A Statistical Analysis of the Fairness of Alternative Handicapping Systems in Ten-Pin Bowling

Sarah Keogh and Donal O'Neill*

August 30, 2012

Abstract

Using data on approximately 1040 games of bowling we examine the fairness of alternative handicapping systems in ten-pin bowling. The objective of a handicap system is to allow less skilled bowlers to compete against more skilled opponents on a level playing field. We show that the current systems used in many leagues do not achieve this objective and we propose a new optimal system which equalizes the playing field across all potential match-ups.

*Donal O'Neill is a Professor of Economics at the National University of Maynooth, Ireland (donal.oneill@nuim.ie); this work was carried out while Sarah Keogh was a graduate student at the National University of Ireland, Maynooth (sarah_kgh@yahoo.ie). We are grateful to Mark Hinchcliffe for providing us with the data used in this analysis and to Paddy Gibbons, Paddy Kilduff, Mark Delany, Olive Sweetman, two anoynmous referees, the editor and seminar participants at NUI Maynooth for helpful comments on the paper.

1 Introduction

Many sports adopt a handicap scoring system in determining the competition winner. In sports such as golf, polo and bowling, players compete under similar conditions and their recorded scores are then adjusted up or down once the game has been completed. In other sports, such as chess and horse racing, handicapping involves altering the conditions under which the game is carried out. Despite differences in the nature of handicapping, the stated objective of adopting such a system is to level the playing field so that weaker players find it worthwhile to take part in tournaments.

Formally the US Bowling Congress (USBC) defines a handicap league as follows

> "A handicap league is one in which a handicap is added to a bowler's score to place bowlers and teams with varying degrees of skill on as equitable a basis as possible for scheduled competition." (USBC 2011)

The Council of National Golf unions, the association charged with determining the handicap scores for golfers in Britain and Ireland, are more explicit in stating that

> "A golf handicap allows players of all levels of golfing ability to compete against each other on a fair and equal basis." (National Congress of Golfing Unions)

Use of the term equal implies that the purpose of these handicap systems is to equalize the probability of victory across ability levels. It is this feature of the handicap system that we examine in the context of ten-pin bowling.

There have been some studies of handicap systems in other sports, for example golf (Lackritz 2011; McHale 2010; Bingham and Swartz 2000). However there has been very little empirical analysis of bowling handicaps. Chen and Swartz (1994) carried out a statistical analysis of 5-pin bowling, a variant of bowling that is played only in Canada. However, the scoring and equipment in this version of the game differs significantly from the more popular ten-pin bowling version. In a very recent paper McCarthy (2011) conducts a statistical analysis of ten-pin bowling. However his analysis is restricted to elite bowlers on the Professional Bowlers Association National Tour, where handicaps are not applied. In this paper we use data from a recreational bowling league to examine the fairness of alternative handicap systems. Section 2 briefly describes the game of bowling. Section 3 discusses our data and section 4 provides a statistical analysis of these bowling scores Section 5 evaluates the fairness of alternative handicap systems and proposes an "optimal" alternative to the systems currently in operation.

2 Ten-Pin bowling

Ten-pin bowling is a competitive sport in which a bowler rolls a bowling ball down a lane with the aim of knocking down as many of the 10 pins as

possible. A game consists of 10 frames, with frames $1-9$ consisting of a maximum of two deliveries and frame 10 a maximum of 3 deliveries. If a bowler knocks down all ten pins using two balls in a frame, then a spare is awarded. If all ten pins are knocked down with the first ball a strike is awarded. A spare (strike) is worth 10 points, plus the points received from the next one (two) ball(s) rolled. A bowler who bowls a spare (strike) in the 10^{th} frame is awarded one (two) extra ball(s), which allow the awarding of bonus points. A bowler's score is the sum of points over the 10 frames.

3 Data Set

Our data are obtained from an adult bowling league in Dublin, Ireland. The league consisted of 10 teams, each team containing two players plus a possible replacement. The teams competed against each other over 18 weeks, with each team member scheduled to play 3 games on any given night. Over the duration of the league we obtained individual bowling scores for 23 bowlers and 1036 games. Summary statistics are given in Table 1.

