Evaluating Scores for Comparing Powerlifters

Brant Imhoff[1] and Michael J. O'Connell[1]

[1]Department of Statistics, Miami University, Oxford, OH 45056, U.S.A.

1 Abstract

In powerlifting, it is important to have a method to fairly represent the relative strength of lifters. Various scores have been developed to comparing lifters with respect to weight class, biological sex, and equipment status. The Wilks Score has been the standard score for powerlifting for the last three decades. Recently, lifters have suggested replacing the Wilks Score with something that more fairly represents lifters across these demographics. Scores such as the IPF's new Goodlift Points (GLP) or DOTS have recently been approved for competition use in other federations. This analysis has two goals: develop novel powerlifting scores and compare them with existing scoring methods. Scores are evaluated through a computational test with 10,000 bootstrap samples of IPF data, computing χ^2 goodness-of-fit test statistics. The Wilks Score is shown to be more biased than novel and existing scores for modern IPF data. The novel P2 Score is the best fit for the general body of IPF lifters and should be used in local competition. GLP is the best score for Elite IPF lifters and should be used in national/international competitions.

2 Introduction

Competitive powerlifting is a strength sport in which an individual attempts to lift the maximum amount of weight possible. A competition is comprised of three different types of lifts: squat, bench press, and deadlift [4, 2]. For each of these three lifts, an individual is given three separate attempts to lift as much as possible [4, 2]. The goal is to have the highest total amount of weight lifted, computed as the sum of the heaviest performed squat, bench press, and deadlift.

Competitors are divided into weight classes, which are further separated by sex [4, 3]. This separation is based on the idea that the more mass an individual holds, the more mass that person should be able to lift. Additionally, classic powerlifting, which is also known as raw powerlifting, competes separately from equipped powerlifting. Raw powerlifting limits a lifter to gear that is not significantly assistive in nature along with the uniform all lifters must wear (an approved singlet, shoes, socks, and a t-shirt). Per IPF (International Powerlifting Federation) raw powerlifting rules, a raw lifter is able to use a supportive belt, knee sleeves, and wrist wraps [4, 10-15]. Equipped lifters are allowed to wear gear that provides more compression and benefits the lifter more than typical lifting gear. These items include a squat/deadlift suit under the singlet, knee wraps (which are different from knee sleeves used in raw lifting), and a bench press shirt [4, 10-12]. Raw and equipped lifters do not compete against one another in conventional powerlifting competitions, even if they fall into the same weight class. Thus, there are not only weight class divisions, but equipped and raw divisions. In addition to weight classes, competitions also allow lifters to enter age divisions. The divisions are teen, junior and collegiate, open (age 24 to 39), and masters (age 40+) divisions [4, 3].

Even within a weight class, determining a winner is not simple. If everyone weighed the same, the person with the highest total kg at the end of the competition would win. However, each class contains a range of weights, so to compare lifters fairly, total weight lifted needs to be scaled by each competitor's body mass. Previous research has suggested that strength is not a linear function of bodyweight. Astrand and Rodahl suggested that strength and body mass have a relationship that would be better characterized by a polynomial function [1]. Robert Wilks, the inventor of the Wilks Formula, provided the previous standard for making this adjustment with a set of scores derived from fifth-order polynomial regression equations [16]. His motivation for creating the Wilks Formula was to best characterize the relationship between body mass and strength in powerlifting [2]. The Wilks Score gave lifters the ability to compare themselves across weight classes. It also allowed for comparisons between sexes outside the parameters of competitions, providing better recognition for female lifters in a sport where males are the majority. This score has been used by all federations as its standard for measuring strength for decades, but has been critiqued in recent years for its biases against middleweights and favorable biases for heavyweights and lightweights. Because of this the IPF moved to a new score called IPF Points, which modeled total kg. lifted as a function of the natural logarithm of bodyweight and adjusts for biological sex and equipment division [13]. The IPF replaced IPF Points with Goodlift Points (GLP) in 2020 [?]. The formula for GLP is an exponential regression model that scales for sex and equipment. Another formula that has been well-received by federations outside the IPF, called DOTS, uses polynomial functions to model strength and adjust for differences between sexes. Other scores such as Wilks and the ones developed in this analysis also scale based on similar variables, just with different distributional assumptions.

2.1 The Wilks Formula and Overfitting

Roberts Wilks last updated his fifth-order polynomial equations in 1995, suggesting no difference in coefficients from his initial model that was fit in 1984 [2]. The primary concern with the Wilks Score is its model form. Polynomial regression equations with high order coefficients, typically anything above a quadratic or cubic term, start to take on extreme values [6]. It is likely that the Wilks equations were overfitting the concurrent powerlifting data. Models that overfit are overestimating the complexity of the relationship between the predictors and the response. Graphically, this would be represented by a fitted curve that is not smooth enough for the data which it is attempting to approximate. Higher-order polynomials that overfit typically perform exceptionally well when evaluating goodness-of-fit to the data on which the models were trained, but perform noticeably worse on new data [11]. It would be unfair to say that the Wilks Score has not been useful. Rather, its use has only come under scrutiny as of late because of the growth of powerlifting. More people are beginning to question if Wilks is the right fit for the powerlifting community. Its use for selecting the best lifter at an elite competition may make sense due to the extreme values a fifth-order polynomial would produce, mimicking the behavior of world-class powerlifters. However, the complexity of a fifth-order polynomial is highly questionable for modeling the relationship between mass and strength for anyone who is not internationally recognized, which is the vast majority of lifters. Thus, this analysis explores other statistical models with smoother functions to create novel powerlifting scores. Additionally, a fifth-order polynomial equation using updated, higher-dimension training data will be used to mimic the Wilks Formula. The goal of building these models is to best explain the relationship between a lifter's total and their demographic characteristics and body mass and compare novel scores to the Wilks Score, as well as determine how and where the Wilks Formula overfits.

2.2 Biases of Modern Powerlifting Scores

Minimizing bias of the powerlifting scores is imperative. People pay money to participate in powerlifting, whether that be in the form of membership fees, meet fees, equipment, or travel expenses. Federations should honor the contributions and dedication of their members by using the fairest scoring method possible, even if in the grand scheme of it all, most people are not competing for titles or prize money. Fair comparison that accounts for features like weight class, sex, age, and equipment choice will continue to grow the sport. Perhaps the most discussed reason for maintaining fairness in scores is the selection of the best lifter at competitions. When titles and prize money are at stake, it's important to those competing that the winner is chosen fairly. These reasons highlight the importance of fairness for the whole community and top lifters.

Prior to discussing methods to evaluate potential biases of the original Wilks Formula,

Table 1: Classification of IPF, USAPL Lifters by Percentiles [15]

Lifter Classification	Elite	Master	Class 1	Class 2	Class 3	Class 4	Class 5	Novice
Quantile	.975	.95	.85	.75	.5	.25	.1	0

an updated Wilks Formula, GLP, DOTS, and the scores developed in this study, it is imperative to define unbiasedness in this context. An unbiased powerlifting formula would rank lifters such that representation by demographic would be the same in smaller samples as demographic representation in the full powerlifting community. More specifically, the proportion of lifters within a certain classification/skill level (seen in Table 1) for the entire IPF sample given a certain demographic or variable (ie. sex, weight class, equipment) should be approximately the same as the proportion of lifters in that same demographic for any given subsample from the IPF data. This idea of unbiasedness should apply particularly to the covariates that were adjusted for in statistical models, which are body weight (which should translate to weight class), sex, and equipment. A statistical representation of this concept of bias is illustrated in the equation below.

Equation for Bias of an Estimator of IPF Proportion

$$
\begin{aligned}
Bias_{score}(\hat{p}) &= E(\frac{d}{n}) - \frac{D}{N} &(1)\\
&= E(\hat{p}) - p &(2)
\end{aligned}
$$

d is the number of lifters from a specified demographic such as weight class, sex, or equipment choice for a specified classification, limited to only lifters in **n**. **n** is the sample size of a specific classification/skill level (Table 1) of lifters based on ranking from a score. **D** is the number of lifters from a specified demographic such as weight class, sex, or equipment choice within the entire IPF sample of size **N**. **N** is the number of total lifters in the IPF sample, which is 66,374 in this analysis.

