this assumption (Banerjee *et. al.*, 1999). In addition, many authors have pointed out shortcomings in both Fleiss's kappa and ICC that could contribute to a difference in the scales (Kraemer, 1980; Davies and Fleiss, 1982).

The purple solid line in Figures 3 and 4 represents scores per round for females, and the green dotted line represents scores per round for males. For both measures, we see that agreement for females is more variable and typically lower than that for males, particularly in the case of the ICC. It is not clear why this is so, or if the same variability would be seen in other competitions. However, it would be interesting to discuss possible variables that could result in such differences, and how one might go about gathering data to test any hypotheses.

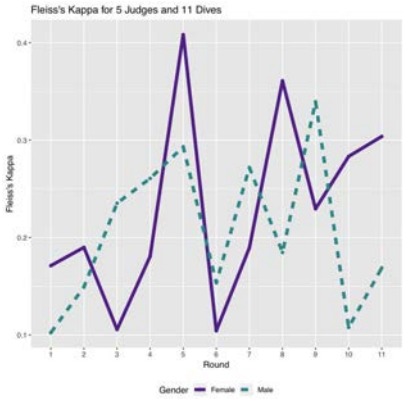

Figure 3: Fleiss's Kappa measure of agreement for each round of 11 dives

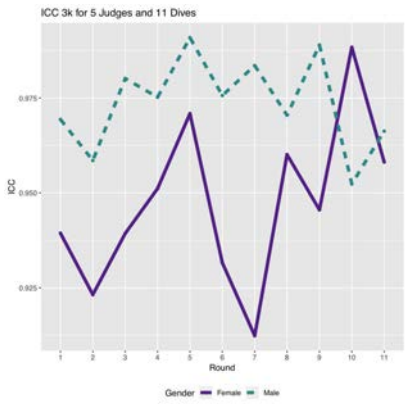

Figure 4: ICC measure of agreement for each round of 11 dives

Kraemer (1980) developed a measure of agreement when there are multiple raters who examine multiple measures on each subject. She also allowed for a subject to be classified into more than one category by a single rater, and her coefficient can be used for ordinal categorical data. Kraemer's coefficient can be calculated using the formula (Arbuckle, 2012).

$$\hat{\kappa}_0 = \frac{r_1 - r_T}{1 - r_T} \tag{4}$$

where r_1 is the unweighted average of the intrasubject correlation coefficients (computed using Spearman's rank correlation coefficient) and r_T is the average Spearman rank coefficient among all pairs of observations in the sample. Unfortunately, there does not seem to be an implementation in R of Kraemer's statistic. Therefore, we do not consider it further for the purposes of this paper.

3. Modeling Patterns of Agreement and Disagreement

As we can see, the diving data set has many features that make measuring agreement with "classical" methods somewhat unsatisfactory. A partial list is given below.

- For most data sets, each rater (judge) rates the same set of subjects *only once*. The diving data set contains 11 ratings from each of 5 judges.
- Diving data set ratings are ordinal.
- Dives are done in different orders for different divers.
- Not every diver performs the same dives.
- Covariates such as degree of difficulty and seed may affect ratings.

Given all of these intricacies of the data, it seems that a single number that measures agreement may be insufficient. Many researchers have examined how to model patterns of agreement, such as chance versus beyond-chance components. Log-linear models to compare patterns of agreement among observers or for different values of covariates have been investigated (Tanner and Young, 1985; Graham, 1995), as well as latent-class models ((Aickin, 1990; Uebersax and Grove, 1990; Agresti,1992), and agreement plus linear-by-linear association (Agresti, 1988).

More recently, Nelson and Edwards (2015) present models for agreement of data with multiple raters where the ratings are on an ordinal scale. They apply their measure to radiologists' assessment of the stage of cancer. The model that they use is a probit model, where a statistic for agreement is calculated from the cumulative probability score of a measure being classified in category c or lower.

