DEPARTMENT OF ECONOMICS AND FINANCE

COLLEGE OF BUSINESS AND ECONOMICS

UNIVERSITY OF CANTERBURY

CHRISTCHURCH, NEW ZEALAND

A Method for Inferring Batting Conditions in ODI Cricket from Historical Data

Scott Brooker and Seamus Hogan

WORKING PAPER

No. 44/2011

Department of Economics and Finance
College of Business and Economics
University of Canterbury
Private Bag 4800, Christchurch
New Zealand

A Method for Inferring Batting Conditions in ODI Cricket from Historical Data*

Scott Brooker[1] and Seamus Hogan[2]

[1] *Yellow*
srbrooker@gmail.com

[2] Corresponding Author
Department of Economics and Finance
University of Canterbury
seamus.hogan@canterbury.ac.nz
(03) 364-2524

Abstract:

This paper is part of a wider research programme using a dynamic-programming approach to modelling the choices about the amount of risk to take by batting and bowling teams in One Day International cricket. An important confounding variable in this analysis is the ground conditions (size of ground, nature of pitch and weather conditions) that affect how many runs can be scored for a given amount of risk. This variable does not exist in our historical data set and would regardless be very difficult to accurately observe on the day of a match.

In this paper, we consider a way of estimating a distribution for the ground conditions using only the information contained in the first-innings score and the result of the match. The approach uses this information to estimate the importance of ground conditions in the determination of first innings total scores. We assume a functional form for a model of first innings scores and we estimate the parameters of our model using Monte Carlo methods. We test the impact of a significant rule change and we apply our findings to selected matches before and after the new rules came into play.

Keywords: ODI Cricket, Batting Conditions, Nuisance Variable

Acknowledgements:

We would like to thank the late Sir Clive Granger, Liam Lenten, Pete Mayell, Dorian Owen, and Les Oxley for helpful comments on earlier drafts of this paper, and *New Zealand Cricket,* for providing much of the data used in this paper.

This paper is based on material contained in Brooker's Ph.D. Thesis (Brooker, 2011). Brooker would like to thank *Sport and Recreation New Zealand* (SPARC), *The Tertiary Education Commission* (TEC), and the College of Business and Economics Research Committee at the University of Canterbury for financial support that made this research possible.

A Method for Inferring Batting Conditions in ODI Cricket from Historical Data

1. Introduction.

The outcomes that take place on a sports field are obviously heavily influenced by the ability and performance on the day of the athletes taking part; however, these are not the sole determinants. In many sports outcomes are also influenced by random influences, ranging from human error by match officials to the proverbial "rub of the green". For empirical researchers interested in analysing sports data, most of these external influences can simply be modelled as exogenous sampling error. There is one influence, however, that is potentially less benign—the impact of weather and venue conditions at the time of the sporting event. In many sports, particularly those played outside, the ease with which player skill and effort can translate into positive outcomes can depend heavily on these conditions. If the variation in conditions during the course of a match is small relative to the variation in conditions between different matches, then conditions within a match cannot reasonably be modelled as independent draws from some random distribution.

One sport where this issue can be particularly problematic is one-day-international (ODI) cricket. In ODI cricket one team bats and has a single "innings" in which it seeks to score as many runs as possible. The innings ends when the other team has bowled 300 deliveries to the batsmen, or when ten batsmen have been dismissed, whichever comes first. The teams then change roles and the other team has an innings of 300 deliveries or 10 dismissals with which to try and achieve a higher score.

ODI cricket has been the subject of a lot of empirical research in the academic literature of statistics, operations research and economics, in part because of enthusiasm for the game of researchers in those areas, but also because of its highly quantitative nature, with the state of the game being quantifiable after each of the up-to 600 deliveries that constitute a match.

Statistical analysis of ODI cricket typically consists of estimates of distributions of likely outcomes as a function of the state of a game at a particular point. For instance, the Duckworth-Lewis system currently used in all ODI matches to make adjustments to target

scores when bad weather forces an interruption in a match with a consequent reduction in the time available for play, originated as an academic paper (Duckworth and Lewis, 1998) that used statistical analysis to model the likely additional runs scored in the remainder of an innings as a function of the balls already bowled and the number of wickets lost. Other papers in this tradition include Clarke (1988), Preston and Thomas (2000), and Carter and Guthrie (2004).

A limitation in all these analyses is the lack of information regarding the ease of batting conditions. As we explain in the next section, variation across matches in the ground at which a match is played and the weather conditions at the time of the match can have a large effect on how easy it is for teams to score runs when batting. In the absence of data concerning these conditions, empirical models, such as in the papers cited above, will find that the effect of playing in difficult conditions and the effect of playing badly will be confounded in the data, with subsequent limitations on the interpretation of the models.