An estimator is unbiased if its bias is 0. Smaller bias indicates that an estimator is close in approximating the true parameter which it is attempting to estimate [12]. Given this definition of bias, if the subsample proportion of a specific demographic of lifters matches the full sample proportion of that same demographic in the entire IPF, the method of estimation is statistically unbiased. For example, consider the subsample of 105 kg. Elite lifters and the Wilks Score. If all lifters are ranked by Wilks and the top 2.5% (Elite lifters) are taken, then the subsample proportion is the proportion of Elite lifters who are in the 105 kg. weight class. If the proportion of Elite 105 kg. lifters matches the full-sample proportion (for the whole IPF) of 105 kg. lifters, then Wilks would be unbiased for the 105 kg. Elite lifters. For this analysis, bias will be evaluated for different models and compared using a test described later. This evaluation takes place after each model is developed into a score which is applied to each lifter where the lifters are then ranked by each score, then divided into samples by lifter classification quantiles (see Table 1) based on their lifting abilities according to each score.

Previous discussions of bias for scores such as the Wilks Score have focused on how the distribution of de facto 'elite' lifters matches the entire lifter population, whether that be the top 100, top 1%, etc. Per IPF and USAPL standards, an elite lifter would be anyone in the top 2.5% of the federation when ranked by score (Wilks, GLP, DOTS, etc.). This idea of comparing elite lifters to the rest of the IPF can also be applied to the other seven lifter classifications based on quantile values (see Table 1). If the distribution of the top lifters should match the IPF distribution, so should the distribution for novice lifters, as well as everyone in between. Other scores such as GLP and DOTS take different approaches to evaluating bias.

GLP are based on biomechanics principles that model the relationship between body weight and relative strength with an exponential regression equation [10]. The creators of GLP use cross-validation to select a final model, which has not been detailed in the development of other powerlifting scores. The evaluation of bias of GLP focuses on record distances. For each weight class, the distance from the IPF record in that class was calculated. Then, all lifters in the data were ranked by GLP, as well as other scores such as DOTS. For the top 50% of lifters, the Spearman correlation between record distance and powerlifting score was calculated separately for quantiles in increments of 10% (0-10% from WR, 10-20% from WR, etc.). This approach suggests that ideal behavior of a powerlifting score ranks lifters such that their distance from the world record is linear and decreasing. The practitioners then compare performance for all powerlifting scores separately for sex-separated equipment divisions for total and bench press, respectively, leading to eight unique comparisons of scores. These eight comparisons are made, ranking each powerlifting score by its average Spearman correlation for the five separate quantile-based correlations for each score. This method of evaluation ranks GLP ahead of all other scores, followed by DOTS, Wilks-2, Wilks, and IPF Points. The practitioners perform this same process for calculating coefficient of variation for each score instead of Spearman correlation. Again, GLP is determined to be superior to other scores given its coefficient of variation was smallest on average across the eight comparisons. They suggest this is sufficient to prove that scores are homogenous for across weight classes. There are multiple flaws in these evaluations. Chief among these is the lack of granularity in evaluation. To suggest homogeneity of rank across weight class, rank should be evaluated with respect to weight class in some way. Comparing rank correlation with record distance or coefficient of variation by ranking fails to show homogeneity of distribution for weight classes by not performing calculations on a by-weight-class basis or providing weight class distribution by ranking. Low coefficients of variation and strong, negative rank correlations can still be achieved without homogeneity of weight class distribution across ranking groups, which makes GLP's evaluation invalid.

The DOTS Formula does not provide anything revolutionary in its form, coming from a polynomial model similar to the Wilks Formula. The slight difference is that it uses fourth-order terms for body weight in its formula opposed to Wilks' fifth-order terms. Its methods of evaluation are not novel either. Tim Konertz, the creator of DOTS, provides theoretical underpinnings for his model, but fails to provide any novel valid statistical validation of his approach. Konertz looks at the coefficient of variation for the same eight comparisons of IPF lifters as the GLP committee, suggesting DOTS has the lowest coefficient of variation. One improvement he makes to the GLP evaluation is looking at the coefficient of variation for male world records on a weight class basis. However, this use does not provide full distributional information for each powerlifting score, as it only looks at the coefficient of variation for a score using the records for each weight class. Konertz also asserts that ”DOTS points were evaluated in several national and international competitions” and ”showed significantly fairer relative winner distributions, as IPF points or Wilks” [9]. He evaluates performance of DOTS, Wilks, and IPF Points at the 2019 IPF Championships and 2019 IPF Bench Press World Championships based on scores required to be the best lifter and place top 3 within each sex and equipment division, while considering the distance to the world record of the hypothetical scores. His idea is that a scoring system should rank lifters according to how close their performance approaches or exceeds world records. While this seems like one intuitive way to approach fairness for powerlifting scores, it also assumes that world records serve as the same barometer regardless of weight class. In simpler terms, it assumes that breaking the 120+ kg world record is just as difficult as breaking the 83 kg world record for lifters in each of those respective weight classes. This is a strong assumption that would require further investigation to be met. Even with Konertz making this assumption, his calculations for distances from world records are not properly scaled. His results are displayed in distance from WR by weight class in kilograms when they should be displayed as percentages of the world record. Naturally, a heavier weight class will have a higher WR than a lighter weight class. For example, it assumes that 10 kg from the world record is the same in the 59 kg weight class is the same as a 10 kg distance from WR for the 120+ kg weight class. Thus, these distances below or above the WR should be scaled as percentages from the current WR to adjust for differences between weight classes. A replication of Tim Konertz’s evaluations of the 2019 IPF World Championships was completed with the addition of GLP, an Updated Wilks, and the novel scores created in this analysis. The results and implications of the replication of Konertz’s analysis are displayed in the supplementary materials.

In Validation of the Wilks Powerlifting Formula, bias was evaluated graphically by scatterplots of body weight and Wilks Scores with separate curves for each sex for totals and individual lifts [16]. If the points were distributed symscoreally about the fitted lines, then it was determined that Wilks was an unbiased measure of strength for that lift/total by sex for the body weight ranges represented in the data. This worked well because of the small subsample used in the analysis, which included 27 female lifters and 30 males. This sample was selected from IPF World Championships from 1996 to 1998, similar to the data that trained the polynomial regression model for the Wilks Score. Thus, it was trivial to evalute bias using a scatterplot given that there were so few data points around the fitted regression lines. It was determined that Wilks was unbiased for total kg. lifted given that the data were roughly uniformly distributed about the fitted regression line [16]. More recent works have suggested the Wilks Formula should be updated due to bias [8] and changes in weight classes and overall athlete body composition [5]. In this paper, evaluating biases is not as simple given the amount of data (66,374 unique IPF-affiliated lifters). Additionally, the samples of interest are the various lifter classifications (Elite, Master, Class 1, etc.) based on quantiles of total kilograms lifted or scores such as GLP or DOTS (see Table 1).

3 Methods

3.1 Data

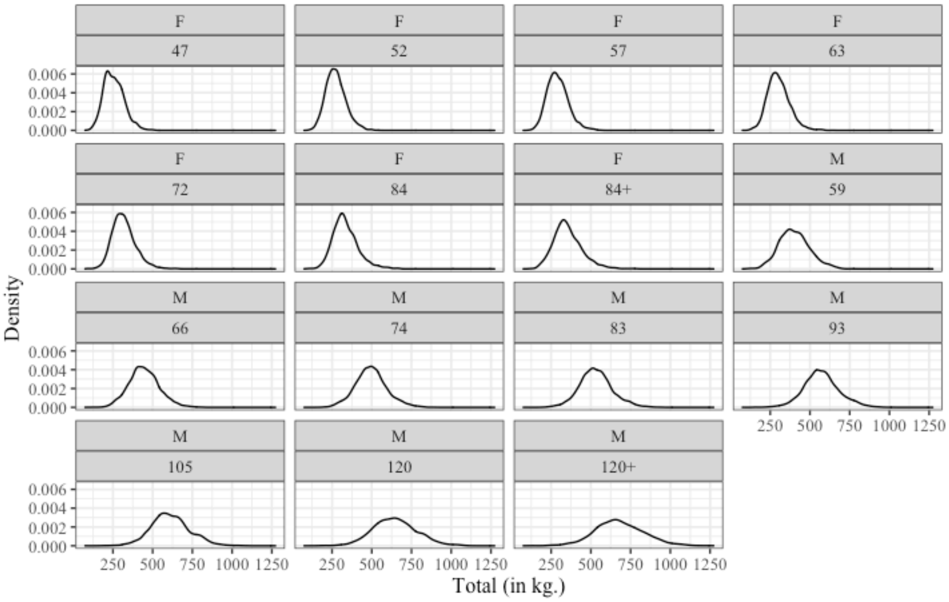
The IPF sample was obtained from the open source database Open Powerlifting [14]. This database is maintained and updated regularly by a group of developers with the help of powerlifting enthusiasts. Data for this project were downloaded in .csv format from Open Powerlifting. The initial data set included approximately 1.4 million observations, each representing a competition for a powerlifter. Competition data were compiled from September 1964 to April 30, 2019. Lifters had each of their attempts recorded, as well as their best attempt in each of the three lifts to form a total. Negative values for attempts indicated missing a lift, while positive values indicated a successful attempt. Additionally, measures such as body weight in kg., biological sex, equipment, and age were collected. For each row, scores such as Wilks, DOTS, and GLP were given.

