Did Lennox Lewis Beat Evander Holyfield? Methods for Analyzing Small-sample Inter-rater Agreement Problems

David J. Algranati*
Daniel L. Cork†
Herbert K. H. Lee‡

August 16, 2000

CarnegieMellon
Department of Statistics
Technical Report #732

Presented at the Joint Statistical Meetings, Indianapolis, Indiana

Abstract

On March 13, 1999, a highly-anticipated prizefight between heavyweight champions Evander Holyfield and Lennox Lewis was ruled a draw by the three official judges for the bout. Many observers of the fight felt that Lewis had clearly outperformed Holyfield; dissatisfaction with the result—particularly the pro-Holyfield scorecard of judge Eugenia Williams—fueled speculation that the fight had been fixed and prompted local and state investigations. In this paper, we examine whether the official judges scored the fight in a significantly different way than other professional observers of the fight; we do so by comparing the official scorecards with those returned by other sources, including the fight's broadcast team and other sports media outlets.

We look at methods of analyzing the round-by-round scoring within the context of inter-rater agreement. The literature on inter-rater agreement typically considers a large number of samples rated by a small number of judges, and relies on asymptotic results for tests. In our case, the sample size is clearly too small for any asymptotics to apply. Instead, we investigate a number of techniques that can be applied to small-sample inter-rater agreement problems, including logistic regression, an exact test, and some Bayesian approaches.

Keywords: Boxing, Logistic Regression, Bayesian Statistics, Inter-Rater Agreement

*dja@stat.cmu.edu; Department of Statistics and Heinz School of Public Policy and Management, Carnegie Mellon University, Pittsburgh, PA 15213

†dcork@nas.edu; Committee on National Statistics, National Research Council, Washington, DC 20418

‡herbie@stat.duke.edu; Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708

1 Introduction

Boxing is a sport with a history rich in legendary champions and thrilling battles. Unfortunately for boxing, that history is also ridden with controversies and scandals. At times, scandal erupts from actions by fans, most notably the 1993 "Fan Man" incident in which a paraglider crashed into an outdoor Las Vegas boxing ring, disrupting the second fight between Riddick Bowe and Evander Holyfield. At other times, actions of referees are the impetus for controversy. For instance, the infamous "long count" in 1927 may have cost Jack Dempsey his chance to regain a championship during his second bout with Gene Tunney. More recently, in 1990, Nevada referee Richard Steele elected to halt the first Meldrick Taylor versus Julio Cesar Chavez bout with only two seconds remaining in the contest, giving Chavez a victory by technical knockout. Had Steele let the bout end, Taylor would have won the fight based on his substantial lead on the judges' scorecards.

Perhaps the most frequent source of controversy in boxing is the scoring of the bouts. Unless a boxing match ends in a knockout or other stoppage, the fight is decided based on the scorecards of three official judges. The combination of many viewers' unfamiliarity with the boxing scoring system, the somewhat arcane nature of that scoring system, and the passions that observers may feel for one fighter over another frequently leads to judges' decisions being hotly contested. "There is a lot of things in boxing you could complain about," trainer/commentator Teddy Atlas noted during a recent boxing broadcast, "but bad decisions are killing this industry" (ESPN2, 2000). Indeed, legendary "bad decisions" have tainted boxing for decades. In 1947, Jersey Joe Walcott appeared to score an easy victory in his first bout with Joe Louis, knocking the aging "Brown Bomber" down twice in the early rounds. However, two ringside judges gave Louis the victory—a decision that sports columnist Shirley Povich (1994) lambasted as "first-degree robbery" and "the great Madison Square Garden holdup". More recently, the 1993 fight between Julio Cesar Chavez and Pernell Whitaker was judged a draw despite the fact that most observers found it difficult to believe that Chavez had won more than a couple of rounds.

In 1999, boxing fans were offered the chance to experience two highly anticipated prizefights, each of which have subsequently been dubbed "bad decisions" (to varying degrees) in the media and public opinion. On March 13, American heavyweight Evander Holyfield and British heavyweight Lennox Lewis met in New York's Madison Square Garden to contend for the first undisputed, unified heavyweight championship in over five years.[1] To many observers, Lewis appeared to outfight a sluggish Holyfield; however, when the judges' scorecards were read, the fight was ruled a draw. British judge Larry O'Connell scored the fight even, while South African judge Stanley Christodoulu's scorecard favored Lewis. Most controversially, judge Eugenia Williams of New Jersey declared Holyfield the winner, seven rounds to five; among the rounds that Williams awarded Holyfield was the fifth round, in which Lewis appeared most dominant of a tentative, defensive Holyfield. A tremendous outcry over this seemingly wrong decision touched off investigations within boxing circles and in New York's state government, and virtually forced a rematch between the fighters in November 1999 (which Lewis won by decision, though neither fighter was thought to have turned in a stellar performance).

Several months after the first Lewis-Holyfield bout, boxing's best welterweight (147 pounds) fighters—Oscar de la Hoya and Felix "Tito" Trinidad—put their undefeated records on the line on September 18 in Las Vegas. The fight, marketed as the "Fight of the Millennium", promised to be not only a clash of highly skilled boxers but a particularly heated war between two of boxing's most powerful promoters; Trinidad is promoted by Don King, while Top Rank's Bob Arum had backed the "Golden Boy" de la Hoya through his professional career. The fight ended with a majority decision favoring Trinidad, a ruling that failed to disturb some observers due to de la Hoya's unusual tactic in the late rounds of the fight. Apparently

[1] "Undisputed" refers to a fighter who is recognized as the champion in a weight division by all three of boxing's major sanctioning bodies.

believing himself to be well ahead of Trinidad on the judges' scorecards, de la Hoya largely kept away from Trinidad in the last three rounds, throwing relatively few punches and avoiding any possibility of a Trinidad victory by knockout. Still, the judges' decision was assailed as a "bad decision" by some in the press who agreed with de la Hoya and his cornermen that the "Golden Boy" had won enough early rounds to secure a victory.

Such "bad decisions" are usually believed by boxing observers and insiders to be the result of underhandedness or foul play, usually by the promoter of the bout. Several sordid episodes in boxing's past make it easier to ascribe corrupt practices to the sport. In 1960, Senator Estes Kefauver's commission on organized crime explored the influences of organized crime on boxing, hearing testimony from fighters who said that they were offered money to throw fights (Newfield, 1995). In 1977, the ABC network abruptly canceled its plans to televise a U.S. Boxing Tournament when it was learned that the man in charge of *Ring* Magazine's fighter ratings—the criterion for entering the tournament—had received \$5,000 in payments from boxing promoter Don King (Newfield, 1995). And, most recently, the International Boxing Federation (IBF) sanctioning body was indicted on racketeering charges, accused of taking over \$300,000 in exchange for manipulating their rankings of fighters; promoter Bob Arum was fined and sanctioned by the Nevada State Athletic Commission after testifying in the IBF case that he had paid \$100,000 to secure the IBF's approval of a 1995 prizefight (Juipe, 2000).