Many of the procedures used in our analysis depend on the randomness of the underlying samples. To examine the possibility that a player's current score is affected by their scores in earlier games we compute the autocorrelations in scores across games for each individual at varying time lags. The estimated autocorrelations provide very little evidence of dependence across games, with only 3 out of the 138 estimated autocorrelations being statisti-

cally significantly different from zero at the 5% significance level. To test if the score of the opposing team had an impact on your team's performance we estimated the correlation between a team's own score and their opponent's score. In only one case can we reject the null-hypothesis at the 5% significance level, while a combined 5% significance test fails to reject the null-hypothesis of no correlation across all teams. The lack of correlation in individual scores over time and between own scores and an opponent's score support the assumptions of randomness and independence adopted in the remainder of our paper.

4 Analysis

In order to evaluate the fairness of alternative handicap systems we first establish the distribution of bowling scores across games. Since the final score in a game of bowling is the sum of the scores obtained over a number of frames, it seems reasonable to consider the normal distribution as an initial approximation to the distribution of raw bowling scores.

4.1 Raw Scores

We test for normality of raw bowling scores using two formal tests; the Shapiro-Francia (SF) test (Shapiro and Francia 1972; Royston 1983) and the Skewness-Kurtosis (SK) test (D'Agostino et al. 1990; Royston 1991). To test the overall hypothesis that the scores are normally distributed across all

players we combine the 23 individual tests using two combination methods; Fisher's method (Fisher 1970) and Liptak's method (Won et al. 2009; Chen and Nadarajah 2011). Both the SF and SK tests reject the null hypothesis that the distribution of scores is normally distributed for all players at the 1% significance level using either combination method.

4.2 Transformed Scores

Having rejected normality of the raw scores we consider whether there exists a transformation of the raw scores that is well represented by a normal distribution. In particular, we consider Box-Cox transformations of the raw scores (Box and Cox 1964; Spitzer 1980). We obtain a maximum likelihood estimate of the Box-Cox parameter, λ, equal to -0.148 with a standard error of 0.199. Since we cannot reject $\lambda = 0$ (the log transformation) at a 5% significance level we adopt this transformation throughout the remainder of the paper.

Again we test for normality using both the SF and the SK test. With the SF test individual normality of the transformed scores is rejected for only two of the 23 players at the 5% significance level. Furthermore the combined tests fail to reject the null-hypothesis of normality across all players at conventional significance levels; the p-values are .068 and .14 for Fisher's and Liptak's method respectively. For the SK test individual normality is also rejected for only two of the players at the 5% significance level, yet, somewhat surprisingly the combined tests reject normality across all players

at the 5% significance level. However this latter result is largely driven by player 10. When this one player is omitted from the analysis the p-value for Fisher's combined test increases from .017 to .0481.

When looking at the outcome of games between players of different ability levels it will simplify calculations if we can assume that the variance of transformed bowling scores does not depends on a player's ability. To test this we ran a linear regression of the variance of transformed scores against average scores. The slope coefficient from this regression was -0.002 with a standard error of 0.012. Thus for the remainder of this paper we assume a constant variance of transformed bowling scores across players.

5 Handicapping

In bowling leagues the range of abilities can be wide. As a result leagues often adopt handicap systems to make tournaments more competitive. At present there are a variety of systems used in Ireland. Under a handicap system the final score of a player with average ability m_i (determined over previous games played by the player) is adjusted by adding a bonus to their raw score. The bonus is given by $w(P - m_i)$, where w is a weighting factor and P is a scratch score. The most common weighting factor is .8, though weighting factors as low as .66 are used. Although the scratch score used to calculate the handicap affects the number of pins added to a player's score, the choice of scratch score will not affect the probability of victory provided

the scores of both players are adjusted under the handicap. Typical scratch scores are 220 or 200.

Clearly the higher the weighting factor the more generous the system is. Under a system with $w = .8$ and $P = 220$ a player with an average of 120 would be credited with an extra 80 pins. Reducing w to .66 would result in this player receiving only an additional 66 pins. With $w = .8$ and $P = 220$ a bowler with an average of 160 would have to beat a lower ability opponent, with an average say of 120, by more than 32 points in order to win the game. Increasing the weighting factor to $w = 1$ would increase the required margin of victory to 40 points.