This limitation has been recognised in the literature. In his seminal paper, Clark (1988) notes that estimates should take into account playing conditions. Duckworth and Lewis (2005) are critical of the proposed alternative to the Duckworth-Lewis target-adjustment method proposed by Carter and Guthrie (2004), stating that the Carter-Guthrie approach does not take ground conditions into consideration. The model proposed in Duckworth and Lewis (1998), however, implicitly assumes that all variation in first-innings scores is due to variation in ground conditions, when in truth the variance in scores comes from a combination of variation in ground conditions and variance performance on the day. Again, it is the absence of data on conditions that forces researchers to adopt either one of these extreme points of view of either ignoring variation in conditions or ascribing too much importance to it.

In this paper, we proposed a method by which information about ground conditions can be inferred in historical data. Our method doesn't provide a point estimate, but rather a distribution of possible values using information from the match to update prior information using Bayes' rule. We believe that the information revealed by this method, while still imperfect, can greatly improve existing empirical models of ODI cricket, particularly models of how to adjust the target score in the event that matches are shortened due to bad weather.

In the following section, we give a brief overview of the role that ground conditions can play in an ODI cricket match. In Section 3, we outline the theory for how a posterior distribution of values for ground conditions can be inferred from observable match data. In

Section 4 we describe the data used in this paper, which we then apply to the theoretical model to provide some general results in Section 5. Section 6, presents some diagnostic analysis of our results to demonstrate the usability of our ground conditions measures in other empirical work. Section 7 presents a discussion of possible extensions to the method.

2. The Role of Ground Conditions in ODI cricket.

In this paper, we assume that the reader has a basic understanding of the structure of a game of One Day International (ODI) cricket; for the uninitiated, we provide a brief description of the game in the Appendix.

There are five main factors that influence the first-innings score in an ODI match as well as the likelihood of each score being a winning one. These factors are

- the skill levels displayed by the players on both teams;
- luck;
- ground size;
- pitch conditions; and
- weather conditions.

The skill measure refers to both the ability and the execution on the day of the players, with high scores being likely when batters perform well relative to the performance of the bowlers and fielders.

Luck plays a role in the outcome of a match; for example, poor umpiring decisions can have a marked influence, as can uncontrolled aerial shots that fall safely rather than going directly to the fielder.

On a small ground, it is relatively easier for the batsmen to hit the ball out of the playing field for boundaries and for this reason scores tend to be higher on small grounds than on large grounds. A mitigating factor here is that there are generally fewer twos and threes run as batsmen more often have to settle for single runs due to the ability of the fielders to reach the ball faster on a smaller ground. A fielding side should, however, be at a minimum indifferent if they were given the option to change from a small ground to a larger ground, as the larger ground simply creates more options for possible field settings, as well as making it more difficult for the batsmen to hit boundaries.

Pitches are extremely variable in their nature. The moisture content, the type of soil used, the hardness, the amount of grass and any cracking present on the pitch all have an impact on how the ball behaves when it bounces on the pitch. Any movement or change of

direction of the ball after hitting the pitch makes batting more difficult, as does inconsistent bounce, extreme pace off the pitch and extreme lack of pace off the pitch. Pitches are very individual; therefore, it is not appropriate to assume that all pitches at a particular ground will behave in the same way.

A fascinating aspect of the game of cricket is the tendency of the ball to "swing", or change direction, in the air after it has been bowled. This swing, if present, makes batting significantly more difficult and is likely to lead to lower scores. On a cloudy or humid day the ball generally swings significantly more than on sunny dry days. For this reason the weather is our final factor influencing the outcome of the game.

It is useful to categorise these factors into two groups, based on the degree to which they are the same for both teams on any given day. The skill level is clearly team-specific and luck should be completely random; therefore, we combine these factors into a category entitled "performance". The size of the ground obviously does not change during the game, and while pitch and weather conditions might change somewhat over the course of a match, we assume that these factors vary to a far lesser degree within a match than between separate matches. We assign these three factors to a category entitled "conditions".

The aim of this paper is to estimate what the average score would have been on the pitch used for a particular match. That is, we ask the question: If a team with average batting ability were to play a large number of games against a team with average bowling and fielding ability in the same conditions, what would the average score be. This average is the theoretical value for "conditions" for the match, and deviations in the first-innings score from that average can be attributed to the various factors that we include in our variable termed "performance".

In the next section, we describe the identification strategy we use to infer a value for conditions in each match, using information from the match itself—specifically, the first-innings score and the result of the match. Before describing this approach, it is worth briefly discussing why we seek to infer conditions only from match data rather than external information about ground conditions. In particular, it has been suggested to us, that, because the size of any particular ground changes rarely if at all over time, and the type of soil at the ground remains consistent over time, that one could infer a lot about average ground conditions simply by regressing first-innings score on a set of dummy variables for the ground where the match is played. There are two reasons why this would not be a useful approach.

First, there are a surprisingly large number of games on which ODI cricket matches have been played over the period of our dataset (the decade from 2000-2009), with many ground hosting a single match.

Second, even with a constant size and soil type, there can be a lot of variation over time in how easy it is to bat on a particular ground, partly from variation in the weather on the day of the match, but also because weather conditions in the lead up to the match will typically affect how much the ball will change direction after hitting the pitch. Furthermore, this relationship between weather and conditions is not highly predictable, so that even if historical data on weather conditions before and during a match were available, it would be impossible to quantify it into a stable relationship with batting conditions.