Allegations of influencing judges are not limited to one promoter, and have been leveled by both commentators and the boxers themselves. Former heavyweight champion and current television boxing commentator George Foreman suspected internal favoritism in Michael Moorer's 1994 majority decision win over Evander Holyfield, since both fighters were promoted by the Duva family's Main Events Promotions. "The Duva people have got too many pieces of paper on too many different fighters," Foreman noted on-air. "Evander should have won that fight" (Pay-Per-View, 1994). In 1999, Johnny Tapia lost his title in a close, exciting fight with Paulie Ayala; asked for his comments immediately after the fight, Tapia implied that the scoring of the bout was influenced by the fact that Ayala was still under contract with Bob Arum's Top Rank Promotions while Tapia's Top Rank contract expired with the Ayala fight. "You know what happened. Everybody knows what happened," Tapia mused. "It was my last bout with Top Rank" (SET Pay-Per-View, 1999). As a final example, boxing writer Wallace Matthews (1994) envisioned a broad conspiracy in Julio Cesar Chavez's technical split decision victory over Frankie Randall in 1994, saying that the decision should "go down on the police blotter as yet another heist masterminded by the Unholy Trio of Chavez and his two henchmen, Don King and (president of the World Boxing Council [WBC]) Jose Sulaiman."

Allegations of incompetence or foul play are easy to cast around, but they are very difficult to prove. Boxing has no central panel or review board to investigate the decisions of world championship bouts. Moreover, the investigation of "bad decisions"—and, for that matter, statistical analyses of the sport of boxing in general—are hampered by the subjective, rather imprecise nature of boxing scoring. The intangible factors that lead to boxers winning rounds are hard to quantify and easy to misconstrue. In this paper, we use data on 1999's landmark "bad decisions"—the first Holyfield/Lewis fight and the de la Hoya/Trinidad bout—to explore the variability in boxing scoring and address the quality of the decisions.

The data we consider here are the round-by-round scorecards for both bouts returned by the official judges, augmented by round-by-round scorecards reported by broadcast teams and other media outlets. The scorecards compiled by unofficial observers provide some frame of reference in assessing how extreme the official judges' scoring of particular rounds may be; more colloquially, they can help gauge whether the official and unofficial observers were all "watching the same fight". Admittedly, there are valid criticisms that can be leveled at media scorecards of boxing matches—for instance, fight broadcasters must necessarily divide their concentration between the action in the ring and the mechanics of providing interesting live commentary (Feour, 2000), and reporters may converse (and be swayed by) peers during the fight unlike the official judges. Still, professional boxing reporters have experience in watching and critiquing live boxing matches, and utilizing their round-by-round judgments—even though their training differs in most cases from the official judges—is a reasonable first step in examining variability in scoring.

Cast in this light, our problem is a case of inter-rater agreement—a concept in statistics that seeks to analyze agreement and consistency among a set of raters in evaluating a common set of subjects. Much of the current inter-rater agreement literature is predicated upon a small number of raters relative to the number of observations, and relies on asymptotic results to perform reliable inference. In the context of boxing scoring, the number of raters (official and unofficial judges with round-by-round scores) is roughly equal to the number of observations (at most twelve rounds, in a modern world title fight), which in turn is sufficiently small as to cast doubt on asymptotic results. Accordingly, we explore a number of methods for assessing inter-rater agreement with a small number of samples and a relatively large number of raters—in this case, ten sets of scores for twelve rounds of boxing.

Our analysis proceeds in Section 2 with further detail on the mechanics of boxing scoring. In Section 3, we outline methods for dealing with inter-rater agreement, including exact tests, logistic regression, and a Bayesian formulation. Sections 4 and 5 apply these methods to the Holyfield/Lewis and de la Hoya/Trinidad fights, respectively, while Section 6 offers concluding comments and directions for future research.

2 The Boxing Scoring System

The Marquess de Queensbury rules, which form the foundation of modern boxing, specify only the most elemental aspects of the sport—the contest is divided up into rounds, the combatants are to wear gloves, and so forth. The exact rules governing the conduct of a particular boxing match depend in part on the official sanctioning body authorizing the fight and on the boxing or athletic commission of the state or nation where the fight takes place.

Boxing's three major sanctioning bodies—the IBF, WBC, and World Boxing Association (WBA)—have standardized certain rules for world title bouts. Under these uniform rules, championship bouts are scheduled for twelve, three-minute rounds. Three judges are appointed to score the bout on the "10 point must" system, which we discuss in more detail later in this section. The judges are seated along the edge of the boxing ring; they do not sit together and cannot consult one another. After each round, the judges submit their score for that round to the referee (who does *not* score the bout), and the referee gives the scores to a clerk at ringside to record and tally them.

If the fight goes the full twelve-round distance without being stopped due to a knockout, disqualification, or other reason, then the fight outcome is decided based on the judges' scorecards. Each judge's round-by-round scores are summed together; if Judge A has more points for Boxer X than for Boxer Y, then Judge A has scored the bout for Boxer X (the margin in points by which Boxer X beats Boxer Y on Judge A's scorecard does not matter). Alternately, Judge A's scorecard can show victory for Boxer Y (if Boxer Y has more points) or an even contest (Boxers X and Y have equal points).

Based on the individual judges' totaled scores, Boxer X wins the fight if two of the ringside judges score the bout in his favor. A draw between the two boxers can result in one of two ways: either at least two of the judges score the bout even (the same amount of points for Boxer X and Boxer Y), or one judge scores the bout for Boxer X, a second judge scores it for Boxer Y, and the third scores the bout even.

2.1 Scoring Criteria

In determining which boxer wins a particular round, judges are instructed to consider four criteria. The first is the number of *clean punches landed* by each boxer, where clean punches are those thrown with a clenched fist that land above the waist on the stomach, chest, and the front or side of the head. Typically, punches to the shoulders, elbows, arms, or gloves of the opponent do not count as a clean punch landed (because the opponent has blocked the punch). The second criterion is *defense*, the avoidance of punches via blocking, bobbing and weaving (movement of the head and waist), and good foot work (Kaczmarek, 1996). Third, boxers are evaluated for their *effective aggressiveness*—whether the fighter forces the action of the bout by pursuing his opponent and initiating action, while landing punches. A final criterion is the concept of *ring generalship*. Professional boxing judge Tom Kaczmarek (1996) observes that "ring generalship applies to the fighter who uses skills beyond his punching power to control the action in the ring"; this includes the use of feints and footwork, as well as the ability to make the opponent fight a style he would prefer not to fight. A judge's determination of how well a boxer has satisfied these criteria must be based only on what that judge observes and believes; for instance, while commercial parties compute breakdowns of punches thrown and landed and these statistics are frequently cited by on-air broadcasters, they are neither official fight statistics nor are they available to the ringside judges.

That there are four criteria for judging whether a boxer won a round is clear; how these criteria should be weighted in making the decision is not. A Web page on boxing scoring hosted by the cable network HBO and authored by former professional boxing judge Harold Lederman—who unofficially scores fights for the network's boxing telecasts—implies that "each of these criteria is supposed to count for 25%". But the page then contradicts the premise of equal weights on the criteria, noting that "all good judges will tell you that clean punching counts a lot more than everything else in scoring a round of professional boxing".[2] Kaczmarek (1996) considers effective aggressiveness and clean punches landed as one category, and deems this composite category the most important scoring criterion; ring generalship is the next most important criterion to Kaczmarek, followed by defense. In the absence of clearer standards, it appears that individual judges—like observers scoring boxing matches on their own—are free to weight the criteria as they choose. Consequently, different boxing styles and interpretations of the criteria introduce variability into judges' scorecards and the resolution of bouts.