More generally consider a game between two players with ability levels m_x and m_y, respectively. For exposition we assume $m_y < m_x$ and refer to player y as the underdog and player x as the favorite. To establish the probability that the favorite wins, we need to determine $Pr(X > Y + h)$, where $h = w(m_x - m_y)$. Letting $f(x, y)$ denote the joint distribution of X and Y, we can write this probability as :

$$\Pr(X > Y + w(m_x - m_y)) = \int_{w(m_x - m_y)}^{\infty} \int_{0}^{x - w(m_x - m_y)} f(x, y) dy dx. \qquad (1)$$

If we further assume that X and Y are independent, together with $Z = \ln(x) \sim N(u_x, \sigma^2), W = \ln(y) \sim N(u_y, \sigma^2)$, where $u_j = \ln(m_j) - \frac{\sigma^2}{2}$, $j = x, y$,

we can derive $f(x, y)$ as :

$$\frac{1}{2\sigma^2\pi}\exp\left[-\frac{\left(\ln(x)-\ln\left(m_x\right)+\frac{\sigma^2}{2}\right)^2+\left(\ln(y)-\ln\left(m_y\right)+\frac{\sigma^2}{2}\right)^2}{2\sigma^2}\right]\frac{1}{xy}. \quad (2)$$

The required probability can then determined using either numerical methods or Monte-Carlo sampling techniques. In our calculations we use the pooled variance across all players, $\hat{\sigma}_p^2 = .01957$, as our estimate of σ^2. We choose seven values of m_i, ranging from 100 to 220 in intervals of 20, and calculate the probability of the favorite winning for each of the 21 unique ability match ups.

5.1 Evaluating the Current System

The first three entries in Table 2 give the probabilities of the favorite winning across each match-up using the raw unadjusted scores (first entry in each cell), adjusted scores using $w = .66$ (2nd entry in each cell) and adjusted scores using $w = .8$ (3rd entry in each cell). Looking first at the probabilities based on raw scores we see that, even with modest differences in ability, the probability of the favorite winning is consistently over .75 and can quickly rise to above .90. We also see that the probability of victory depends on the ability levels of the players involved, and not just the difference in their abilities. For instance the probability of a player with ability level 160 defeating a player with ability level 120 is .93, however it is only

.85 when the players' abilities are 220 and 180. This non-linearity follows from the log normality of the underlying data and as a result one must be careful when extrapolating findings based on one part of the distribution (for example those based on scores by professional bowlers) to other parts of the distribution (for example the performance of players in recreational leagues). With log-normality the outcome probability depends on the ratio of ability levels rather than the difference. Therefore the probability of a player with ability level 160 defeating a player with ability level 120, is equal to the probability of a player with ability level 220 defeating a player with ability 165.

The strong tendency of high ability bowlers to win when raw scores are used motivates the use of handicap systems. The 2nd and 3rd entries in each cell give the probability of the favorite winning using a handicap system with $w = .66$ and $w = .8$ respectively. As expected the adjustments reduce the favorite's advantage. However the playing field remains far from level. The probability of a player with ability level 180 beating a player with ability level 100 is .83 when $w = .66$ and .70 when $w = .8$. Discussions with league organizers suggest that this has been a feature of Irish bowling leagues in practice, with the same players and teams filling the top places every year. In the next section we therefore consider what adjustments to the handicap system, if any, can produce fairer outcomes. However, before we do that we first examine the fit of our theoretical predictions to the outcomes observed in practice.

Our estimated probabilities are based on our assumptions of independence, a constant variance and log normality. To assess the validity of these assumptions we compare the theoretical probabilities to the actual outcomes in the bowling league under analysis. To do this we use the fact that in addition to the individual scores bowled each night we also know the average score of each player in the 20 games bowled prior to the commencement of the league. It is this average, along with the scratch score and weighting factor, that determine the score adjustment under a given handicap system. At the start of the league we know for instance that player 7 had an average score of 220 and player 15 had an average score of 160. Since this was a team league these players never competed directly against each other. Nevertheless both players bowled in all 54 games, with their games taking place on the same evenings, at the same time and in the same bowling alley. One way to check the accuracy of our predictions is to compare the scores recorded by both of these players in games played at the same time *as if they were competing against each other.* We can then determine the victor in each of these 54 games and compare the observed probabilities with our theoretical predictions. Using the raw scores we find that the high ability player would have emerged victorious in 53 of the 54 games. This gives an estimated probability of victory for the high ability player equal to .98, which is close to our predicted estimate of .95. We then repeat the analysis this time adjusting scores using a scratch score of 220 and a weighting factor of .8 (similar to what is actually used in the league). With these adjusted scores the number

of games won by the favorite falls to 35, giving a probability of victory equal to .648. This observed probability is once again very close to our theoretical probability of .621. Our estimated probabilities therefore seem to fit the observed probabilities very well.