Accordingly, in this paper we adopt the strategy of assuming that there is no useful information from knowing the ground at which the game was played. In the final section, however, we discuss how ground information could be combined with our measures in further research.

3. Outline of our Approach:

In this section, we provide an extremely stylised model of a game of ODI cricket to illustrate our basic approach. Throughout this paper, we refer to the team batting first as "Team 1", and the team batting second as "Team 2".

We model a game of ODI cricket as follows: Initially, the ease of batting conditions, χ, which is common across the match is drawn from a distribution, F, with density, f. Team 1 then draws a value of its performance, ρ_1, from a distribution, G_1, (density, g_1) and conditions and performance are summed to give that team's score, S_1,

$$S_1 = \rho_1 + \chi.$$

Without loss of generality, assume that the mean of the performance distribution, G_1, is zero; that is, the interpretation of performance is how much better the team performs than the average performance one can expect given the conditions. We assume that S_1 is observable to the modeller, but the components, ρ_1 and χ, are not.

Team 2 then also draws a value of its performance, ρ_2, from a second-innings performance distribution, G_2 (density, g_2), which combines with the common conditions to give a second-innings outcome, S_2,

$S_2 = \rho_2 + \chi.$

Even though we assume conditions are the same for both innings, we don't require that the performance distributions, G_1 and G_2, be the same. This is due to the team batting second having a known target score, resulting in their being able to adjust their risk strategy depending on the target. The fielding team does have some control over the overall risk strategy of the innings, in terms of bowling style and field placement, but significantly more control over the risk strategy is available to the batting team. This is obvious to any cricket watcher as we almost always see the scoring rate increase and the survival rate decrease towards the end of the first-innings, which is what the team batting first would generally prefer as its overs begin to run out. The effect of this is that a team chasing a low target relative to conditions will choose to bat more conservatively than they would if they were unaware of the target score, and a team chasing a high target will bat more aggressively. Teams chasing an average target will typically adjust along the continuum between conservative and aggressive strategies as their innings progresses as a function of how well they are doing. We model these effects by assuming that the second-innings performance distribution is a uniform rightward shift from the first-innings performance distribution. That is,

$$G_2(\rho_2) = G_1(\rho_2 - \gamma). \tag{1}$$

This is essentially assuming that, whatever the target, Team 2 starts with γ runs already scored.

We make two assumptions about the distributions, F, G_1, and G_2.

First, we assume that χ, ρ_1, and ρ_2 are distributed independently on each other. This implies that the variation in scores due to performance on a pitch where scoring is difficult will be similar to the variation on a pitch where scoring is easy. While one might expect that the variance of performance would be proportionate to the level of conditions, our experience watching cricket suggests that this is not, in fact, the case.

Second, we assume that each of the three distributions is normal, which implies normality in S_1 and S_2. This seems a reasonable approximation an *a priori* grounds for the performance distributions, G_1 and G_2, because of the central limit theorem. The performance measure is a combined measure of batting team performance and fielding team performance. The batting team performance is composed of the individual performances of up to 11

batsmen and the fielding team performance is composed of the individual performances of up to 11 bowlers and fielders. Each player may not play an equal part in determining the overall performance of the teams, but generally speaking the central limit theorem would imply that there are more ways of putting together the 22 performances in a way that gives an average overall performance than there are ways of putting them together to get an extremely good or extremely poor performance. Furthermore, with 300 individual balls in an innings, performance will also vary from ball to ball even within the overall performance of an individual player. The most extreme performances would require an extremely good performance from all required members of one team and an extremely poor performance from all required members of the other team. This would be much less likely than an average total performance, which could be caused by almost unlimited combinations of good batting and bad bowling from various players, or vice versa, completely cancelling each other out. This is true even if the individual player batting and bowling performance distributions were uniform.

We can make a similar argument for the normality of the conditions distribution. Conditions are a combination of a number of individual factors such as the nature of the pitch, ground size and weather conditions. These main factors are likely to have smaller factors underpinning them, with each sub-factor requiring a draw from a distribution for each match. It is, however, not as obvious that our conditions distribution should have a normal distribution as it is for our performance distribution, due to at least some factors, such as rainfall and soil type, being relatively constant at a particular venue or at least correlated with a particular country. Later in the chapter we show that normality is a reasonable assumption for S_1 and S_2, which increases our confidence in the normality of χ.

Of course, it can't be literally true that the distribution of conditions and performance are normal, since negative scores are not possible. This is extremely unlikely over the range of the data, however.[1] The log-normal distribution, while having the desirable property of being bounded at zero, does not fit the data well.

[1] The probability, given the mean and variance of our full data set, of our assumed normal distribution generating a score in any given match less than zero is 0.000016. This means that over our dataset of 784 matches, the probability of all our observed scores being greater than zero is 98.8%.