The scope of the initial data set was narrowed to include only IPF lifters given the interest of comparing Wilks, DOTS, Goodlift (which was created using only IPF lifter data), and novel scores from statistical models developed in this study that were trained only on IPF data. These data were filtered to include lifters in only IPF-affiliated federations and current IPF weight classes (implemented in 2011) for full powerlifting meets, excluding push/pull and bench only competitions. Further, these data were filtered to include only lifters eligible for the Open division (ages 14+) and of reasonable body weight (at least 40 kg). Competition results with missing values for body weight, biological sex, and equipment were excluded. The categories of single-ply and multi-ply from the Equipment variable were reduced to one equipped category in a new variable EQ. The new EQ variable included only two categories: classic/raw and equipped. This was done in the interest of simplifying statistical models without losing much information given the IPF does not currently allow multi-ply equipment [4, 10]. The final IPF sample contained 66,374 observations and 41 variables.

Variables of interest for statistical modeling include those for which bias will be evaluated: weight class, sex, and equipment. Strength can be evaluated on the basis of each of the aforementioned variables. In Figure 1 a density plot of total by weight class displays approximately bell-shaped, unimodal distributions across all weight classes. Although, variability appears to be higher in the male weight classes.

Weight classes are mutually exclusive on the basis of sex, so males and females do not compete against one another. Equipped and raw/classic lifters do not compete against one another, but the Wilks Score does not reflect the differences between these lifters. Goodlift Points were created with the assumption that equipped and raw/classic lifters have different distributions of strength, so their coefficient estimates are different. Figure 2 displays density plots for total demonstrating a difference in distribution between classic and equipped lifters

Figure 1: Density of Total by Weight Class



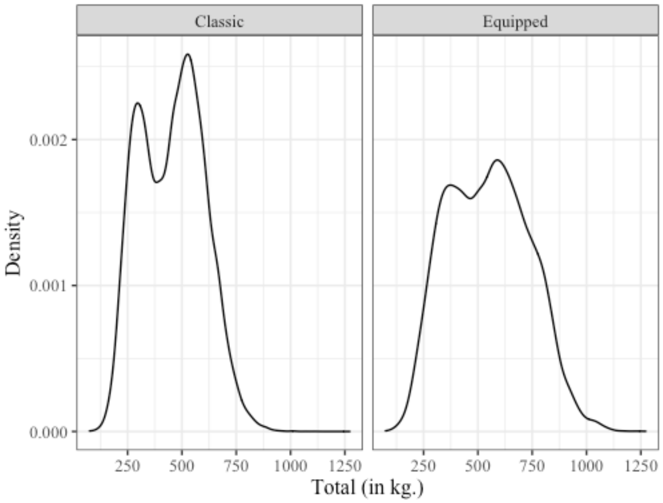
for the IPF sample of 66,374 lifters.

Both distributions are bimodal, but the equipped lifter sample looks to have more variability than the classic/raw sample. Equipment should benefit a lifter by providing additional strength, which is what differentiates equipped lifting from raw lifting. This idea in addition to the difference in the distribution of strength (in the form of total kg. lifted) between equipped and classic/raw lifting provides justification for exploring statistical models that differentiate between equipped and classic lifters.

3.2 Models

The Wilks formula was originally developed with data that are no longer representative of modern powerlifters [3]. To update this formula, a fifth-order polynomial regression model was fit using newer data with the same regression equation as the Wilks formula, given below. The coefficients from this model were used to generate updated Wilks scores in the same fashion in which Wilks scales using a ratio of observed over predicted and multiplies by a constant. Additionally, another fifth-order polynomial regression model that adjusts for

Figure 2: Density of Total by Equipment



equipment was fit to compare to the Wilks and Updated Wilks scores. This regression model included additive terms for body weight, age, sex, and equipment, as well as interactions. We will call this new score the P5.

The general form of the Wilks and updated Wilks regression models are displayed below, with total kg. lifted as the response variable. These models use polynomials for body weight in kg. up to the fifth-order and biological sex, an indicator variable, as well as interactions to model the mean behavior of total kg. lifted. The DOTS Formula also uses sex-adjusted polynomial functions of body weight to approximate total kg. lifted. DOTS differs from Wilks in that it uses quartic functions of body weight instead of body weight to the fifth power. The regression formulas for Wilks/Updated Wilks and DOTS are described below.

Wilks Polynomial Regression Equation

$$\hat{Total} = \beta_0 + \beta_1 S + \sum_{i=1}^{5} \gamma_i W^i + \sum_{i=1}^{5} \delta_i W^i S$$

Total is in kg., W represents body weight in kg., S is an indicator for Sex where 0 = Female, 1 = Male

DOTS Polynomial Regression Equation

$$\hat{Total} = \beta_0 + \beta_1 S + \sum_{i=1}^{4} \gamma_i W^i + \sum_{i=1}^{4} \delta_i W^i S$$

Total is in kg., W represents body weight in kg., S is an indicator for Sex where Male = 1 and Female = 0

While updating the Wilks Formula is one potential remedial measure for biases, another is to utilize other covariates. Of the candidate models, the best model determined by Adjusted R^2, testing RMSE, and AIC was also a fifth-order polynomial regression, but with additional predictors. Wilks utilizes bodyweight as a predictor, and has two separate formulas: one for men, one for women. The alternative equation for what will be dubbed the P5 uses the same idea as Wilks, but adjusts for equipment and sex. The polynomial regression equation for P5 is as follows:

P5 Polynomial Regression Equation

$$\hat{Total} = \beta_0 + \beta_1 E + \beta_2 S + \beta_3 E \cdot S + \sum_{i=1}^{5} \gamma_i W^i + \sum_{i=1}^{5} \delta_i W^i E + \sum_{i=1}^{5} \zeta_i W^i S$$

Total is in kg., W represents body weight in kg., S is an indicator for Sex where Male = 1 and Female = 0,

E is an indicator where Equipped = 1 and Classic/Raw = 0

In addition to the P5 model, other models were developed to adjust for equipment, which is excluded from the Wilks formula. The P2 is derived from a polynomial regression model that includes a quadratic term for bodyweight in addition to covariates for sex and equipment. It is imperative to note the convex form of the P2 Score as a function of body weight, resulting in increases in scores for extreme heavyweights (200 kg body weight). To make the P2 Score a concave function of body weight and adjust for superheavyweights, a cubic spline with a knot at 200 kg. was added to the original P2 Score. The new form of the P2 with the cubic spline will be called the P2SPL Score. These models are described below:

P2 Polynomial Regression Equation

$$\hat{Total} = \beta_0 + \beta_1 E + \beta_2 S + \beta_3 ES + \sum_{i=1}^{2} \gamma_i W^i + \sum_{i=1}^{2} \delta_i W^i E + \sum_{i=1}^{2} \zeta_i W^i S + \sum_{i=1}^{2} \theta_i W^i SE$$

Total is in kg., W represents body weight in kg., S is an indicator for Sex where Male = 1 and Female = 0,

E is an indicator where Equipped = 1 and Classic/Raw = 0

A score that has recently received attention due to its replacement of IPF Points, is called Goodlift Points (GLP). Goodlift Points (GLP) utilizes an exponential function of body weight instead of a linear term, but also adjusts for sex and equipment and their interactions. This model form takes the opposite approach of its predicessor, IPF Points, which specified Total kg. lifted as a function of the natural logarithm of body weight. GLP utilizes separate models of the same form to model total and bench press.