5.2 Improving the current System

We now examine what changes, if any, to the handicap system would produce more equitable outcomes. As noted earlier the key feature of a handicap system in determining the probability of victory is the weighting factor. To examine which weighting system produces the most equitable outcomes we consider a scaled goodness of fit statistic based on the 21 comparisons in Table 2 :

$$\chi^2(w) = \sum_{k=1}^{21} \frac{(P_k(w) - E_k)^2}{E_K}, \tag{3}$$

where $P_k(w)$ is the derived probability of the favorite winning in match-up k and E_K is the probability of success on a level playing field. Since we are considering head to head games $E_K = .5$. The goodness of fit statistic therefore measures the deviation of the observed distribution from a Bernoulli distribution with equal probabilities and depends on the handicap system through its dependence on w. By choosing w to minimize χ^2 it is possible to determine the handicap system that produces the *fairest* outcomes across all potential match-ups. Figure 1 plots the value of $\chi^2(w)$ against w. From this we see that the handicap system that produces the fairest outcomes uses a

weighting factor $w = .98$. The fourth entry in each cell of Table 2 gives the probabilities of the favorite winning across each possible match-up using the optimal weight. The probability of the favorite winning is almost exactly .5 in all 21 unique match-ups considered. Thus despite the nonlinearities evident in the distribution of bowling scores it is possible to achieve fair outcomes across all games with this simple linear handicap system.

To compare the distribution of victories generated by alternative handicap systems to a Bernoulli distribution with parameter $p = .5$ we use the test statistic $X^2(w) = 2n \sum_{k=1}^{21} \frac{(P_k(w)-.5)^2}{.5}$, where n is the number of games used to estimate $P_k(w)$. This test statistic approximates to a chi-squared distribution with $df = 21$. To estimate $P_k(w)$ we simulated n games for each of the 21 pairings in our analysis using the distribution given in (2) for various values of w. With $n = 1000$ we obtain a value for $X^2(.8)$ equal to 1628 which is far in excess of the 5% critical value of 32.67. In contrast a similar exercise carried out using the optimal weight $w = .98$ provided a $X^2(.98)$ equal to 1.84. Further analysis revealed that we would reject the null-hypothesis of equality at the 5% significance level for any w less than .965 but fail to reject at this significance level for $w \in [.965, 1]$. It is interesting to note that this latter range includes the 100% weighting ($w = 1$). This is the weight recommended by the USBC (2011). The probabilities of the favorite winning using the USBC weight are given in the final entry in each cell of Table 2. As expected the USBC weight produces a more level playing field than the system currently used in many leagues ($w = .8$), though by definition

this weight cannot level the playing field to the same extent as our proposed optimal system. Moreover, under the USBC scheme the odds of victory shifts in favour of the less skilled player. Some authors (McHale 2010) argue that such a system distorts incentives. In particular league organizers are wary that such a scenario would encourage better players to "sandbag"; this is a situation where players bowl below their ability level for a period in order to obtain bigger handicaps for upcoming tournaments. An illustration of the problems that can arise from such behavior is provided by a January 2012 High Court lawsuit filed in Ireland. In the lawsuit the Golfing Union of Ireland (GUI) were sued for €10m by a member whose handicap was reduced following suspicion of "sandbagging." The player sued the GUI for defamation, claiming that the decision to adjust his handicap was tantamount to being branded a cheat. Although the lawsuit was unsuccessful the GUI face legal costs estimated at between €150,000 and €160,000. Our proposed optimal weighting scheme not only produces fairer outcomes than the USBC scheme but maintains a small incentive for players to improve skills levels.

Finally the framework we have developed allows league organizers to judge the trade-off between fairness and incentives when choosing an appropriate handicap system. Some league organizers might be willing to sacrifice some fairness in return for greater incentives for skill-development. In this instance organizers might consider a weighting factor slightly less than the optimal. For example with a weighting factor $w = .95$ the estimated probabilities of the favorite winning range from .50 to .55 across the match-ups we consider.