Let ω denote the result of the match with $\omega = 1$ if Team 1 wins and $\omega = 0$ if it loses, so that $\omega = 1$ if $S_2 < S_1$, which by assumption is equivalent to $\rho_2 < \rho_1$.[2] We assume that the second-innings outcome is a non-observable latent variable, but the result of the match is observed. There are two reasons that we cannot simply use the second-innings score as an observable measure of the second-innings outcome. The first is that a match ends as soon as Team 2 has overtaken Team 1's score, so that instances where Team 2 heavily outperforms Team 1 do not show up in the data as a big difference in scores. The second reason is that the optimal adjustment in the level or risk taken by Team 2 when batting can result in a small difference in performance showing up as a very large difference in scores as they get forced by the game situation into taking highly risky strategies.

The information available to the modeller, then, is the first-innings score, S_1, and the result of the match, ω. The idea of this paper is to find a posterior density for conditions, f_p, conditional on these two pieces of information. Let H_χ and h_χ denote the distribution and density of S_1 conditional on a particular value of conditions, χ, and let $\Pr(\omega \mid S_1, \chi)$ denote the probability that Team 1 achieves the result, ω, given its score, S_1, and the match conditions, χ. From Bayes' rule we have

$$f_P(\chi \mid S_1, \omega) = \frac{f(\chi) h_\chi(S_1 \mid \chi) \Pr(\omega \mid S_1, \chi)}{\int h_\chi(S_1 \mid \chi') \Pr(\omega \mid S_1, \chi') dF(\chi')}. \tag{2}$$

Note that the density, h_χ, and the probability, $\Pr(\omega \mid S_1, \chi)$, can be inferred from the distributions of performance, G_1 and G_2:

$$h_\chi(S_1 \mid \chi) = g_1(S_1 - \chi), \tag{3}$$

$$\Pr(\omega = 1 \mid S_1, \chi) = G_2(S_1 - \chi) = G_1(S_1 - \chi - \gamma), \tag{4}$$

$$\Pr(\omega = 0 \mid S_1, \chi) = 1 - G_2(S_1 - \chi) = 1 - G_1(S_1 - \chi - \gamma), \tag{5}$$

so that Equation (2) can be written entirely in terms of the distributions, F and G_1:

$$f_P(\chi \mid S_1, 1) = \frac{f(\chi) g_1(S_1 - \chi) G_1(S_1 - \chi - \gamma)}{\int g(S_1 - \chi') G_1(S_1 - \chi' - \gamma) dF(\chi')}, \text{ and} \tag{6}$$

[2] Note that in this theoretical model with continuous distributions, the probability of a tie—i.e. of Team 2 scoring exactly the same number of runs as Team 1—is zero. In reality, ties are possible. We describe in Section 5 our way of dealing with the small number of ties in our database.

$$f_P(\chi \mid S_1, 0) = \frac{f(\chi) g_1(S_1 - \chi)\big(1 - G_1(S_1 - \chi - \gamma)\big)}{\int g_1(S_1 - \chi')\big(1 - G_1(S_1 - \chi' - \gamma)\big) dF(\chi')}. \quad (7)$$

2.2 Identifying F and G_1.

Equations (6) and (7) describe the distribution of conditions that we infer for each match in our dataset from the first-innings score and the result of the match. These equations, however, require us to know the prior distribution of conditions, *F*, and the distribution of first-innings performance, G_1.

To infer these, we make use of the assumption that *F* and G_1 are normal distributions, so that $\chi \sim N(\mu_\chi, \sigma_\chi^2)$, $\rho_1 \sim N(\mu_\rho, \sigma_\rho^2)$, and the first-innings score is also normally distributed, $S_1 \sim N(\mu_{S_1}, \sigma_{S_1}^2)$.

Since *F* and G_1 are assumed independent, we have

$$\mu_{S_1} = \mu_\chi + \mu_\rho, \text{ and} \quad (8)$$

$$\sigma_{S_1}^2 = \sigma_\chi^2 + \sigma_\rho^2. \quad (9)$$

We can estimate μ_{S_1} and $\sigma_{S_1}^2$ from the mean and variance of first-innings score in our dataset. By assumption, $\mu_\rho = 0$ so $\mu_\chi = \mu_{S_1}$. Let δ denote the fraction of the variance in first-innings scores arising from variance in conditions so that

$$\sigma_\chi^2 = \delta \sigma_{S_1}^2, \text{ and}$$

$$\sigma_\rho^2 = (1 - \delta)\sigma_{S_1}^2.$$

The final step needed to identify a posterior distribution for conditions, then, is to estimate the decomposition parameter, δ, and the magnitude of the second-innings advantage, γ.