Goodlift Exponential Regression Equation for Total or Bench Press

$$T\hat{o}tal = \beta_0 + \beta_1 S + \beta_2 E + e^{-S \cdot E \cdot W}$$

Total/Bench Press is in kg., W represents body weight in kg., S is an indicator for Sex where Male = 1 and Female = 0, E is an indicator where Equipped = 1 and Classic/Raw = 0

Once all models were fit, their fitted values for total kg. lifted were used to create new powerlifting scores. The statistical models created in this analysis predict a lifter's total based on the relationship specified in the model form, implying that scaling still must be completed. Because of this, an ad-hoc method of scaling will be used. This problem of creating an adjusted score from a linear model is sparse in the statistical literature, but is familiar to the powerlifting community. Methods used previously by Robert Wilks in his fifth-order polynomial regression conversion to the Wilks Score can be utilized here. The general form of his scaling method is often glossed over, with no explanation or justification. From a statistics perspective, the scaling appears strange. However, when coupling the context of the problem with statistical ideas, the idea becomes simple and intuitive. If the model is believed to explain the relationship between strength and some predictors of choice, then comparing actual results to the model's predictions of strength provides a relative measure of strength. In simpler terms, the ratio of a lifter's observed total to predicted total can be thought of as how many times stronger a lifter is than what the model suggests they should be. This ratio of relative strength makes practical sense if the model is believed to be truly explanatory of strength and also provides a solution to the issue of unfair scaling for different scores. Thus, the general form of a powerlifting score is as follows:

$$Score\ =\ C * \frac{Total\ kg.}{Tot\hat{a}l\ kg.}$$

C is a positive, arbitrary constant used to inflate the **Total kg.**:**Totâl kg**. ratio, providing further space between individuals' scores. **Total kg.** is the actual total kilograms lifted for a given powerlifter. **Totâlkg**. is the predicted total kilograms lifted by a statistical model for a given powerlifter

3.3 Using Goodness-of-Fit to Assess Bias

The ultimate goal of this analysis is to measure the goodness of fit to the full IPF sample of the Wilks Score, DOTS, GLP, Updated Wilks, and novel scores generated from the models described in the previous section. However, the full IPF sample is dynamic, continually growing along with the strength of its competitors. If the IPF data weren't changing so frequently and requiring the updating of lifting scores to meet expectations of fairness, then the IPF sample of 66,374 lifters that trained the models developed in this analysis could be considered the true population. This would make the goodness of fit evaluation much simpler in practice, given each score's performance is limited to this one sample.

Evaluating goodness of fit can be completed in a variety of ways. This study utilizes χ^2 goodness-of-fit statistics to measure how fairly each of the six scores ranks IPF lifters. Jeong et. al. used the bootstrap method to conduct non-parascore hypothesis tests with χ^2 tables for two categorical variables [7]. Their computational process is similar to the one conducted here, commonly using statistical models fit to an initial sample, the implementation of bootstrap samples, and calculation of χ^2 test statistics for each bootstrap sample to evaluate goodness of fit of each score. The difference between the methods used here and those used by Jeong et. al. is the way inference is conducted. The analysis of powerlifting scores relies on the comparison of the distributions of χ^2 test statistics and confidence intervals from bootstrap samples, while Jeong et. al. compressed their results to p-values [7].

While bias and goodness-of-fit are inherently two separate concepts, the two are naturally intertwined in this analysis of powerlifting scores. The goal of this analysis is to quantify bias of powerlifting scores by evaluating differences in goodness-of-fit statistics. The parameter of interest to be estimated is the count of lifters within each cross-section of skill level (ranking quantile based on powerlifting score) and demographic. The count parameter is based on the population proportion of lifters who fall within that demographic of interest for the entire lifting federation. The point estimate of count is the observed number of lifters who fall into the skill level and demographic cross section given the powerlifting score used. Thus, if a powerlifting score is able to generate a point estimate for count that matches the count parameter based on the whole federation, the score produces an unbiased estimate. Goodness-of-fit statistics become useful in quantifying bias in this application for two reasons. First, there are many cross-sections (r x c) of lifter classifications (skill level) and demographic level for each table. A bias calculation becomes convoluted in this scenario and does not provide one summative measure for each powerlifting score. Second, the goodness-of-fit statistic is a function of bias for each individual cell, given that it takes the difference between point estimate for count and count parameter (see equation below). The χ^2 statistic takes this function of bias and sums it for the whole table, providing a summative measure of bias or goodness-of-fit statistic for a specific powerlifting score. Comparing these statistics then allows for comparisons in bias between powerlifting scores.

For each bootstrap iteration, six χ^2 tables (one for each score) are generated. Each powerlifting score's χ^2 table can be used to judge how well the score ranks lifters with respect to a specified demographic variable. Perhaps the most important of these demographic variables is weight class. A χ^2 statistic for a powerlifting score based on weight class distribution would display counts of lifters within each weight class for each skill level, comparing it to hypothesized counts based on proportions within each weight class for the full federation.

Table 2: Proportion of All IPF Lifters by Weight Class (Original IPF Sample)

Weight Class (kg.)	47	52	57	63	72	84	84+	59	66	74	83	93	105	120	120+
Proportion of IPF Lifters	.01	.03	.05	.06	.07	.05	.04	.03	.05	.10	.14	.15	.11	.07	.04

The proportion of lifters within each weight class for the original sample of IPF lifters is displayed in Table 2. For each bootstrap sample, these proportions are calculated then used to obtain the hypothesized counts of lifters within each cross section of χ^2 tables.

χ^2 Test Statistic Calculation

$$\chi^2 = \sum_{All\,Cells} \frac{(Bootstrap\,Sample\,Count\;-\;Hypothesized\,Sample\,Count)^2}{Hypothesized\,Sample\,Count}$$

Hypothesized Sample Counts are the expected counts of lifters within a classification/skill level; These are calculated by taking the product of the classification's sample size and the proportion of lifters in the demographic level of interest within the full IPF bootstrap sample of size 59,047.

The test statistic above can be calculated for each of six tables for each iteration of the bootstrap sample. The lower the test-statistic, the less biased that powerlifting score is for the full sample. Each score has a test statistic for each iteration of the bootstrap. In the following section, the implementation of χ^2 goodness-of-fit statistics is coupled with the use of bootstrapping.

3.4 Comparing Methods with a Bootstrap Test

Now that bias has been defined, the methods of evaluating biases for each score may be outlined. The goal of developing an alternative to Wilks and IPF Points is two-fold. First, it is desired to provide a score that more fairly distributes lifters regardless of classification amongst demographics, whether it be by weight class, sex, or classic/equipped status. Second, a score should accomplish the aforementioned duty for different data sources that are representative of the IPF population, not just the IPF sample on which the models were trained. The abundance of data from the past few decades of powerlifting inherently has led to the development of models which are more flexible than Wilks' fifth-order polynomial. The popularity of the sport will also provide a stream of new data in the future. In an attempt to better approximate how fairly powerlifting scores rank IPF lifters, the statistical tool of the bootstrap may be used to quantify the variability of the goodness of fit of these scores. The bootstrap involves taking the data at one's disposal and resampling it randomly with replacement, matching the same sample size as the original data. A sample statistic is then taken from each bootstrap sample. The goal of bootstrapping is to build a distribution around the sample statistic in order to better encapsulate the quantity's variability.

In practice, the idea is to take the 66,374 unique best totals of IPF lifters and bootstrap a large number of samples. For each bootstrap sample, each lifter is then ranked by each of the six scores. Once ranked, lifters are divided into classifications (see Table 1) and for each weight class, the expectation of the proportion of lifters within that classification/quantile (seen above in Equation for Bias of an Estimator of IPF Proportion) is taken. This is completed for each of the six scores, then a table below is generated for each of the six scores with proportions or counts. The cross-section at each cell refers to a sub-sample within each weight class and lifter classification for each of the six scores. An example of this table may be found in the supplementary materials.

This analysis utilizes 10,000 iterations of boostrap resampling. Each iteration produces 6 χ^2 statistics (one for each of six powerlifting scores) for each of 3 demographic analyses (weight class, sex, and equipment). From these test statistics, a bootstrap distribution of test statistics can be assembled for each score. Each bootstrap distribution for each score should have one test statistic per iteration of the bootstrap. Assembling boostrap distributions for each score provides an idea of how well a score does in terms of distributing lifters by ranking. From these distributions, differences in test statistics can be taken to see which score outperforms another. Particular contrasts of interest would be all scores with Wilks. A positive difference in test statistics would indicate a larger test statistic for the reference/first score, suggesting the subtracted/second score matches the IPF sample better.