Although this weighting scheme falls a little short of the optimal level of fairness, organizers may nevertheless be willing to except this small loss in fairness in return for the greater incentives for skill-development that follow.

6 Conclusions

Handicap scoring systems are used in many bowling leagues with the explicit purpose of allowing all players to compete on a level playing field. This paper examines the distribution of ten-pin bowling scores and uses the findings to examine the fairness of alternative handicapped scoring systems. We show that many of the current systems still leave large biases in favour of high ability players. Using a 100% weighting factor produces a more level playing field but goes too far in the sense that players have no incentive to improve their skill level. We derive an optimal weighting factor equal to .98 and show that this system allows lower ability players to compete across all match-ups yet still provides some incentives for players to improve their game.

References

Bingham, D. and T. Swartz (2000), "Equitable Handicapping in Golf," *The American Statistician*, 54, 170-177.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, 26, 211-252.

Chen, Z and S. Nadarajah (2011) "Comments on 'Choosing an Optimal Method to Combine p-values,' by Sungho Won, Nathan Morris, Qing Lu and Robert C. Elston, Statistics in Medicine 2009; 28:1537–1553," *Statistics in Medicine*, 30, 2959–2961.

Chen, W. And Swartz, T. (1994), "Quantitative Aspects of Five-Pin Bowling," *The American Statistician*, 48, 92-98.

D'Agostino, R. B., Belanger, A. and D'Agostino Jr. R. B. (1990), "A Suggestion for Using Powerful and Informative Tests of Normality," *The American Statistician*, 44, 316-321.

Fisher, R.A. (1970), *Statistical Methods for Research Workers*, 14th edition Hafner Publishing Company, New York.

Lackritz, J (2011), "A New Analysis of the Golf Handicap System: Does the Better Player Have an Advantage," *Chance*, 24 (2), 24-34.

McCarthy, D (2011), "Estimating Different Sources of Variation and Predicting Tournament Outcomes in Professional Bowling," *Chance*, 24 (3), 37-46.

McHale, I (2010), "Assessing the Fairness of the Golf Handicap System in the UK," *Journal of Sports Sciences*, 28, 1033-1041.

Royston, J. P. (1983), "A Simple Method for Evaluating the Shapiro-Francia W' Test of Non-Normality," *The Statistician*, 32, 297-300.

Royston, J. P. (1991), "Comment on sg3.4 and an Improved D'Agostino Test," *Stata Technical Bulletin*, 3, 13-24.

Shapiro, S. S. And Francia, R. S. (1972), "An Approximate Analysis of Variance Test for Normality," *The Journal of the American Statistical Association*, 67, 215-216.

Spitzer, J. J. (1982), "A Primer on Box-Cox Estimation," *The Review of Economics and Statistics*, 64 (2), 307-313.

United States Bowling Congress (2011)

http://usbcongress.http.internapcdn.net/usbcongress/bowl/rules/pdfs/2011-2012PlayingRules.pdf

Won, S., N. Morris, Q. Liu and R. Elston (2009) "Choosing an Optimal Method to Combine P-Values," *Statistics in Medicine*, 28, 1537–1553

Table 1. Summary of raw bowling scores.

Player	Games Bowled	Average	Standard Deviation	Player	Games Bowled	Average	Standard Deviation
1	51	170	24	13	45	156	21
2	51	199	27	14	33	216	32
3	38	183	22	15	54	161	17
4	41	155	27	16	54	169	25
5	34	192	28	17	54	181	28
6	10	179	30	18	54	154	21
7	54	220	25	19	54	208	29
8	54	194	24	20	54	210	27
9	54	189	28	21	33	215	32
10	54	194	22	22	23	208	38
11	54	177	28	23	29	205	30
12	54	176	29				

Table 2. The simulated probabilities of the favorite winning in head to head bowling matches between players of different ability levels. The probabilities are estimated using equations (2) and (3) provided in the text with $\hat{\sigma}_p^2 = .01957$. The results are reported for five handicap systems; from top to bottom these correspond to $w = 0$; $w = ..66$; $w = .8$; $w = .98$ (optimal weight); $w = 1$