2.2 Scoring Using the 10 Point Must System

After considering the criteria described in Section 2.1, a judge must assign a score for a particular round in a boxing match using what is known as the "10 point must" system. Under this system, the winner of the round receives ten points, and the loser of the round receives nine points or less. In the event that the judge decides that each boxer is equally deserving—that there is no "loser" of the round—the judge has the discretion to score the round "even", assigning 10 points to each fighter.

On the surface, the ability to assign the loser of a round "nine points or less" seems to give individual judges wide latitude in their scoring; in practice, this is not the case. Kaczmarek (1996) suggests that a round won by a "'close' to a 'definite' margin" should be scored 10–9. Rounds are almost always scored 10–9, unless a fighter is knocked down during the round, in which case the round would be scored 10–8. Were a fighter to be knocked down by his opponent two or three times, the corresponding score would be 10–7 and 10–6, respectively.[3] Judges vary in their propensity to score rounds even (10–10); additionally, if one fighter thoroughly dominates his opponent but does not score a knockdown, some judges will score the round 10–8.

Another factor which impacts scoring is that the referee is empowered with the ability to penalize fighters for rules infractions. These penalties are typically one point off the offending fighter's score in the round in which the infraction occurred. Typically, point deductions will only occur for repeated infractions; the referee will usually warn the offending fighter unless the first infraction is particularly egregious.

[2]The Web page is accessible at `http://www.hbo.com/boxing/cmp/scoring.shtml`.

[3]Kaczmarek (1996) describes 10–6 rounds as "very rare". Based on watching over 4,000 boxing matches, the authors have knowledge of only one round which was scored 10–6. Typically, bouts are stopped if a fighter has been knocked down three times in one round.

An important feature of the boxing scoring system is that "each round should be judged as an isolated event, regardless of what has transpired earlier" (Kaczmarek, 1996). Once a judge turns in a score for a particular round, the slip with the score is turned over to a ringside official by way of the referee. Because they are not keeping a running scorecard in front of them, judges do not have the opportunity to change scores they assigned in previous rounds or to base their scores on previously-scored rounds. Indeed, judges may only become aware of the implications of their aggregate scores at the same time the public does—when they are announced at the end of the fight. In the Holyfield-Lewis bout, British judge Larry O'Connell maintained after the fight that he was as surprised as anyone when his scorecard turned out to show a draw. "You might get a gut feeling about how the fight is going, but you have to ignore it," O'Connell said in an interview. "You keep your score for each round. You can't add up the totals—that's unethical—so I didn't know what my final score was until I heard it read out" (Uncredited, 1999).

2.3 Discrepancies in Scores

Though they watch the same fight, weigh the same criteria, and follow a common system of scoring, the scorecards turned in by official ringside judges can vary widely—and the official judges' opinion may be wildly at odds with the verdicts rendered by boxing commentators, journalists, and the public. Why do these discrepancies exist?

Perhaps the greatest cause of discrepancy between the official scores of a bout and other observer's "scores" is that many people are not scoring the fight round-by-round, on the 10 point must system.[4] Instead, many people may watch a bout and form a general impression of the contest; they may rely on a gut feeling about which boxer won a round rather than considering all four criteria. One can readily conceive of scenarios under which failure to commit round-by-round scores under the 10 point must system can clash with more general perceptions a fight's winner. For example, Boxer A wins 5 rounds by a "definite" margin, and Boxer B wins 7 rounds by a "close" margin. On the 10 point must system, Boxer B wins the bout, 115-113, while an observer just making one, overall summary impression is likely to favor Boxer A.

Others—particularly viewers of televised boxing broadcasts—may be swayed by punch statistics, such as the summaries of "jabs" and "power punches" thrown and landed that are tallied by the company CompuBox and featured on ESPN and HBO broadcasts (Hauser, 1999). These punch statistics are not official (the judges do not consider them in making their decisions) and are recorded from a vantage point different from (and perhaps inferior to) the ringside location of official judges. Moreover, they reflect—at best—only two of the four scoring criteria (clean punches landed and defense).

The most striking, and probably most costly, example of such scoring fallacies might be the tactical errors that many believe transpired in Oscar de la Hoya's corner during his fight with Felix Trinidad. By the last three rounds of the fight, de la Hoya and his cornermen believed themselves to be safely ahead on points on the judges' scorecards; "I thought I had it in the bag," de la Hoya told a post-fight press conference (Fernandez, 1999). After the fight was decided in Trinidad's favor, de la Hoya and his advisors frequently cited the punch statistics as evidence that they had been wronged. "How does the guy who landed 100 less punches win the fight?" asked veteran trainer Gil Clancy, who served in de la Hoya's corner. "How does that happen?" (Kriegel, 1999). To another writer, Clancy argued that de la Hoya "landed more jabs and a higher percent of power punches. I always thought the business of boxing was about hitting the other guy more than he hits you. I do not understand the judges' decision" (Izenberg, 1999). Specifically, the CompuBox statistics showed that de la Hoya landed 41% of his total punches (263 of 648) compared to Trinidad's 36% (166 of 462); but, by "[believing] that punch statistics are as irrefutable as second-grade

[4]Or, if many people are performing such scoring, they are not making it publicly available. The two bouts analyzed in this paper were two of the highest-profile bouts in 1999, yet we were only able to find a handful of scores which were broken down, round-by-round.

arithmetic answers"—believing them to be a true and indisputable indicator of who won the fight—de la Hoya made a terrible "miscalculation" (Fernandez, 1999).

Other discrepancies in boxing scoring may arise due to the lack of standardization in applying the scoring criteria. As mentioned in Section 2.1, the judges may assign different weights to the different scoring criteria. Another problem is that while clean punches landed and defense are fairly objective, effective aggressiveness and ring generalship contain subjective elements (e.g., the point at which a boxer's aggression becomes "effective" is a decision each judge makes individually). Even the seemingly objective criteria of clean punches landed and defense could be legitimately viewed differently by judges due to the fact that the judges are situated at different points at ringside. When a fighter's back is to a judge, it becomes more difficult to discern if punches are being landed or blocked.

As mentioned in the introduction to this paper, a common belief is that discrepancies in scores arise because the judges are either incompetent or corrupt. As a general rule, the more important a bout, the more senior and experienced the judges are who are assigned to that bout. This would seemingly decrease the chances that the referees in important bouts, such as the ones examined in this paper, are incompetent.

3 Methods for Analyzing Small-sample Inter-rater Agreement

In this paper, we consider round-by-round scorecards recorded by both official boxing judges and media observers. This is a problem of inter-rater agreement that differs substantially from more common problems in the field. Much of the extant literature on inter-rater agreement deals with agreement matrices—structures with one dimension per rater so that, for example, cell [1,1,2] of the matrix indicates the number of times that the first two raters awarded a score of 1 and the third rater awarded a score of 2. Statistics like Kappa (Cohen, 1960) are then applied to measure agreement.

Such agreement matrices can be very useful when the number of raters is small, and the number of observations is sufficiently large to invoke asymptotic arguments typically used to evaluate the size of Kappa. But agreement matrices become less useful as the number of raters increases (making the matrix hopelessly sparse) or when the number of observations is small. In our case, the number of raters is roughly equal to the number of observations, which in turn is small; we must consider alternate methods.