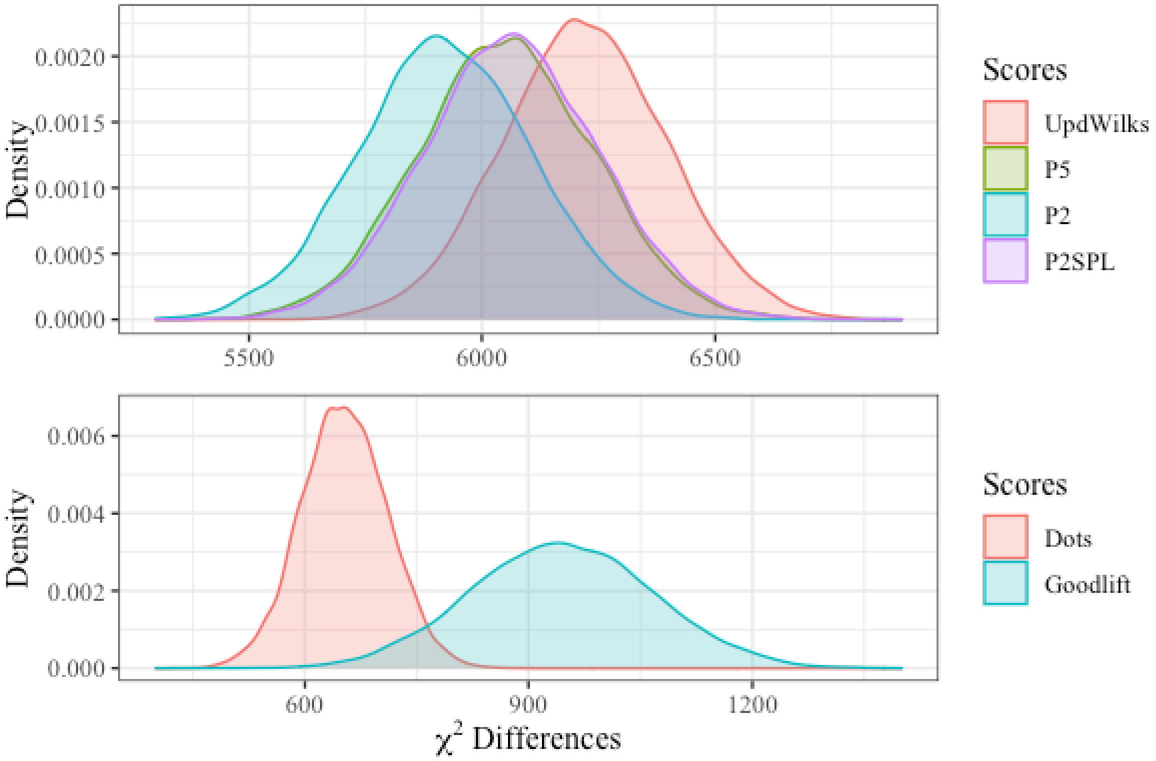
To make interpretation easier, Wilks will be the reference score. Thus, a positive difference in test statistics would suggest that the score being compared to the Wilks Score more closely matches the IPF sample. Conversely, a negative difference would indicate that the reference/first score, Wilks, matches the IPF sample better in that iteration of the bootstrap. Distributions of differences in test-statistics and quantile-based confidence intervals are used to demonstrate pairwise comparisons in the Results section.

4 Results

4.1 Bias by Weight Class

For each bootstrap sample, lifters were given a Wilks Score, Updated Wilks Score, P2SPL, etc. All lifters were then ranked by each score for each iteration of the bootstrap, then divided into classifications determined by quantiles (see Table 1). Within each classification, the number of lifters within each weight class for that classification were obtained. Each cell in a bootstrap sample's contingency χ^2 table corresponds to this count of lifters in a given weight class for all of the classification levels. After completing 10,000 bootstrap samples of IPF lifters, χ^2 test statistics were obtained for each score in each sample. Further, differences in test statistics between scores were taken to make comparisons of interest. The original Wilks Score performance was compared to all other scores given that Wilks was the gold standard of powerlifting scores for decades and has only recently been replaced by other scores. As of 2020, the IPF's standard for best lifter was determined by GLP, replacing the Wilks Score as the competition standard. Thus, comparing GLP to Wilks and all other scores remains pertinent for future IPF competitions. DOTS should also be compared to Wilks given it has served as its replacement in federations such as the USPA, BVDK, and Swiss PL. Visually, the performance of scores compared to one another can be visualized with density plots or histograms of the differences in χ^2 test statistics. Differences were calculated as the reference score (Wilks) minus the new score. Therefore, a positive difference in test statistics would

Figure 3: Differences in χ^2 Test Statistics for Weight Class between Wilks Score and Other scores



suggest that the reference score had higher χ^2 values than the new score. This would mean the new score performed better in bootstrap samples than the reference score, suggesting that the new score would be better at matching the observed IPF representation by weight class. Conversely, if differences were negative this would indicate that the reference score, Wilks, is better than the new score in terms of bootstrap sample performance.

In Figure 3 below, densities of differences in test statistics between the Wilks Score and other scores are displayed for 10,000 bootstrap samples of 66,374 unique IPF lifters. In χ^2 tests for goodness of fit, the better the observed counts match expected counts in each cell of a table, the lower the test statistic that table will produce. Thus, the lower the test statistic, the better the score for that bootstrap sample. In this case, each table for a specific score has counts for the skill level (classifications found in Table 1) by each weight class. To compare scores, the difference between two scores for the same bootstrap sample provides an indication of performance.

Each of the densities are entirely positive in χ^2 differences, suggesting all scores are less biased than Wilks on the basis of weight class. The density curves with the highest positive densities are the three novel scores (P2, P2SPL, and P5) and the Updated Wilks. This indicates that these scores provide noticeably better fit for IPF lifters by weight class when compared to the Wilks Score. GLP and DOTS also perform favorably compared to Wilks on the basis of weight class. The curves for differences in χ^2 statistics for GLP and DOTS are entirely positive, but lower in magnitude compared to the novel scores. A higher positive magnitude suggests a score is less biased for all skill levels of IPF lifters. Thus, the novel scores and Updated Wilks Score are the least biased for all skill levels of lifters for weight class.

Density plots of differences in χ^2 statistics can be supplemented by 95% quantile-based bootstrap confidence intervals. The lower and upper limits for these confidence intervals are the 2.5 and 97.5 percentiles of the density curves displayed in Figure 3. These intervals found in Table 3 can be viewed the same as the density curves. Positive values suggest that the score is less biased than Wilks on the basis of weight class, while negative values indicate more bias than Wilks. If an interval contains 0, the difference between the score of comparison and the Wilks Score is inconclusive. All scores display entirely positive intervals in Table 3, suggesting they are all less biased than Wilks by weight class. The magnitude of positivity is much higher for Updated Wilks, P2, P2SPL, and P5 Scores than GLP and DOTS. The novel scores and Updated Wilks have 95% confidence intervals ranging from the mid-to-high 5000s to the low 6000s, which is noticeably higher than the intervals for DOTS and GLP in the low hundreds. When attempting to distinguish between scores for IPF lifters in general, there is insufficient evidence to suggest a difference between the novel scores and Updated Wilks. This is because their curves and intervals overlap. The same applies when distinguishing between GLP and DOTS.