2.3 Estimating δ *and* γ.

To see how we can infer the relative contribution of conditions and performance variances to the observed variance in first-innings scores, consider a special case of the stylised game described above in which both the variance of conditions and the second-innings advantage is zero. In this case, a team scoring at the 90th percentile in the distribution of first-innings scores, say, will have, by definition, performed at the 90th percentile of performance and will have a 90% probability of winning. That is, the graph of Team 1's probability-of-winning versus first-innings score would be identical to the cumulative distribution of first-innings scores. In contrast, let the variance of conditions be positive. Now

a team that scores at the 90th percentile of scores will, on average, have had a better-than-average performance, but will also, on average, be playing in better-than-average conditions and its probability of winning will be lower than 90%. The graph of Team 1's probability-of-winning versus first-innings score will then be flatter than the cumulative distribution of first-innings scores, and the greater the variance of conditions, the greater will be the difference in these two graphs.

This insight is the key to our estimation procedure. We estimate a probit regression of the probability of winning versus the first-innings score, and use the difference in variance between the implied estimated distribution and the variance in first-innings scores to identify the variance of conditions.

Specifically, let f_{S_1} denote the posterior density function of conditions, given a first-innings score of S_1 and not other information, and let $h(S_1)$ be the unconditional density of S_1. We then have (using Equation (3)),

$$f_{S_1}(\chi \mid S_1) = \frac{h_\chi(S_1 \mid \chi) f(\chi)}{h(S_1)} = \frac{g_1(S_1 - \chi) f(\chi)}{h(S_1)}.$$

Let $J(S_1)$ denote the probability that Team 1 wins given a score of S_1 and no information about conditions. We have

$$J(S_1) = \int \Pr(\omega = 1 \mid S_1, \chi) \cdot f_{S_1}(\chi \mid S_1) d\chi = \int G_1(S_1 - \chi - \gamma) \cdot \frac{g_1(S_1 - \chi) f(\chi)}{h(S_1)} d\chi. \quad (10)$$

$J(S_1)$ denotes a probability, but, as it describes an increasing function from the real line onto the unit interval, it also describes the cumulative density function of some distribution that we can interpret as the distribution of the (unobserved) second-innings outcome, S_2. Let μ_{S_2} and $\sigma^2_{S_2}$ denote the mean and variance of this distribution. Monte-carlo investigation of Equation (10) confirms that the distribution is normal with

$$\mu_{S_2} = \mu_{S_1} + \frac{\gamma}{1-\delta},$$

$$\sigma^2_{S_2} = \sigma^2_{S_1}\left(\frac{1+\delta}{1-\delta}\right).$$

Rearranging these gives

$$\delta = \frac{\sigma_{S_2}^2 - \sigma_{S_1}^2}{\sigma_{S_2}^2 + \sigma_{S_1}^2}, \text{ and} \tag{11}$$

$$\gamma = \left(\mu_{S_2} - \mu_{S_1}\right)(1-\delta). \tag{12}$$

Now the mean and variance of the first-innings distribution can be estimated directly from a sample of first innings scores. Since S_2 is a latent variable, we can't observe the second-innings distribution, *J*, directly, but it can be inferred from a probit regression of ω on S_1.

This describes the estimation procedure of the paper. We estimate the means and variances of the first-innings and second-innings distributions from a dataset of ODI matches describing the first-innings score and the result of the match and then use Equations (11) and (12) to infer the decomposition of variance between conditions and performance and the magnitude of the second-innings advantage. This gives us sufficient information to then calculate a posterior distribution of match conditions for each match in our database using Equations (6) and (7).

In the rest of this paper, we implement the procedure described in this section and test some of our maintained assumptions.

4. Description of the data.

The research described in this paper requires two pieces of information: the first-innings score; and the result of the match. This information is publicly available on www.cricinfo.com. We select our time period as the decade of the 2000s; from January 1, 2000 until December 31, 2009. There was a total of 1405 official ODI matches played during this decade.

In order to ensure a robust analysis, there are some additional factors to consider when selecting the data set. As at the date of writing, there are sixteen countries with official ODI status[3]. It is generally accepted among cricket followers that there is a significant gap between the top-eight ranked countries in the world and the remaining countries. We therefore only select matches played between two top-eight countries in our data set. Additionally, to perform the analysis we need an estimate of the distribution of first-innings scores in

[3] These teams are Australia, Afghanistan, Bangladesh, Canada, England, India, Ireland, Kenya, Netherlands, New Zealand, Pakistan, Scotland, South Africa, Sri Lanka, West Indies and Zimbabwe.

completed innings. On occasion, rain interferes in the game of cricket, resulting in a shortened match or even causing the complete abandonment of the match. These matches have the potential to distort our analysis. In order to be included in our data set, at match must meet all of the following criteria:

- the match was played between January 1, 2000 and December 31, 2009, inclusive;
- the match was between two top-eight countries;
- the first innings was not shortened in any way other than the batting team being bowled out before their full allotment of 50 overs had been used; and
- the match was not abandoned without the declaration of a winner.

The total number of matches meeting all these criteria is 784. This forms our dataset for this paper.

Tables 1 and 2 outline the number of matches involving each team and in each venue country. These data show that we have a good distribution of matches.