The methods we consider and apply in this paper are all designed to examine the legitimacy of individual scorecards. They vary in complexity in their assumptions; by performing the batter of procedures, we are—in a sense—exploring the robustness of the individual tests' findings. We begin with frequentist and Bayesian "exact" tests that assume a particularly simple structure, yet still yield probabilities for individual scorecards. We then add more realism to the modeling with a logistic regression model that allows for individual judge effects and biases to be estimated; we fit both a classical logistic regression model as well as a Bayesian analog. Finally, a direct Bayesian approach assumes underlying latent variables that correspond to the magnitude by which a fighter won a round; this approach allows us to estimate individual judges' biases and to account for rounds scored even. These methods are described in more detail in the following sections.

3.1 Exact Tests

A basic approach to modeling boxing scorecard data begins by making two assumptions: that no even (tie) rounds are permissible and that the probability that a fighter wins a particular round depends only on the round, not on the judge. The first of these assumptions is not terribly unrealistic—as discussed in Section 2.2, boxing scoring is effectively binary since most rounds in boxing are scored either 10–9 for Fighter A or 10–9 for Fighter B. Indeed, only six of the 252 round-scores in the datasets we use in this analysis are scored even. The second assumption—that there are no individual judge effects—is less realistic, but we will add this layer of complexity later.

Under these assumptions, we can model each judge's score on each round as conditionally independent Bernoulli variables. Let x_{ij} be 1 if judge j scores round i for Fighter 1, and 0 if they score it for Fighter 2. We can estimate p_i, the probability a judge will score round i for Fighter 1, by the proportion of the judges that score round i for Fighter 1. Hence, following the ideas of Fisher (1935) (such as in his tea-tasting example), we can compute the probability of observing each sequence of scores (scorecard), and see how unlikely the particular scorecard of a judge is. Since there are only twelve rounds in a championship boxing match, we can easily program a computer to compute the probabilities of all 2^{12} possible scorecards and compute test statistics and p-values.

The key concern is computing the probability of observing a scorecard as "extreme" as that given by a particular judge. There are two obvious ways in which we could define "extreme", and we will explore both of them. First, a scorecard could be extreme in that it has at least as many rounds scored for Fighter 1 as the scorecard in question (or symmetrically for Fighter 2). Alternately, a scorecard could be considered extreme if it has at least as many rounds scored for Fighter 1 and it has a probability of being observed that is no higher than that of the scorecard in question.

There is a Bayesian analog of the above analysis. We model the rounds as independent and the judges as exchangeable. We treat the scores for round i as Bernoulli with probability p_i and put a prior on the p's. In the case of a uniform prior, we can get the posterior for p_i in closed form and again compute the probability of seeing a scorecard as extreme as that given by a particular judge. For more complicated, hierarchical models for p_i, we can use Markov Chain Monte Carlo (MCMC) either to get point estimates for p_i (followed by again computing all possible scorecards), or to directly estimate the probability of observing as extreme a scorecard as that of a particular judge.

3.2 Logistic Regression

We then consider methods that continue to assume that no even rounds are permissible but allows for individual judge effects to be estimated. Following the methods of Agresti (1990), we fit a logistic regression model to the scores with factor covariates for the judges and the rounds. Let p_{ij} be the probability that judge j scores round i for Fighter 1. Let α_i be an indicator covariate for round i (omitting the first round for identifiability), and let β_j be an indicator covariate for judge j (also omitting the first judge). Then the model is

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \tag{1}$$

where μ is an overall level and ε is the error term. After fitting the model, we can use the fitted values for β_j to compare the judges. Judges with unusually large or small β's could be seen to be biased in favor of one of the fighters. For identifiability, we set one α_i and one β_j to zero.

In evaluating the coefficients to attempt to determine significance or select a model, we typically need to rely on asymptotic approximations. Clearly these are inappropriate for a sample as small as ours, so that we cannot assume normality of the fitted coefficients or a chi-square distribution for the deviance. However, McCullagh and Nelder (1989) point out that the chi-square approximation for the difference in deviance (for comparing models) is usually still good even when the assumptions are not met. Our experience with these models shows no evidence to the contrary, so we rely on this approximation for valid analysis of classical logistic regression models.

This model has the obvious Bayesian analog, where a hierarchical prior is used for the parameters. One can then compare the posteriors for the β's to check for possible biases in the scores of the judges. The

Bayesian approach has the additional advantage in that MCMC provides a direct estimate of the posterior for parameters, so that we do not need to rely on normality assumptions for them.

In theory, the generalized linear model framework allows us (in both the classical and Bayesian cases) to add a level to the model to deal with ties in the scores. However, the relative scarcity of ties in this already small sample makes it infeasible to incorporate the ties in the data into analysis with these methods.

3.3 A Direct Bayesian Model

Our final approach follows the approach of Johnson (1996) and Johnson and Albert (1999). Consider a latent variable that measures the true difference in performance between the two fighters in a boxing match in a particular round—that indicates the "true" winner of the round. Under this framework, an individual judge's score for that round reflects their perception of that underlying difference in performance; assessment can be made as to whether an individual judge is consistently biased in favor of one fighter or another.

For each round i, denote the difference in performance between Fighter 1 and Fighter 2 by y_i, and let $y_i \sim N(0,1)$. The scale is arbitrary, and we fix the standard deviation at 1 for identifiability. If both fighters are truly even in performance in the round, then $y_i = 0$. Large, positive y_i suggest dominance by Fighter 1, while a negative y_i suggests that Fighter 2 scored the best performance.

Obviously, y_i is unobserved; instead, what we observe are individual judges' perceptions of the value of y_i. For judge j, we model the perceived difference in performance on round i as $y_{ij} \sim N(y_i + b_j, \sigma_j^2)$, where b_j is a bias term for judge j. If $y_{ij} > 0$, judge j will score round i for Fighter 1, i.e., $x_{ij} = I_{\{y_{ij}>0\}}$. Finally, we put priors on b_j and σ_j^2 to get the full model

$$\begin{aligned}
y_i &\sim N(0,1) \\
y_{ij} &\sim N(y_i + b_j, \sigma_j^2) \\
b_j &\sim N(0, \tau^2) \\
\sigma_j^2 &\sim \Gamma^{-1}(a_1, a_2)
\end{aligned}$$

where τ, a_1, and a_2 are hyperparameters chosen based on historical data (or chosen to reflect prior beliefs).

The model is easily fit with MCMC. One can then look at the posteriors for the various b_j and σ_j to compare the judges. Judges with b_j centered away from zero can be deemed to be "biased" in favor of one fighter. The variance parameter σ_j speaks to the consistency of an individual judge, in that a larger σ_j implies higher variability from the "true" result, or at least from the result implied by the other judges.

Frequentist versions of this model also exist (Quinn, 1989; Henkelman et al., 1990; Uebersax and Grove, 1993). These tend to involve latent variables for class membership, where membership is determined by a score variable analogous to x_i above. The problem with the frequentist approach here is that more data are necessary to estimate all of the parameters of the model than in the hierarchical model case. In fact, they generally require multiple observations for at least some of the combinations of covariates (here, judges and rounds). Since we only have one observation from each judge for each round, this form of model is not appropriate for this paper.

The terminology used to describe this model raises a subtle point that should be made clear. "Bias" is a somewhat unfortunate term, given that it could be interpreted pejoratively. We emphasize that we use term "bias" in accordance with the usual statistical definition of the term. In noting that a particular judge's fitted bias favors one fighter, we emphasize that we mean that their scorecard tends to favor one fighter relative to the true latent difference in performance y_i; it should *not* be interpreted as saying that this deviation from the latent y_i is necessarily the result of any corrupt or improper behavior.