In their formulation, each of J experts, $J = 1, \ldots, j$ independently assigns each of I subjects, $i = 1, \ldots, I$ to one of C ordinal categories, $c = 1, \ldots, C$, yielding classifications $Y_{ij} = c$. A GLMM for modeling agreement with ordinal classifications calculates the cumulative probability of a score being classified into category c or lower:

$$\Phi^{-1}(Pr(Y_{ij} \leq c|u_i, v_j)) = \Phi\left[\alpha_c - (u_i + v_j)\right] \tag{5}$$

An observed agreement statistic κ_m is given by a linear transformation of ρ_c and ρ_0, rescaled to be between 0 and 1. Formally, $\kappa_m =$

$$\left(\frac{C}{C-1}\right)\int_{-\infty}^{+\infty}\sum_{c=1}^{C}\left[\Phi\left(\frac{\Phi^{-1}(\frac{c}{C}-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)-\Phi\left(\frac{\Phi^{-1}(\frac{c-1}{C})-z\sqrt{\rho}}{\sqrt{1-\rho}}\right)\right]^2\phi(z)dz-\frac{1}{C-1} \tag{6}$$

where z is a $\mathcal{N}(0,1)$ random variable. The form for ρ, ρ_c, ρ_0 and their relationships are given in Nelson and Edwards (2015), along with a proof in Appendix 3 of their paper.

A Bayesian ordinal model has been developed in the context of multiple judges for wine tasting (Cao, 2014). The assumption for this model is that there is an underlying continuous measurement of each subject's ability, denoted as θ_i. Let y_{ij} the observed ordinal score assigned by judge j on subject i. $y_{ij} = s$ if $c_{s-1} < x_{ij} < c_s$, where c_{s-1} and c_s are category cutoffs for score s. Then the continuous estimate of the i^{th}'s subject's ability is

$$x_{ij} = \alpha_j + \beta_j\theta_i + e_{ij} \tag{7}$$

where $e_{ij} \sim \mathcal{N}(0, \sigma_j^2)$ is the judgment error made by judge j on the rating of subject i. The Bayesian Ordinal Model allows for determination of whether bias, discrimination ability, or random variation are responsible for disagreement among judges. Here, bias (α_j) measures whether a judge is relatively stringent, neutral, or generous with regard to other judges. Discrimination (β_j) measures a judge's ability to distinguish between good and bad dives. Variation measures the amount of judgment error (σ_j^2). Judge performance can be evaluated based on the correlation between his/her assigned ratings and the estimates of latent quality, θ_i.

For the remainder of this paper, we examine the use of the probit model and the Bayesian ordinal model on the diving data.

4. Results

Neither of these models as they currently stand has a place for covariates; therefore, we ignored the degree of difficulty for each dive and modeled males and females separately. In

Table 1: Measures of Agreement based on the Nelson and Edwards Probit Model

Statistic	Males	Females
$\hat{\rho_0}$	0.537 (0.031)	0.331 (0.026)
$\hat{\rho_c}$	0.176	0.204
κ_m	0.066 (0.006)	0.035 (0.003)

addition, we modeled each round as if it were a separate diver, leading to 132 observations for males and 143 for females. There are 12 different ordinal score levels for females and 15 for males.

Table 1 gives the results of statistics derived from the ordinal probit model. Standard errors for each measure, where obtainable, are given in parenthesis. For this model, κ_m measures the overall assessment of chance-corrected agreement based on ρ_0 and ρ_c, scaled to be between 0 and 1. Because κ_m is close to 0, there is little agreement among the judges' ratings. However, the low agreement is more likely a result of combining all rounds for all divers than it is for actual non–agreement. ρ_0 measures the observed (exact) agreement between many experts classifying the same sample of subjects using an ordinal categorical scale. Here, ρ_0 is moderate at 0.537 for males and 0.331 for females. ρ_c is the agreement corrected for chance. ρ_c is slightly greater for females than it is for males. Although there is some discrepancy using the probit model, there is also some support for what was seen in Figures 3 and 4; there is slightly less agreement among female divers than for male divers.

The Bayesian Ordinal Model given in (7) is another way of examining these data, in that we can see where sources of disagreement among judges arise. The results for this model are given in Table 2. Each statistic is calculated separately for males and females. The first pair of columns gives the bias statistic for each judge. The second pair of columns shows the discrimination statistic, and the third pair of columns shows the signal to noise ratio. The 0's for Judge 1 indicate that all bias measurements are relative to Judge 1. Since the bias measurements are all positive, this indicates that Judge 1 is stricter compared to the other judges. There was some evidence for the strictness of Judge 1 in Figures 1 and 2.

For discrimination, we see that scores for males are greater than those for females. This indicates that judges were better able to discriminate the performance of males than they were for females. Perhaps this is easiest to explain with the signal to noise ratio, which is much greater for males than for females. This result supports the other measures in that there seems to be more variability in agreement for female divers.