Table 1: Number of matches played by each team

Country	Bat First	Bat Second	Total
Australia	130	98	228
England	90	78	168
India	106	123	229
New Zealand	82	103	185
Pakistan	103	104	207
South Africa	83	109	192
Sri Lanka	120	85	205
West Indies	70	84	154

Table 2: Number of matches played in each country

Country	Matches
Australia	122
England	81
India	99
New Zealand	73
Other	91
Pakistan	53
South Africa	110
Sri Lanka	83
West Indies	72

Over the ten-year period of our data set, the rules of ODI Cricket changed significantly three times. The rule changes predominately concerned the restrictions on where the bowling

captain can place his fielders. At the beginning of our data set, the fielding captain could have no more than two fielders outside an oval drawn 30 metres from the wickets for a period of 15 overs at the start of the match. For the remainder of the innings, five fielders were allowed outside the oval. In approximately July 2005[4], this was reduced to the first ten overs of the match but the bowling captain also had to select two other blocks of five overs in which the restrictions would apply. These blocks of overs are known as "powerplays". At this time the "supersub" rule was introduced, which would allow each side to make one player substitution at any stage of the game. In March 2006 the supersub rule was cancelled, while the powerplay rule continued. Finally, in October 2008 the powerplay rule was changed to enable the batting side to control when one of the two blocks of powerplay overs was taken. The number of games in our dataset played under each of the four rule regimes are 441, 58, 193, 92. As described in Brooker (2011), there is some evidence that these rule changes have brought about structural breaks in the data, but the sample sizes for all but the first of these regimes are simply too small for us to be able to meaningfully estimate each regime separately. Accordingly, in this paper we group all 784 matches as a single dataset and treat the changing rules as one of the factors leading to variation in batting conditions across the games. As more games are played under the current set of rules, it will become feasible to resestimate the parameters of the model using just those games.

5. Results.

First-Innings Data:

Our estimation procedure relies on the maintained assumption that both the conditions and performance distributions are normal, implying that the distribution of first-innings scores is also normal. We can't test the base assumptions directly, but we can investigate whether the implied normality of first-innings scores is a good approximation. Figure 1 below shows the frequency of the first-innings scores, in bins of thirty runs. The summary statistics for these data are given in Table 3.

[4] At the time of the rule change the old rules were still used for some games for a short period of time.

Figure 1: Distribution of first-innings scores

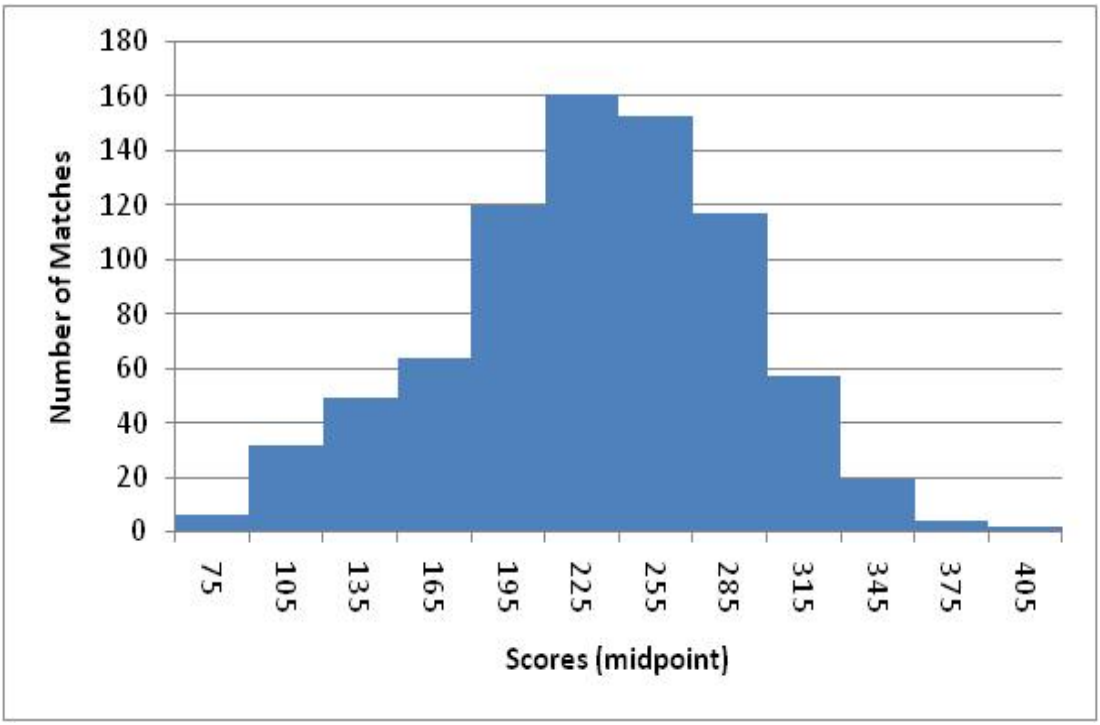


Table 3: Summary Statistics for First-innings Scores

Statistic	Value
n	784
Mean, (μ_S)	243.3
Median	247.5
Variance, (σ_S^2)	3412.5
Skewness	-0.228
Kurtosis	2.888
Excess Kurtosis	-0.112002