	Individual Round Scores												
	1	2	3	4	5	6	7	8	9	10	11	12	**Total**
Official Judges													
Williams	L	L	H	H	H	L	L	H	H	H	H	L	115-113
O'Connell	L	L	H	L	L	H	E	H	H	E	H	L	115-115
Christodoulu	L	L	H	L	L	L	L	H	H	H	E	L	113-116
Media & Unofficial Scorecards													
HBO-Lederman	L	L	H	L	L	L	L	L	H	H	L	L	111-117
Sportsticker	L	L	H	L	L	L	L	L	H	H	L	L	111-117
Boxing Time	L	L	H	L	L	L	L	L	H	H	L	L	111-117
Sportsline	L	L	H	L	L	H	L	L	L	H	L	L	111-117
AP	L	L	H	H	L	L	L	L	L	H	L	L	111-117
ESPN	L	L	L	L	L	L	L	L	L	H	L	L	109-119
Glyn Leach	L	L	H	E	L	E	L	H	H	H	H	L	115-115

TABLE 1: Scorecards for the First Holyfield-Lewis Fight. Rounds in favor of Holyfield denoted "H", in favor of Lewis denoted "L", and even rounds "E". All non-even rounds scored 10–9 in favor of indicated fighter.

3.3.1 Dealing with Ties

This final, direct Bayesian approach is particularly advantageous because it provides a natural way of dealing with the few observed ties (even rounds) that occur in boxing scoring data. Specifically, a cut-off value t can be introduced, defining an interval around $y_{ij} = 0$ wherein a judge can decide that the two fighters were evenly matched and score the round even. Judge j scores round i for Fighter 1 if $y_{ij} > t$, scores the round as a tie if $-t \leq y_{ij} \leq t$, and scores the round for Fighter 2 if $y_{ij} < -t$. While we might want to use separate values of t for each judge, ties are relatively rare, and there is not sufficient data to estimate all these parameters. We elect to set a common t across all judges, and put an exponential prior on t.

4 Analysis of the Holyfield-Lewis Fight

Scorecards for the Holyfield-Lewis Fight are shown in Table 1. In the table, an "H" indicates a round for Holyfield, an "L" a round for Lewis, and "E" an even round (tie). The total score reported in the final column lists the score for Holyfield first, Lewis second. The first three rows are the scorecards of the official judges—Eugenia Williams of the United States, Larry O'Connell of the United Kingdom, and Stanley Christodoulu of South Africa. The last seven lines are other professional observers of the fight, including wire service reporters from Sportsticker and the Associated Press (AP). HBO was the live broadcaster of the fight; the HBO-Lederman scorecard is the one compiled by the network's unofficial scorer, Harold Lederman. Glyn Leach is the editor of the British boxing magazine *Boxing Monthly*, which published his scorecard. Note that three of the unofficial judges—HBO-Lederman, Sportsticker, and Boxing Time—returned exactly the same scorecard, in favor of Lewis.

4.1 Exact Tests

The exact test methodology outlined in Section 3.1 does not allow for even rounds. Accordingly, we treat even rounds in Table 1 as rounds for Holyfield; since Lewis was perceived as the winner by the media and the public, failure to score these rounds for Lewis could be reasonably interpreted as round victories for Holyfield.

As described in Section 3.1, we calculate the probability of scoring round i for Holyfield; with the full set of p_i, we can then compute the probability of observing each of the 2^{12} possible scorecards. We find that the probability of observing official judge Williams' scorecard (seven rounds for Holyfield, five for Lewis, in the exact order as on her scorecard) is only 0.0019. The probability of observing judge O'Connell's even scorecard is also 0.0019; treating even rounds as rounds for Holyfield, O'Connell's scorecard also features seven Holyfield rounds and five Lewis rounds, albeit permuted differently than on Williams' card.

The fundamental question is how extreme these two official scorecards are, with respect to all 2^{12} possible scorecards. We find that the probability of observing a scorecard with at least as many rounds scored for Holyfield as on the Williams and O'Connell cards is 0.0256 (aggregating the individual probabilities of 1586 possible scorecards). Under an alternate definition of "extreme", we find that the probability of observing a scorecard with at least as many Holyfield rounds as on the Williams and O'Connell cards, but with probability no greater than that of the Williams/O'Connell scorecards (0.0019), equals 0.0182 (based on 1585 scorecards, the only difference between the previous set being the sequence of rounds exactly equal to Glyn Leach's card). From this analysis, we see that both Williams and O'Connell have produced scorecards which are highly unlikely to be observed by chance (the p-value is less than 0.05).

We can also use the fitted probabilities to estimate the probability that Holyfield won at least six rounds—the probability that Holyfield tied or beat Lewis, assuming that round i is won by Holyfield with probability p_i and that no ties are permitted. Adding up the estimated probabilities of all possible scorecards with six or more rounds for Holyfield, we estimate this probability to be 0.1320. Accordingly, the likelihood that the bout should end in an outcome other than a Lewis victory is small but non-trivial; it is not totally unreasonable for a judge to have scored the fight a draw or for Holyfield, even though the exact combinations of scores registered by Williams and O'Connell are quite improbable.

Putting a uniform prior distribution on p_i and estimating p_i by MCMC—the Bayesian analog of the classical exact test just described—we find that the Williams and O'Connell scorecards are not quite as extreme as the classical test suggests. The posterior probability of observing a scorecard with at least as many rounds for Holyfield as Williams is 0.072, and the probability of at least as many rounds on a card with no higher posterior probability than that of Williams is 0.067. Note that these values are larger than the previous versions because the fitted values for p_i have been shrunk toward their prior mean of 0.5. Still, these probabilities are rather small, and we may legitimately question the objectivity of the Williams and O'Connell scores. Several hierarchical models and prior specifications on p_i were tried, and the results were similar.

4.2 Logistic Regression

The preceding exact tests are interesting, but do not account for individual judge effects. To achieve a more realistic model, we start by fitting a logistic regression as in Equation 1. We use a single parameter for rounds 1, 2, and 12, as all were scored unanimously for Lewis. We also use a single parameter for HBO, Sportsticker, and Boxing Time, as all produced identical scorecards. Finally, we set the coefficients for round 1 and for the HBO/Sportsticker/Boxing Time scorecards to zero for identifiability of the model (this coefficient was chosen because HBO is a common element to both this fight and the de la Hoya/Trinidad fight). An analysis of deviance table finds that the round effect is highly significant (p-value = 5×10^{-12}) and that after taking the differences in rounds into effect the coefficients for judges are still highly important with a p-value of 1.5×10^{-6}). The coefficients themselves are not extremely large in magnitude; ESPN has the strongest negative effect at -18.5 (t-value 58.2), while Williams, O'Connell, and Glyn Leach each register the largest positive coefficient of 4.21 (t-value 1.86). The main result of this model is that the judges vary significantly in their propensity to score rounds for a particular fighter.

	Posterior Mean	Posterior SD	Posterior Median	Credible Interval 2.5%ile	97.5%ile
Judge Effects β_i					
Williams	3.885	1.396	3.864	1.228	6.663
O'Connell	3.882	1.399	3.859	1.217	6.688
Christodoulu	1.975	1.418	1.983	-0.823	4.732
Sportsline	-0.344	1.568	-0.307	-3.500	2.643
AP	-0.301	1.574	-0.276	-3.515	2.718
ESPN	-4.384	2.164	-4.238	-9.027	-0.567
Glyn Leach	3.845	1.427	3.821	1.116	6.653

TABLE 2: Bayesian logistic regression results for Holyfield-Lewis I, with coefficients relative to the HBO/Sportsticker/Boxing Time scorecards.