Table 2: Statistics calculated from the Bayesian Ordinal Model

Judge	Bias (α_j)		Discrimination (β_j)		STN Ratio (β_j^2/σ_j^2)	
	Males	Females	Males	Females	Males	Females
1	0	0	0.436	0.225	11.129	3.248
2	0.047	0.075	0.421	0.273	11.466	5.221
3	0.048	0.054	0.411	0.264	9.748	5.113
4	0.054	0.049	0.418	0.261	11.748	5.251
5	0.085	0.068	0.388	0.283	10.157	4.907

5. Conclusion and Future Work

Measuring agreement among the five judges for these data proved to be a more difficult task than expected. The data have several features that render classical measures of agreeement, such as Cohen's kappa, Fleiss's kappa, and even Kraemer's kappa, unsatisfactory. Because of the intricacies of the data, we chose to model the pattern of agreement with a probit model (Nelson and Edwards, 2015) and a Bayesian ordinal model (Cao, 2014), instead of using simple measures of interrater agreement,

Although different measures and models were used for these data, some common themes emerged. For example, all of the measures hinted at more variability in agreement for females than for males. However, all measures ignored covariates. The increase

in variability for females may be eliminated if covariates were accounted for in the model. Furthermore, the rounds were treated as if they represented scores for different divers. There are not 132 male divers in this competition. There were 12 males with 11 dives each, and similarly for females. In future work, it would be important to model the nesting of divers within rounds, as well as account for the degree of difficulty in the dives and the fact that different dives are performed by each diver.

REFERENCES

Agresti, A. (1988), "A model for agreement between ratings on an ordinal scale". *Biometrics*, 44, 539-548.

Agresti, A. (1992), " Modeling patterns of agreement and disagreement". *Statistical Methods in Medical Research*, 1, 201-218.

Aickin, M. (1990), "Maximum likelihood estimation of agreement in the constant predictive model and its relation to Cohen's kappa". *Biometrics*, 46, 269-287.

Arbuckle, L. (2012). "Function to compute multi-response, multi-rater kappa?", *R-Help*[online], Available at `http://r.789695.n4.nabble.com/Function-to-compute-multi-response-multi-rater-kappa-td4348947.html`.

Associated Press, (2002), "NBC commentators surprised, shocked by judges", February 12, 2002.

Banerjee, M., Capozzoli, M., McSweeney, L., and Sinha, D. (1999), "Beyond kappa: A review of interrater agreement measures". *The Canadian Journal of Statistics*, 27, 3 - 23.

Cao, J. (2014), "Quantifying randomness versus consensus in wine quality ratings". *Journal of Wine Economics*, 9, 202 - 213.

Clarke, L. (2014), "Report: U.S., Russian judges accused of conspiring to help certain figure skaters." *The Washington Post*, February 8, 2014.

Davies, M., and Fleiss, J.L. (1992), "Measuring Agreement for Multinomial Data". *Biometrics*, 38, 1047-1051.

Easterbrook, J. (2004), "Olympic Gymnastics Judges Booed." *Associated Press*, August 25, 2004.

Emerson, J.W., Seltzer, M., and Lin D. (2000), "Assessing Judging Bias: An Example from the 2000 Olympic Games". *The American Statistician*, 63, 2, 124-131.

Fleiss, J.L. (1971), "Measuring Nominal Scale Agreement among Many Raters". *Psychological Bulletin*. 76, 378-382.

Franklin, W. (2018). "High School Diving Competition Requirements", *ThoughtCo.*[online]. Available at `https://www.thoughtco.com/important-aspects-of-high-school-diving-1100253`.

Graham, P. (1995), "Modeling covariate effects in observer agreement studies: The case of nominal scale agreement". *Statistics in Medicine*, 14, 299-310.

Kraemer, H.C. (1980), "Extension of the Kappa Coefficient". *Biometrics*, 36, 207 - 216.

Maclure, M. and Willett, W.C. (1987), "Misinterpretation and misuse of the kappa statistic". *American Journal of Epidemiology*, 126, 161-169.

Nelson, K. and Edwards, D. (2015), "Measures of Agreement between many Raters for Ordinal Classifications". *Statistics in Medicine*, 34, 3116 - 3132.

Pajek, M.B., Cuk, I., Pajek, J., Kovac, M. and Leskosek, B. (2013), "Is the Quality of Judging in Women Artistic Gymnastics Equivalent at Major Competitions of Different Levels?" *Journal of Human Kinetics*, 37, 173-181.

Shrout, P.E. and Fleiss, J.L. (1979), "Intraclass correlations: uses in assessing rater reliability". *Psychological Bulletin*, 86, 420-3428.

Uebersax, J.S. and Grove, W.M. (1990), "Latent class analysis of diagnostic agreement". *Statistics in Medicine*, 9, 59-572.