The Jarque-Bera test statistic for normality is

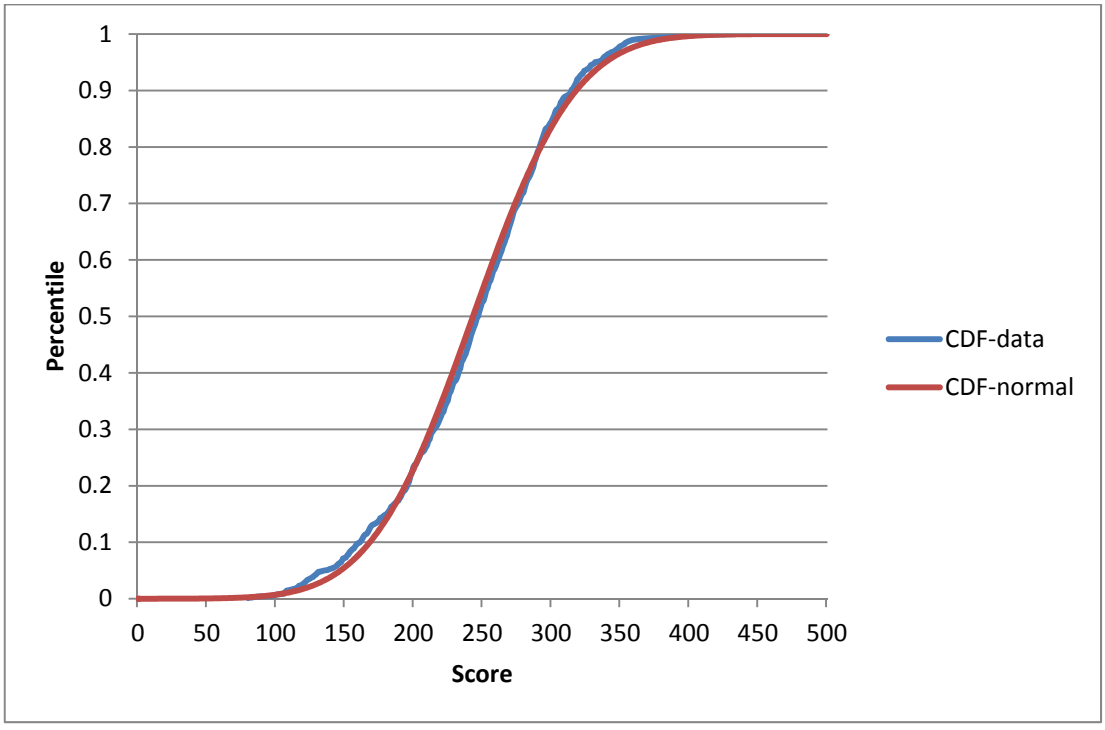
$$JB = \frac{n}{6}(\varsigma^2 + \frac{1}{4}K^2)$$

where n is the number of observations in the sample, ς is the sample skewness and K is the sample excess kurtosis. The Jarque-Bera statistic has an asymptotic chi-square distribution with two degrees of freedom. This chi-square distribution is an approximation of the true distribution of the Jarque-Bera statistic and is prone to making Type I errors. We identify the true distribution of the Jarque-Bera statistic for a sample size of 784 by Monte Carlo simulation. We generate 784 values from the standard normal distribution, calculate the

skewness and excess kurtosis before finally calculating the JB statistic. Repeating this process 10,000 times gives us a distribution of 10,000 JB statistics under the assumption of normality. We are asking the question, were our data normal, how likely are would we be to get a JB statistic as extreme as the one we observe by random chance alone in a sample of the same size as ours. In our data set, $JB = 7.186640$. This value occurs between the 9692nd and 9693rd observations of our simulated distribution of 10,000 JB statistics and therefore we are able to reject the null hypothesis that the data are normally distributed at the 5% significance level but not at significance levels of 3% or less.

As we have previously noted, the assumption of normality cannot be literally true, and so it is perhaps not surprising that one can reject normality at the 5% significance level given how large a sample size we have. We therefore consider the practical significance of any deviation from normality. Figure 2 below compares the distribution function of the first-innings score data with the distribution function of a normal distribution with the same mean and variance.

Figure 2: Distribution function comparison: Data vs. Normal Distribution



To the eye, the assumption of normality does not appear to be a reasonable approximation. As a final check, we provide a numerical descriptive-statistic measure of the deviation of our sample data from a normal distribution. We sort our data in ascending order

of first-innings score, if the data is normally distributed the i^{th} score should be approximately equal to the inverse normal of i/784 our mean and variance. Ignoring the first and last five observations in a bid to eliminate any outliers, the mean-absolute deviation of our observed score from the theoretical score implied by the normal distribution is 4.1 runs, which is small relative to the data average of 243. We take this as indicating that normality is a reasonable approximation for the data distribution.