Ability	120	140	160	180	200	220
	.82	.96	.99	1.00	1.00	1.00
	.62	.71	.78	.83	.87	.89
100	.57	.62	.67	.70	.73	.76
	.50	.50	.51	.51	.51	.51
	.49	.49	.49	.48	.48	.48
		.78	.93	.98	1.00	1.00
		.60	.68	.75	.80	.84
120		.56	.61	.65	.68	.71
		.50	.50	.50	.51	.51
		.50	.49	.49	.49	.48
			.75	.90	.96	.99
			.59	.66	.72	.77
140			.55	.59	.63	.66
			.50	.50	.50	.51
			.50	.49	.49	.49
				.72	.87	.95
				.58	.65	.70
160				.55	.58	.62
				.50	.50	.50
				.50	.49	.49
					.70	.85
					.57	.63
180					.54	.58
					.50	.50
					.50	.49
						.69
						.56
200						.54
						.50
						.50

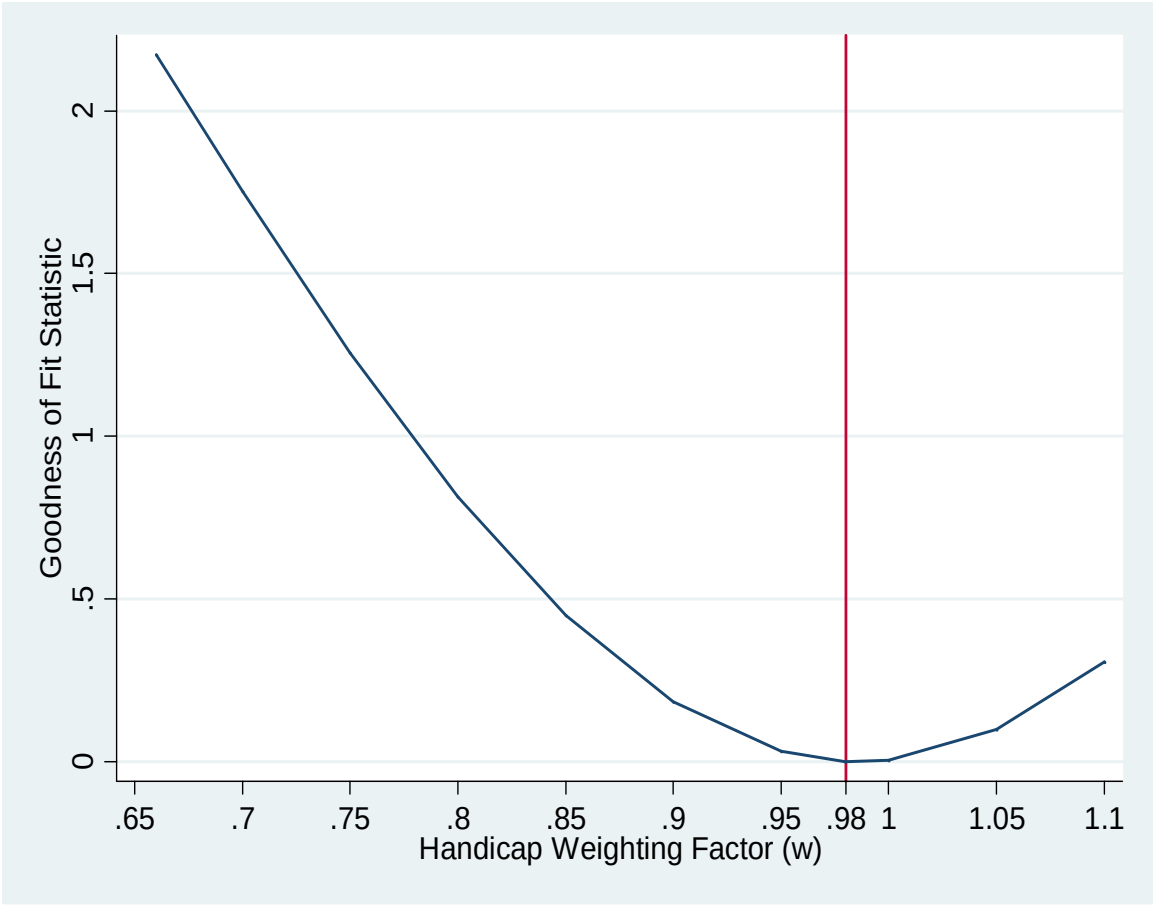

Figure 1: Measure of fairness across alternative handicap systems