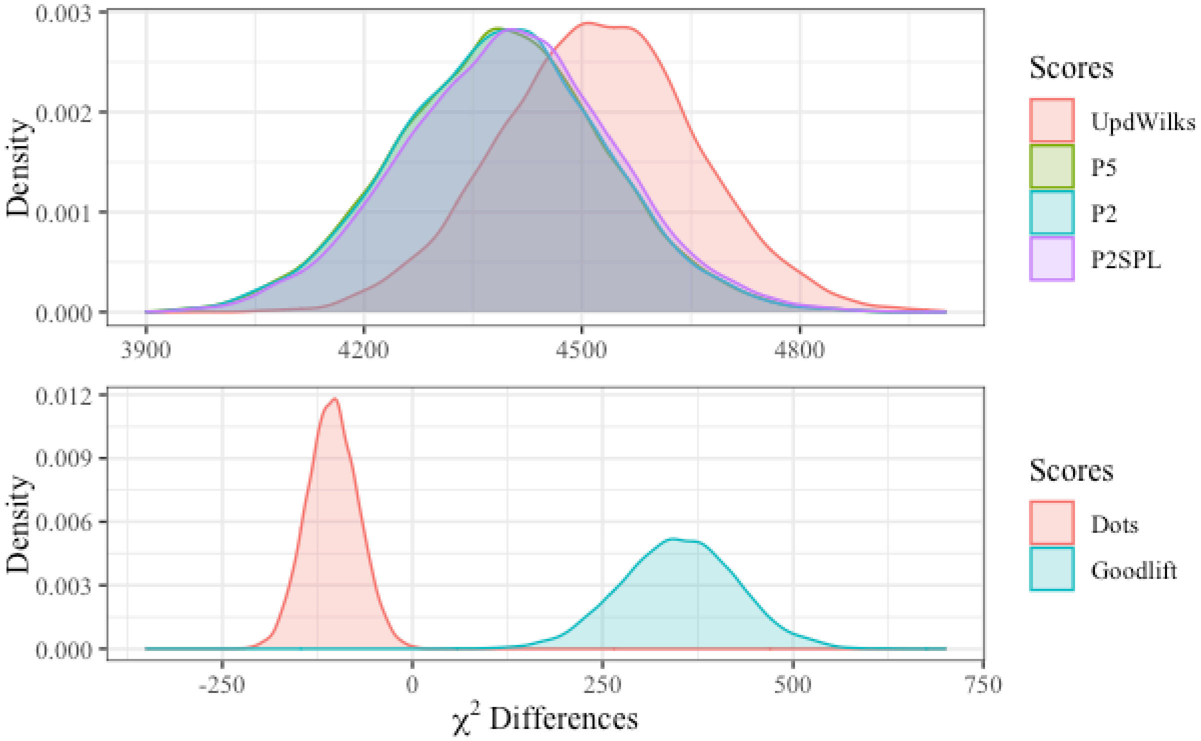
All scores outperform the Wilks Score by weight class when evaluating fairness for all skill levels of IPF lifters. In the interest of providing more skill-level specific results, bias of scores can be further evaluated on a more granular level. While a generalized result applied to the entirety of a federation may suit one federation, another may only be interested in using a score that performs well for the best of the best. This can be accomplished by looking at the average bias for each score over 10,000 bootstrap iterations. These calculations can be made for each skill level of lifter and level of demographic variable. An example of this would be the average bias of the Wilks Score for 93 kg. Elite (top 2.5%) lifters. Heatmaps in the supplementary materials of this publication show the average biases for all scores over 10,000 bootstrap samples. In these heatmaps, red indicates favorable bias (overrepresentation) for that specific group of lifters, while blue indicates unfavorable bias (underrepresentation). White indicates no bias (precise representation). This concept of bias harkens back to the Equation for Bias of an Estimator of IPF Proportion found in Section 2.2.

When discussing biases of powerlifting scores, the focus is often the general lifter population or the best of the best. The best lifter award is often selected within each sex and equipment choice, considering all weight classes. At national or international competitions, these would be the Elite lifters. The IPF defines Elite as the top 2.5% of lifters. Though this standard is ambiguous as to what ranks the lifters (either total or score), it is assumed to be score in this analysis for the purpose of comparing fairness for different powerlifting scores. Thus, the heatmaps for weight class in the Elite lifter category provide evidence as to how fairly a score will choose a best lifter at the highest podiums in powerlifting. When looking at the heatmaps for male and female weight classes, respectively, there are noticeable differences between existing scores (DOTS, GLP, and Wilks) and the novel/updated scores. None of these scores provide a completely unbiased solution to ranking Elite lifters. The Wilks, GLP, and DOTS all tend to favor male lifters. Wilks seems to perform moderately well for Elite lifters, but is inconsistent in where its bias is distributed. This is observed in its moderate bias toward males in the 120 kg and 120+ kg of about 5% and 6%, respectively, while fluctuations in favorable and unfavorable bias vary just about every other weight class. The P2, P2SPL, P5, and Updated Wilks all tend to slightly favor female lifters. The favorable bias is more evenly spread amongst female weight classes, where the unfavorable bias toward males is concentrated entirely in the 83 kg, 93 kg, and 105 kg weight classes. When comparing scores for Elite lifters by weight class, there is no avoidance of bias. Rather, one must select the score with the source of bias that fits their unique criteria of selecting a best lifter. If the goal is simply to make a comparison of non-Elite lifters, the P2 seems to be the most balanced in its issuance of bias when considering the heatmaps for average bias by weight class (supplementary materials). However, if the goal is to compare Elite lifters at a high-level regional, national, or international competition, DOTS or GLP should be the standard. Both of these scores tend to be virtually unbiased for Elite lifters in lighter weight classes when looking at average bias over the 10,000 bootstrap samples (supplementary materials). DOTS has barely any bias in Elite male lifters up through the 93 kg weight class, while GLP has minimal bias up through the 74 kg weight class and the 120+ kg weight class. Both have minimal bias against Elite female lifters in the 47, 52, and 57 kg weight classes. DOTS has favorable bias toward males (of around 2-4%) for classes 105kg and greater, while GLP has the same favorable bias for those in the 83 kg class through the 120 kg class. Both of these scores have low bias against female Elite lifters in weight classes 63 kg and above, ranging from roughly 2-4%.

4.2 Bias by Sex

To evaluate the bias of each score on the basis of sex, the difference between IPF proportions and subsample proportions for both males and females were taken. In the IPF data that have been filtered to include only each unique lifter's best total, of the 66,374 people, 68.7% were male and 31.3% were female. These would be designated as the hypothesized proportions for each sex in the original IPF sample, but may differ based on different bootstrap samples of the IPF data. The idea is that within each sex and classification, the proportion of sex represented in each classification subsample should match the proportion of males and females within each full IPF bootstrap sample. Variability in bootstrap samples of all IPF

Figure 4: Differences in χ^2 Test Statistics by Sex between Wilks Score and Other scores



lifters is inevitable given the high volume of observations, so the null proportion will likely differ from sample to sample.

The density plots of differences in test statistics on the basis of sex yield similar results as the curves for weight class. However, DOTS has an entirely negative curve for χ^2 differences from Wilks, suggesting it performs worse. All other scores have positive density curves, suggesting less bias by sex than Wilks for the IPF. Similar to the weight class density curves, high positive magnitude of χ^2 differences are observed for the novel scores and Updated Wilks. This suggests that these scores provide a much better fit than the Wilks Score when looking at all skill levels of IPF lifters on the basis of sex. GLP also performs better than Wilks, with a positive density curve.

From Figure 4 the Wilks Score is more biased on the basis of sex than all other scores with the exception of DOTS. To support the visualizations, 95% confidence intervals for differences in test statistics are displayed in Table 3. Intervals constructed from the subtraction of other scores' test statistics from those of the Wilks Score are all positive, implying less bias from sample to sample in 10,000 iterations of bootstrapping the IPF lifter sample. Novel scores and the Updated Wilks Score all have 95% confidence intervals from the low 4000s to the high 4000s, suggesting much less bias than Wilks. Further, these intervals overlap, suggesting insufficient evidence for any difference between their magnitude of bias. GLP is also less biased than Wilks and DOTS on the basis of sex, with a 95% confidence interval of χ^2 differences of 201.596 to 498.339.

To provide a closer look at the bias by sex for powerlifting scores, Figure 3 in the supplementary materials displays a heatmap of average bias for each score by lifter classification and sex. It is important to note that the value of comparisons may not be high if the sole desire of the practitioner is choosing a best lifter by each sex. However, comparisons between males and females are inevitably made with the growth and increasing popularity of powerlifting. The situation also may arise in which an invitational meet selects only one best lifter regardless of sex, which would then make the fairness of comparisons between males and females of high interest. With this stipulated, the Elite lifters are more fairly represented by Wilks, DOTS, and GLP than novel scores or the Updated Wilks. The novel scores favor female lifters by about 22% more than their true representation, suggesting men are underrepresented by that same percentage. This same bias for Elite lifters is observed at a rate of 18% with the Updated Wilks. The established scores favored males by an average of 9% (Wilks), 10% (DOTS), and 13% (GLP). However, when looking at non-Elite lifters, the novel scores are much fairer than the preexisting scores. This was also observed in heatmaps for average bias by weight class. Thus, if the goal is to compare lifters of different sexes in non-Elite settings, such as local meets, use a novel score such as the P2. If the goal is to compare a male and female lifter in high-level regional, national, or international competitions, use Wilks, DOTS, or GLP.