Bayesian logistic regression was fit using MCMC, using 10,000 draws from the posterior after 15,000 burn-in samples. Results are shown in Table 2, and are similar to those under the classical logistic regression model. Again, judge effects are relative to the HBO/Sportsticker/Boxing Time scorecard, the effect of which is forced to equal zero. 95% credible intervals for the judge effects show that Williams, O'Connell, and Glyn Leach are more pro-Holyfield than HBO (the interval is strictly positive), and ESPN is more pro-Lewis (strictly negative).

The logistic regression model is less useful the exact test methodology in determining whether a particular judge's scorecard is markedly different from their peers—specifically, to assess whether judge Williams' scorecard is fundamentally out of line compared to all other observers. To test for a single judge, we fit the classical logistic regression model (Equation 1) with parameters for the rounds and a single parameter for fit a model with parameters for the rounds and a single parameter for the judge in question. A model of this sort for examining the scores of Williams yields a coefficient on Williams with associated p-value 0.0127, indicating that judge Williams did seem to produce a scorecard different from the rest of the sample.

Similarly, we can also fit a model to examine whether the official judges—collectively—are significantly different in their scoring than the unofficial judges. Again, we allow for the full set of round effects in the classical logistic regression model, but only a single "judge" effect corresponding to an indicator variable flagging whether the score was returned by an official judge or not. An analysis of deviance table in this case again shows there to be a significant difference, with the p-value being 1.3×10^{-4}.

4.3 Direct Bayesian Model

We now apply the direct Bayesian model of Section 3.3, first to the data treating even rounds as wins for Holyfield, and then to the full data set including ties. Though the analysis that incorporates ties is more realistic, we start with the analysis where ties are treated as Holyfield rounds to facilitate direct comparison of the fitted biases from the direct Bayesian methods with the results of the exact test and logistic regression procedures applied earlier.

For a data set this small, the choice of prior obviously has a great impact on the fitted values of the parameters. However, our experience is that the prior mainly affects the scale of the parameters. The relative sizes of the parameters, as well as the conclusions drawn from the fitted model, are fairly robust to the choice of prior. For the results reported here, we chose a precision parameter of $\tau^2 = 0.1$ for the biases, and inverse-gamma parameters $a_1 = 2.0$ and $a_2 = 0.04$ (prior mean 0.04) for the variance of the judges.

The key structural difference between this model and the preceding logistic regression models is the underlying latent variable measuring the fighters' true difference in performance. Hence, by examining the

	Mean Bias	Std Dev	$P(b_j > 0)$
Official Judges			
Williams	0.30	0.15	0.977
O'Connell	0.30	0.15	0.974
Christodoulu	0.09	0.14	0.758
Media & Official Scorecards			
HBO-Lederman	−0.14	0.15	0.165
Sportsticker	−0.14	0.15	0.172
Boxing Time	−0.14	0.14	0.163
Sportsline	−0.15	0.16	0.176
AP	−0.14	0.16	0.183
ESPN	−0.54	0.20	0.003
Glyn Leach	0.30	0.14	0.985

TABLE 3: Summary of Fitted Biases for Judges of Holyfield-Lewis

MCMC output for these y_i, we can estimate the underlying probability of Holyfield having won at least six rounds. This estimate works out to 0.0856. So we find that it is possible, although somewhat unlikely, that Holyfield really won or tied the bout. This gives us further reason to look more closely at the possible biases of the judges.

Table 3 shows the posterior mean bias for each judge, the posterior standard deviation, and the probability that their bias is larger than zero, i.e., that they are biased in favor of Holyfield. The table confirms—and brings into stark relief—the patterns that emerged in the logistic regression analysis. First, the three official judges do appear to have a bias toward Holyfield relative to the unofficial scorers. Second, official judges Williams and O'Connell have particularly pronounced estimated biases; the probability that each of their biases exceeded zero (favored Holyfield) is over 0.97. And, finally, strong estimated biases are not unique to the official judges—ESPN's parameter estimates emphatically favoring Lewis stand out as yielding the largest probability of non-zero bias and the largest estimated bias in magnitude.

When we incorporate the additional parameter t into the direct Bayesian model and account for ties, we obtain the results summarized in Table 4. As in the non-tie analysis, the choice of hyperparameters has an appreciable impact on the scale of the fitted parameters but has little impact on the conclusions drawn from the model.

We estimate the probability that Holyfield won at least six rounds at 0.0667. To put the mean judge biases in Figure 4 in more context, the mean value for the threshold for scoring a tie was 0.0357. We see that the Williams fitted bias now stands out clearly over the other "pro-Holyfield" judges, and that her scorecard is the most extreme in favor of Holyfield. The judges who scored rounds as ties (O'Connell, Christodoulu, and Leach) are seen to have less extreme biases when we take the ties into account. ESPN is seen to be unusually biased in favor of Lewis, to a larger magnitude than the bias of Williams.

In particular, a key result of accounting for ties is that O'Connell's scorecard now looks much more reasonable. He had scored round seven as a tie, whereas everyone else in the sample scored it for Lewis. Counting his round seven as a pro-Holyfield round makes his scorecard much less probable by chance than counting it more accurately as a tie. Correctly assessing the ties lets us see that Williams has the viewpoint most biased in favor of Holyfield. But, as the ESPN card suggests, it was possible for an expert observer to view the proceedings of March 13, 1999, even more strongly in Lewis' favor as Williams judged it in favor of Holyfield.

	Mean Bias	Std Dev	$P(b_j > 0)$
Official Judges			
Williams	0.32	0.15	0.983
O'Connell	0.13	0.19	0.762
Christodoulu	0.06	0.14	0.669
Media & Official Scorecards			
HBO-Lederman	−0.16	0.15	0.138
Sportsticker	−0.14	0.15	0.174
Boxing Time	−0.15	0.15	0.174
Sportsline	−0.15	0.17	0.186
AP	−0.15	0.17	0.184
ESPN	−0.51	0.19	0.001
Glyn Leach	0.22	0.13	0.953

TABLE 4: Summary of Fitted Biases for Judges When Accounting for Ties

5 Analysis of the De la Hoya-Trinidad Fight

We now replicate the analysis on a collection of scorecards from the Oscar de la Hoya-Felix Trinidad fight. Figure 5 shows the scorecards of the three official judges—Glen Hamada and Jerry Roth of the United States, and Bob Logist of Belgium—along with eight unofficial scorecards. HBO broadcast the fight, just as they had broadcast Holyfield-Lewis I, so Harold Lederman again provided the network's unofficial scorecard; the Associated Press and Sportsticker are also common sources between the two datasets. Two of the *Los Angeles Times*' sports columnists—Steve Springer and Tim Kawakami—both produced coverage of the fight that included round scores. One distinctive feature of the de la Hoya-Trinidad dataset is that it includes the score of another current professional boxing judge with world title fight experience, Daniel van de Wiele of Germany. Van de Wiele was not an official judge for de la Hoya-Trinidad. While van de Wiele was not an official judge for de la Hoya-Trinidad, he scored the bout as he viewed it and shared his scorecard with one of the authors via personal correspondence.