4.3 Bias by Equipment

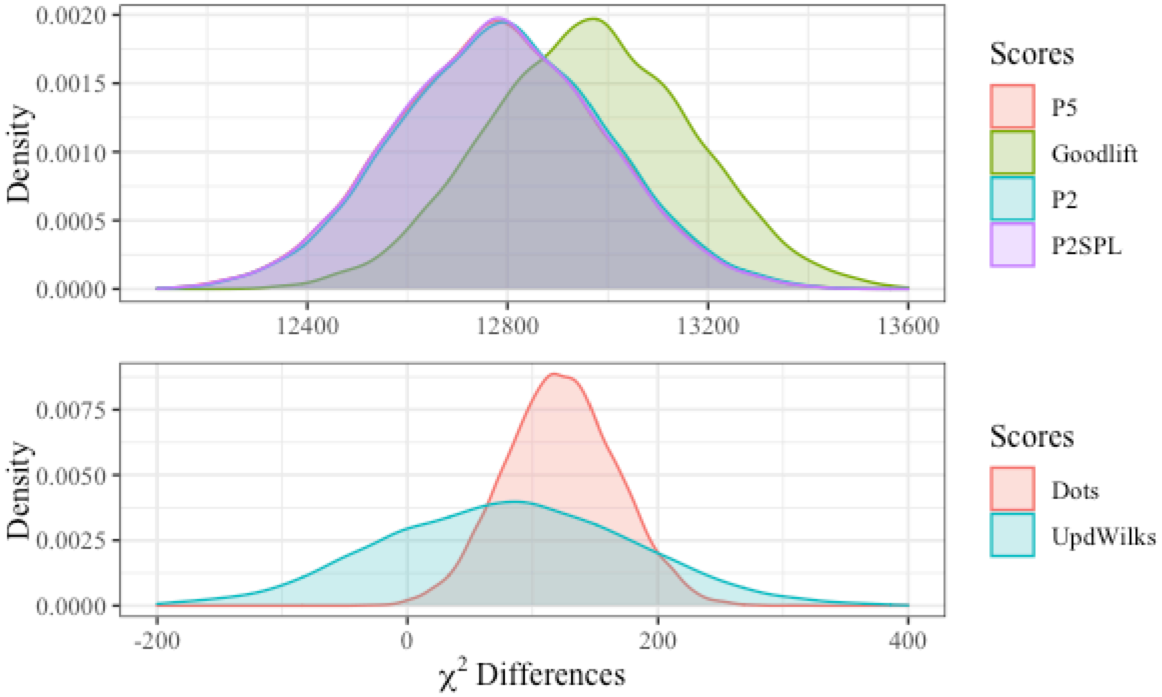
In the sport of powerlifting, raw and equipped lifters compete in the same weight classes, but not against each other. Best lifters at national and international competitions are selected separately for raw/classic and equipped. For these reasons, comparisons of raw and equipped lifters are often disregarded. Returning to the idea of the Wilks Formula, the goal is often to know who the best lifter is regardless of weight class. Best lifter is often determined within boundaries of sex and equipment, so separate awards go to men's equipped, women's raw, etc. Given that GLP and the novel scores in this analysis adjust for sex and equipment choice, these scores may be useful for choosing the most dominant lifter regardless of equipment choice. This may also prove to be useful for lifters who wish to compare their raw and equipped performances in an effort to determine which is best.

An intuitive way to adjust for equipment and measure the effect of it is to include it in statistical models for total kilograms lifted. This would allow for the comparison of raw and equipped lifters. It is widely accepted that equipped lifting should yield a higher total compared to raw lifting for a given lifter, but the degree of strength increase is variable from person-to-person. Statistical models that include equipment, as well as its interactions with other predictors such as bodyweight and sex adjust for differences between raw and equipped lifters.

The full IPF lifter sample is comprised of 72.7% raw lifters and 27.3% equipped lifters. Ideally, an unbiased score should have the same distribution of raw and equipped lifters by each classification within each bootstrap sample. For example, the top 2.5% of lifters (elite classification) should be 60% raw and 40% equipped if a bootstrap sample of IPF lifters is split 60/40 in terms of raw and equipped, respectively. This should apply to all of the quantile-based samples.

In Figure 5 below, there is a noticeable discrepancy in performance between the scores derived from models that adjust for equipment and those that come from models that do not adjust for equipment. The novel scores and GLP have high positive densities of χ^2

Figure 5: Differences in χ^2 Test Statistics for Equipment between Wilks Score and Other scores



differences, suggesting fairer representation of lifters by equipment compared to Wilks. This performance meets intuition given that the aforementioned scores come from models that provide different coefficients for equipped and raw/classic lifters, respectively. Scores such as Wilks, Updated Wilks, and DOTS all perform noticeably worse on the basis of equipment than GLP and the novel scores. It is also important to note that Wilks, Updated Wilks, and DOTS only adjust for body weight and sex. These scores do not provide any adjustment for lifters' equipment division. Despite DOTS' poor performance by equipment compared to novel score and GLP, it should be noted that it does outperform the Wilks Score. The visual findings from comparing the curves of differences in χ^2 statistics can be further supported by 95% confidence intervals found in Table 3.

The discrepancy in bias by equipment is exhibited by vast differences in χ^2 statistic differences from Wilks. GLP, P2, P2SPL, and P5 all have 95% confidence intervals with lower bounds in the mid 12,000s and upper bounds in the low 13,000s (Table 3). These large positive values indicate superior performance compared to the Wilks Score on the basis of equipment for IPF lifters. It also seems that updating the Wilks Score with new coefficients does not seem to provide fairer representation of raw and equipped lifters. This is supported by a 95% confidence interval for χ^2 differences of Updated Wilks from Wilks that includes 0 (Table 3). DOTS is deemed to be fairer than Wilks given that its interval is entirely positive. However, there is insufficient evidence of any difference between DOTS and an Updated Wilks Score because the interval for DOTS is contained entirely inside of the Updated Wilks interval. Further, the densities in Figure 5 provide visual evidence that the bias of the Updated Wilks Score is subject to a higher degree of uncertainty than DOTS given its curve is less steep and has longer tails.

Determining fairness by equipment division is a new idea but can be justified with the growth of the sport. The prominent use of social media among powerlifters has allowed lifters to display their talents, connect with others, and for some, rise to prominence within the community. Given this current state of connection, comparisons between lifters that may not have been compared in competition have become commonplace and will likely continue for the foreseeable future. Further, having powerlifting scores that fairly distribute lifters by equipment division may provide justification for exhibition meets in which raw/classic lifters compete with equipped lifters. It also will allow for a given lifter to compare his/her performance in each of the divisions. With these uses for an equipment adjustment in mind, the average biases for scores by equipment division can be compared to determine which score is fairest. Figure 4 of the supplementary materials displays a heatmap of average bias for scores by lifter classification on the basis of equipment. Scores tend to be fairly balanced for most skill levels, ranking lifters by equipment within each classification such that it closely matches the IPF makeup. However, DOTS, Wilks, and Updated Wilks hold a strong, favorable bias toward equipped lifters at the Elite and Master level. The average representation for Elite equipped lifters was approximately 70% higher than it should have been for all three of the aforementioned scores. Similar results are observed for the Master level lifters, the skill level immediately after Elite. The novel scores and GLP also favor equipped lifters, but only slightly. Average bias for these scores ranges from 11% to 14%. Thus, if the goal of a practitioner is to compare equipped and raw/classic lifters, a novel score such as the P2 or GLP should be used.

4.4 Performance of Scores at the 2019 IPF World Championships

In an effort to distinguish the performance of DOTS and GLP in fairly distributing Elite IPF lifters, a replication of Tim Konertz's record distance analysis from the 2019 IPF World Championships was completed. His methods were discussed briefly in Section 2.2. Konertz took the results for the 2019 IPF World Championships and calculated the distance from the IPF records needed to achieve top 3 or best lifter in each of the four separate sex and equipment divisions. He also evaluated the coefficient of variation for each score. His analysis was replicated (excluding the IPF World Bench Press Championships) and then scaled such that kg distance to world records was in percentages above/below the world record. Further, mean relative distance from the world record needed to win best lifter or place in the top 3 was also calculated for each score within each division. The results from this analysis are displayed in the supplementary materials in Figures 5 and 6, and Tables 2 and 3.

5 Discussion

Visual comparisons of the density curves for differences in test statistics suggested that novel scores yielded the best results in terms of matching bootstrap IPF samples when considering all lifter classifications collectively. The high positive differences from the Wilks Score χ^2 test statistics indicated that these clearly outperform Wilks on the basis of weight class, sex, and equipment. While novel scores were determined to be the least biased from a comprehensive perspective of IPF lifters, it is also necessary to note the difference in performance when evaluating Elite lifters. In Figures 1-4 displaying bias for 10,000 bootstrap samples by lifter classification, it was determined that GLP and DOTS were tied for the least biased by weight class and biological sex for Elite lifters. GLP alone was the least biased by equipment in samples of Elite lifters. The variable of highest importance is weight class, given that the best lifter is selected from all weight classes within sex and equipment divisions. The comparison of least interest is equipment, as it is not yet used in a competition setting. Despite the great performance for GLP for all lifter classifications, it appears virtually indistinguishable from DOTS in its ability to fairly distribute Elite IPF lifters within this computational study using bootstrap sampling.

Powerlifting scores evaluated in this analysis were able to provide fairer rankings of lifters than the Wilks Score. Novel scores, namely the P2, were able to outperform Wilks when trained on larger sample data that were more representative of all IPF lifters. Existing scores that were trained on samples of highly skilled lifters also outperformed Wilks, particularly GLP and DOTS. There are multiple reasons that these scores were able to provide better representation of IPF lifters than the Wilks Score. All of these were updated with new data that are representative of modern powerlifters, an advantage over the Wilks Score. Further, it also seems that model complexity and overfitting may have played a role in the superior fit of the three aforementioned scores given that all were derived from simpler models than the fifth-order polynomial Wilks Formula. GLP and P2 presented the best solutions to ranking Elite lifters and the collective IPF community, respectively. While DOTS is a highly regarded score, it fell short in performance to both the P2 and GLP because it was not selected based on predictive power [9]. Both GLP and P2 were models selected based on cross validation performance, meaning that these models were built to be accurate and account for future variability. Not only should practitioners explore predictive performance of statistical models but venture outside of the realm of polynomial regression. Modern statistical modeling and machine learning provide many alternatives to polynomial regression and nonparametric approximation. These methods are praised for being highly accurate despite lacking the accessibility that a regression model provides with a table of coefficients. Further, evaluations of powerlifting scores should be more careful in assuming the degree to which data are representative and continually update scores every couple of years. The Wilks Score should have been evaluated as soon as the IPF changed weight classes in 2011 and as the demographic of lifters shifted heavily toward classic lifting. This also calls for the uniformity of evaluation of scores. Though this analysis provides its own original method of evaluating powerlifting scores and replication of other methods, it differs from all other evaluations. Governing bodies such as the IPF should establish clear methods to evaluate and update their powerlifting scores in order to maintain standards of fairness in representation. This evaluation and updating of scores should take place every other year to maintain the integrity of the federation's standards. Until then, the powerlifting community must rely on analyses such as Tim Konertz's, the IPF's, and this one to address this issue. In conclusion, the IPF should continue to use GLP in national and international competition. If fairness for all skill levels is desired, a simpler novel score, namely the P2 should be used in local meets to ensure comparison of lifters is fair.

6 Disclosure Statement

The authors report no conflict of interest.

Table 3: 95% Confidence Intervals for Differences in χ^2 Test Statistics

Variable	score Comparison	(Lower Bound, Upper Bound)	Which is better?	Which is least biased?
Weight Class (kg.)	1. Wilks - DOTS	(540.561, 765.229)	DOTS	One of the Novel Scores
	2. Wilks - Updated Wilks	(5873.498, 6568.732)	Updated Wilks	
	3. Wilks - P5	(5685.911, 6418.911)	P5	
	4. Wilks - Goodlift	(709.822, 1184.533)	Goodlift	
	5. Wilks - P2	(5553.475, 6288.030)	P2	
	6. Wilks - P2SPL	(5699.043, 6424.217)	P2SPL	
Sex	1. Wilks - DOTS	(-170.974, -38.138)	Wilks	One of the Novel Scores
	2. Wilks - Updated Wilks	(4251.919, 4794.045)	Updated Wilks	
	3. Wilks - P5	(4100.307, 4673.410)	P5	
	4. Wilks - Goodlift	(201.596, 498.339)	Goodlift	
	5. Wilks - P2	(4103.415, 4674.216)	P2	
	6. Wilks - P2SPL	(4112.549, 4689.565)	P2SPL	
Equipment	1. Wilks - DOTS	(36.712, 210.675)	P5	Goodlift or one of the Novel Scores
	2. Wilks - Updated Wilks	(-115.025, 275.063)	Inconclusive	
	3. Wilks - P5	(12378.800, 13187.730)	P5	
	4. Wilks - Goodlift	(12563.750, 13366.670)	Goodlift	
	5. Wilks - P2	(12381.410, 13195.760)	P2	
	6. Wilks - P2SPL	(12378.290, 13185.840)	P2SPL	

Table 4: P2 Powerlifting Score Formulas

Demographic (Sex, Equipment)	Formula
Male, Classic/Raw	$300 * \frac{Total\ kg.}{-48.9+9.399BW-0.03119BW^2}$
Female, Classic/Raw	$300 * \frac{Total\ kg.}{78.10+4.405BW-0.01707BW^2}$
Male, Equipped	$300 * \frac{Total\ kg.}{-120.59+12.0009BW-0.0388163BW^2}$
Female, Equipped	$300 * \frac{Total\ kg.}{43.48+6.461BW-0.024541BW^2}$

References

[1] Per-Olof Astrand and Kaare Rodahl. *Textbook of work physiology New York*. NY: McGraw-Hill, 1986.

[2] L. Fabien. Telephone interview with r. wilks., 2008.

[3] International Powerlifting Federation. Ipf formula: Why it was time for a new ipf formula. `https://www.powerlifting.sport/rulescodesinfo/ipf-formula.html`, 2019.

[4] International Powerlifting Federation. Technical rules book of the international powerlifting federation. international powerlifting federation. `https://www.powerlifting.sport/fileadmin/ipf/data/rules/technical-rules/english/IPF_Technical_Rules_Book_2020.pdf`, 2020.

[5] Pierre-Marc Ferland, Marc-Olivier Allard, and Alain-Steve Comtois. Efficiency of the wilks and ipf formulas at comparing maximal strength regardless of bodyweight through analysis of the open powerlifting database. *International Journal of Exercise Science*, 13(4):12, 2020.

[6] Andrew Gelman and Guido Imbens. Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, 37(3):447–456, 2019.

[7] Hyeong Jeong, Myoungshic Jhun, and Daehak Kim. Bootstrap tests for independence in two-way ordinal contingency tables. *Computational Statistics and Data Analysis*, 48:623–631, 03 2005.

[8] Anna Khudayarov. Revising the wilks scoring system for pro raw powerlifting. *arXiv preprint arXiv:1903.10694*, 2019.

[9] Tim Konertz. The dots formula - the new wilks. https://drive.google.com/file/d/1GG22jYL3JaalTUydcigkZNCjddtH-0rv/view, 2019.

[10] Oleksandr Kopayev, Borys Onyshchenko, and Anatoliy Stetsenko. Evaluation of wilks, wilks-2, dots, ipf and goodlift formulas for calculating relative scores in ipf powerlifting competitions. https://www.powerlifting.sport/fileadmin/ipf/data/ipf-formula/Models_Evaluation-I-2020.pdf, 2020.

[11] Jake Lever, Martin Krzywinski, and Naomi Altman. Points of significance: model selection and overfitting. 2016.

[12] Richard C Liu and Lawrence D Brown. Nonexistence of informative unbiased estimators in singular problems. *The Annals of Statistics*, pages 1–13, 1993.

[13] Joe Marksteiner. Ipf points - proposed replacement for wilks coefficients. https://www.powerlifting.sport/fileadmin/ipf/data/ipf-formula/IPF_Points_Proposal.pdf, 2019.

[14] OpenPowerlifting. Openpowerlifting data.

[15] USA Powerlifting. Classification standards. https://www.usapowerlifting.com/wp-content/uploads/2014/01/Raw-Classifications-kg.pdf, 2020.

[16] Paul M Vanderburgh and Alan M Batterham. Validation of the wilks powerlifting formula. *Medicine & Science in Sports & Exercise*, 31(12):1869–1875, 1999.