A "D" shows a round scored for de la Hoya, a "T" for Trinidad, and an "E" an even round (a tied round). Total scores are de la Hoya-Trinidad. Note that the official judges (and the unofficial professional who published a scorecard) tend to favor Trinidad while the media tend to favor de la Hoya. Total scores are closer than in the Holyfield-Lewis fight. There is more dissension on individual rounds, and round 3 is the only round on which all eleven scorecards agree. It seems that this fight was a closer fight, and it may not be unreasonable to have scored it either way. We may hope that our methods will come to a similar conclusion.

Note that there is only one tie in the entire data set, round one for official judge Logist. This tie is very important in practical terms; had he scored this round for de la Hoya, his overall score would have been a tie score, and the bout would have been ruled a draw by majority decision. In applying our methods that preclude the use of even rounds, we consider Logist's even round as a round in favor of Trinidad (since it can be argued that this round-score is the one that won the fight for Trinidad); of course, in our final Bayesian analysis, we will handle the even round directly.

Using the exact test methodology to investigate the scorecards of Logist and Roth shows that the probability of observing at least as many rounds for Trinidad (or tied) is 0.21, while performing the simple Bayesian analysis to generate round probabilities to use in the exact test yields a probability of Trinidad draw or victory of 0.26. Hence, the pro-Trinidad scorecards of two of the official judges are not extreme by any

	Individual Round Scores												
	1	2	3	4	5	6	7	8	9	10	11	12	**Total**
Official Judges													
Hamada	D	T	D	T	D	D	D	T	D	T	T	T	114-114
Logist	E	T	D	T	D	D	T	D	T	T	T	D	114-115
Roth	T	T	D	T	D	D	D	D	T	T	T	T	113-115
Media & Unofficial Scorecards													
HBO-Lederman	D	D	D	T	D	T	D	D	T	T	T	T	114-114
AP	D	T	D	D	D	D	T	D	D	T	T	T	115-113
USA Today	D	D	D	T	D	D	T	D	D	T	T	T	115-113
Las Vegas Review-Journal	D	T	D	T	D	D	D	D	D	D	T	T	116-112
Los Angeles Times-Springer	D	D	D	T	D	D	T	D	D	T	T	T	115-113
Los Angeles Times-Kawakami	D	T	D	T	D	D	D	D	D	T	T	T	115-113
Sportsticker	T	D	D	T	D	D	D	D	D	T	D	T	116-112
van de Wiele	D	T	D	T	T	D	D	D	T	T	T	T	113-115

TABLE 5: Scorecards for the de la Hoya-Trinidad Fight. Rounds in favor of Trinidad denoted "T", in favor of de la Hoya denoted "D", and even rounds "E". The first score listed under "Total" is for de la Hoya, the second for Trinidad. All non-even rounds scored 10–9 in favor of indicated fighter.

reasonable standard; the probability of observing scorecards even more partial to Trinidad is appreciable.

Fitting a logistic regression using both round and judge as covariates finds that while the effect of round is significant, the effect of judge is not, and none of the coefficients for the individual judge effects has a significant t-value. Moreover, fitting the model for comparing a single judge (say, Roth) to the others also leads to an insignificant judge effect, as one might expect. Thus, logistic regression analysis suggests that there is not enough evidence to show that any judge was significantly out of line from the others.

Comparing the official judges to the rest of the sample also leads to an insignificant judge effect, although the p-value of the chi-squared test is borderline insignificant at 0.06. Upon further review of the data, the most pro-Trinidad result of the non-official scorecards was that of van de Wiele, who has served as a professional judge in other fights. Regrouping van de Wiele with the official judges does lead to a significant model for a judge effect (p-value 0.0117). Thus, there is some evidence that people who are professional judges scored this fight differently than did people who are professional media fight observers—an intriguing result, but one whose origins the data alone can not unravel.

A Bayesian logistic regression also shows no large differences between the judges. Table 6 shows that all of the 95% credible intervals overlap zero (and each other). Again, the coefficients are relative to the HBO-Lederman scorecard, which has been given a coefficient of zero for identifiability.

Finally, we fit the direct Bayesian model, which is configured such that a positive underlying difference in performance—and a positive bias—suggests a Trinidad victory, while negative differences and biases favor de la Hoya. The estimated biases from this model are shown in Table 7. We find the fitted biases are all small in magnitude and have large standard deviations; again, no single judge stands out as unduly partial to one fighter. The single largest estimated bias in the collection of scorers is in the *Las Vegas Review-Journal* scorecard, for which the probability of a de la Hoya bias is 0.875. But, again, the size of this probability and of the fitted bias does not suggest any form of impropriety. There is enough variability in the data to make all of the observed scorecards reasonable. The estimated probability of Trinidad having won at least six rounds is 0.379, not far from one-half.

	Mean	SD	Median	Credible Interval 2.5%ile	97.5%ile
Judge Effects β_i					
Hamada	0.311	1.507	0.303	−2.624	3.243
Logist	0.860	1.509	0.855	−2.039	3.846
Roth	1.152	1.524	1.151	−1.874	4.168
AP	−1.002	1.502	−0.967	−3.987	1.870
USA Today	−0.885	1.526	−0.874	−3.921	2.028
LV Review-Journal	−2.014	1.560	−1.989	−5.140	1.002
LA Times-Springer	−0.854	1.502	−0.847	−3.814	2.104
LA Times-Kawakami	−0.885	1.538	−0.894	−3.870	2.127
Sportsticker	−1.954	1.514	−1.910	−4.988	0.914
van de Wiele	1.207	1.505	1.196	−1.712	4.131

TABLE 6: Bayesian logistic regression results for the de la Hoya-Trinidad fight.

	Mean Bias	Std Dev	$P(b_j > 0)$
Official Judges			
Hamada	0.06	0.14	0.674
Logist	0.10	0.16	0.745
Roth	0.13	0.13	0.845
Media & Official Scorecards			
HBO-Lederman	−0.07	0.14	0.314
AP	0.06	0.15	0.652
USA Today	−0.04	0.13	0.364
Las Vegas Review-Journal	−0.16	0.14	0.125
Los Angeles Times-Springer	−0.04	0.13	0.382
Los Angeles Times-Kawakami	−0.04	0.13	0.369
Sportsticker	−0.15	0.15	0.159
van de Wiele	0.14	0.14	0.853

TABLE 7: Summary of Fitted Biases for Judges of De la Hoya-Trinidad

	Mean Bias	Std Dev	$P(b_j > 0)$
Official Judges			
Hamada	0.06	0.15	0.675
Logist	0.06	0.15	0.659
Roth	0.14	0.13	0.851
Media & Official Scorecards			
HBO-Lederman	−0.08	0.15	0.303
AP	0.06	0.16	0.649
USA Today	−0.05	0.14	0.350
Las Vegas Review-Journal	−0.17	0.15	0.121
Los Angeles Times-Springer	−0.05	0.13	0.352
Los Angeles Times-Kawakami	−0.05	0.13	0.358
Sportsticker	−0.16	0.16	0.156
van de Wiele	0.15	0.14	0.853

TABLE 8: Summary of Fitted Biases for Judges of De la Hoya-Trinidad

We perform the analysis again, handling official judge Logist's even round directly; the results are summarized in Table 8. Not surprisingly, the only judge whose results are significantly altered by accounting for ties is Logist; his fitted bias drops toward zero, and his probability of bias for Trinidad moves toward one-half. All scorecards are still reasonably consistent with a close fight and with each other. The estimated probability of Trinidad having won at least six rounds is 0.377, again not far from one-half.

6 Conclusions

In summary, our analysis and battery of tests examines the variability of scoring in two of 1999's most high-profile—and controversial—fights. The data do suggest that two of the official judges in the first Holyfield-Lewis fight—Larry O'Connell and, particularly, Eugenia Williams—scored the fight in a significantly different manner than other professional observers. The exact tests suggest that the Williams card is extreme relative to the traditional $\alpha = .05$ cutoff, and her pro-Holyfield card stands out as extreme in the other analyses we have performed. It is worth noting, however, that one of the unofficial judges—ESPN's Al Bernstein—produced a scorecard whose estimated pro-Lewis bias was larger in magnitude than Williams' pro-Holyfield bias. Analysis of scorecards from de la Hoya-Trinidad suggests a picture entirely consistent with random variation in scoring of a very close fight. The p-values for Roth and Logist in the exact test, .21, are well larger than .05. The other methods used reveal that none of the judges have a clear bias, with credible intervals for the judge effects comfortably containing zero. Despite the small sample size, we were able to employ a number of different methods that all reached the same conclusions. This demonstrates both the validity of the methods and the practicality of analyzing data sets of this size.

Let us be clear: by our analysis, we do not argue that Williams' scoring was corrupt and we do not challenge the competence of any of the judges (official or unofficial) whose scorecards are considered here. Indeed, by the simple act of committing to paper (or, at least, Web page) round-by-round scores rather than aggregate impressions, the unofficial judges mentioned here contribute to an informed discussion of fight outcomes and, for that, should be commended.

Rather, our primary finding—and the message we hope is borne out by this analysis—is a simple truth that may seem trivial to statisticians but is often lost on many boxing observers: there is variation in the scoring of boxing rounds. It is possible for different people, viewing a fight from different angles and weighting (already vague) scoring criteria in slightly different ways, to watch the same round of boxing and arrive at different conclusions; such is the nature of close rounds in boxing, and not necessarily the result of corruption. During the fight with Felix Trinidad, Oscar de la Hoya and his cornermen demonstrated the folly of not recognizing that scores can vary; de la Hoya's tactic of avoiding Trinidad in the late rounds seemed predicated on the assumption that he had won the first eight rounds with probability at or near one. But this was far from the case; only one of these early rounds, round three, was given to de la Hoya by all the scorers in our dataset.

As fans of the sport, we are not so naive as to believe that bribery and other impropriety are completely foreign to boxing and that fights have not been fixed; boxing's rich history is populated by both angels and scoundrels. Our analysis does clearly suggest that O'Connell's and Williams' scorecards in Holyfield-Lewis I were extreme observations; exactly why that is the case is not something statistics can explain. But, in boxing, there is far too often a rush to equate disagreement over a decision with the contention that one or more judges was "crooked", "corrupt", "incompetent", or "on the take". What is frequently forgotten by such critics is that a judge's scorecard is effectively the sum over twelve binary variables where, frequently, there is no clear and undisputed winner in a particular round. In a close fight—such as de la Hoya-Trinidad—there is enough variability that it is entirely reasonable to have a total which favors either fighter.

In light thereof, this research speaks to the importance of continuing education on the boxing scoring system. Many fan criticisms of supposedly "bad decisions" would likely be averted if the fan actually scored the bout using the 10 point must system, committing scores to paper as the fight proceeds (and gaining appreciation for how difficult it can be to judge close rounds). Our finding that professional judges scored de la Hoya-Trinidad differently from the set of unofficial judges suggests that broadcasters and fight reporters—like official judges—may benefit from formal training and seminars in boxing scoring, to make sure that all parties are clear on the criteria that should be applied in scoring a round of boxing. Fundamentally, work still needs to be done in the boxing community to achieve consensus on those scoring criteria—whether the existing four criteria are meaningful in practice, and how they should be weighted.

Boxing is a sport that celebrates and cherishes absolutes—decisive knockouts and champions who successfully defend their titles against any challenge. Yet it can be argued that boxing is most thrilling and majestic when two skilled, equally-matched fighters are pitted against each other, where both fighters perform to the peak of their abilities, and when all participants—audience and fighters alike—hang on the words of the ring announcer to learn who won the fight. Ironically, such moments are also times when the sport is highly vulnerable to criticism, as a hard-fought battle can quickly be dubbed a "bad decision". Scoring discrepancies will always occur by human nature—people will always see things in slightly different ways—but continuing to leap from disagreement with a fight outcome to accusations of another "bad decision"—without beginning to recognize the inherent variability in boxing scoring—would be the worst decision of all.

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement*, 20, 37–46.

ESPN2 (2000). *ESPN2: Friday Night Fights* (Television broadcast, January 14, 2000).

Feour, R. (2000). "Showtime Announcers Off Base in Scoring Criticism." *Las Vegas Review-Journal* (August 20).

Fernandez, B. (1999). "Trinidad's win matter of opinion." *Philadelphia Daily News* (September 20).

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Hauser, T. (1999). "How Reliable are 'Punch Stats'?" Boxing column, HBO, `http://www.hbo.com/boxing/columns/hauser/cmp/19990402.shtml`.

Henkelman, R. M., Kay, I., and Bronskill, M. J. (1990). "Receiver Operator Characteristic (ROC) Analysis without Truth." *Medical Decision Making*, 10, 24–29.

Izenberg, J. (1999). "De La Hoya Done In by Fatigue, Judges...and Trinidad." *New Jersey Star-Ledger* (September 20).

Johnson, V. E. (1996). "On Bayesian Analysis of Multirater Ordinal Data." *Journal of the American Statistical Association*, 91, 42–51.

Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*. New York: Springer.

Juipe, D. (2000). "Promoter Arum Hit With $125K Sanction." *Las Vegas Sun* (August 16).

Kaczmarek, T. (1996). *You Be the Boxing Judge: Judging Professional Boxing for the TV Boxing Fan*. Pittsburgh: Dorrance.

Kriegel, M. (1999). "Bad Advice Puts Oscar in Corner." *New York Daily News* (September 20).

Matthews, W. (1994). "Chavez's Great 'No Mas' Theft." *Boxing Illustrated*, 21.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.

Newfield, J. (1995). *Only In America: The Life and Crimes of Don King*. New York: William Morrow.

Pay-Per-View, T. (1994). *The Will to Win, the Power to Destroy: Holyfield vs. Moorer* (Television broadcast, April 22, 1994).

Povich, S. (1994). "Joe Walcott: Many Names, Endurance Most of All." *Washington Post* (March 6), D3.

Quinn, M. F. (1989). "Relation of Observer Agreement to Accuracy According to a Two-receiver Signal Detection Model of Diagnosis." *Medical Decision Making*, 9, 196–206.

SET Pay-Per-View (1999). *Warriors: Tapia vs. Ayala* (Television broadcast, June 26, 1999).

Uebersax, J. and Grove, W. M. (1993). "A Latent Trait Finite Mixture Model for the Analysis of Rating Agreement." *Biometrics*, 49, 823–835.

Uncredited (1999). "Reviled Judge Keen to Do Rematch: Briton Admits He Was Surprised His Card Showed Holyfield-Lewis Draw." *Ottawa Citizen* (March 17), B7.