Second-Innings Results:

We estimate the following probit regression,

$$\Pr(\omega = 1 | S_1) = \Phi(\alpha + \beta S_1),$$

for which the estimated parameters are

$$\hat{\alpha} = -3.292 \text{ and } \hat{\beta} = 0.013.^{5} \tag{13}$$

The function, Φ, is the cumulative standard normal distribution so that

$$Z = \frac{S_1 - \mu_{S_2}}{\sigma_{S_2}} = \alpha + \beta S_1. \tag{14}$$

Setting $S_1 = \mu_{S_2}$ in Equation (14) gives

$$\mu_{S_2} = \frac{-\alpha}{\beta}, \tag{15}$$

and hence that

$$\sigma_{S_2} = \frac{1}{\beta}, \text{ so}$$

$$\sigma_{S_2}^2 = \frac{1}{\beta^2}. \tag{16}$$

Putting our estimated parameters, (13), into Equations (15) and (16) gives

$$\hat{\mu}_{S_2} = \frac{-\alpha}{\beta} = 247.981, \text{ and} \tag{17}$$

[5] Six of the 784 games in our dataset resulted in a tie. Rather than complicating the model by estimating an ordered probit to account for this small number of tied games, we simply repeat each tied match in the data set as one win and one loss and give each of these observations a weight of 0.5. All other observations have a weight of one in the regression, meaning that each match has a total weight of one.

$$\hat{\sigma}^2_{S_2} = \frac{1}{\beta^2} = 5673.117. \quad (18)$$

Estimates of the Conditions and Performance Distributions.

The sample first-innings mean and variance from Table 3, and the estimated second-innings mean and variance from Equations (17) and (18) give us the necessary information to estimate δ and γ. Putting this information into Equations (11) and (12) gives

$$\delta = 0.249, \text{ and}$$

$$\gamma = 3.526.$$

From this estimate, that roughly 25% of the variation in first-innings scores is attributed to variance in the conditions under which matches were played and 75% to variation in the relative performance of the batting team relative to the bowling team, we can parameterise the two distributions as

$$F(\chi) \sim N(243.387, 849.076), \text{ and} \quad (19)$$

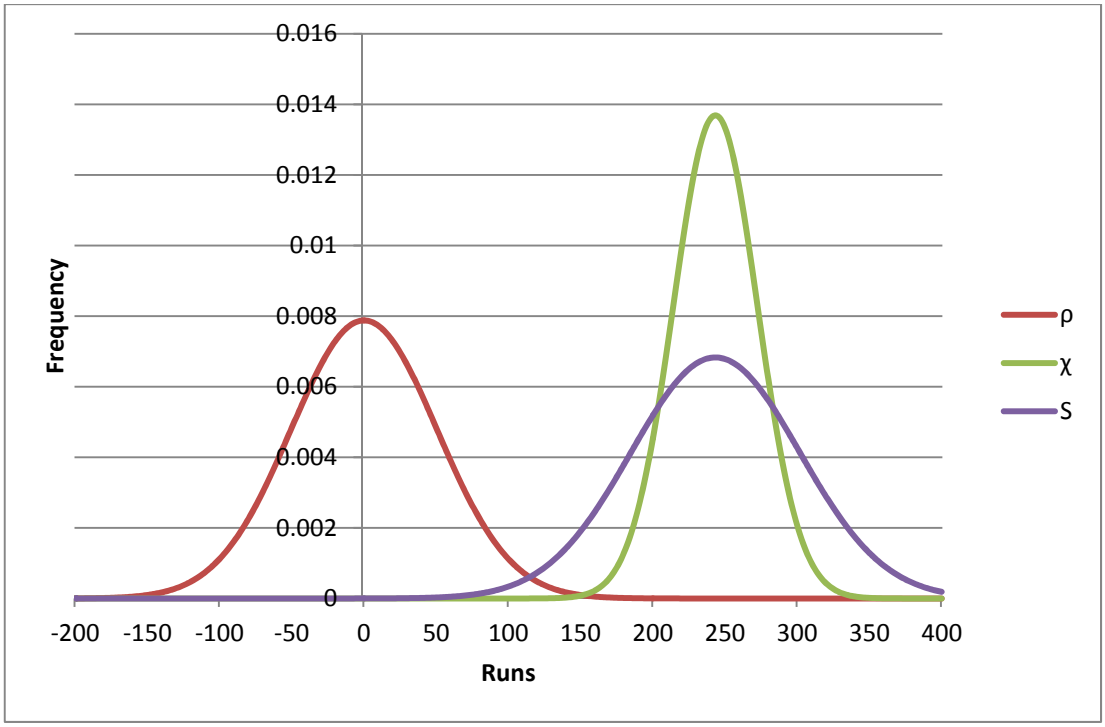
$$G(\rho) \sim N(0, 2, 563.412), \quad (20)$$

which leads to a combined distribution for first-innings scores

$$H(S_1) \sim N(243.387, 3, 412.488).$$

The densities for three distributions are shown in Figure 3.

Figure 3: The performance, conditions and score distributions



Selected Results:

The parameterised distributions given by (19) and (20) give us the information needed to infer a posterior distribution for conditions in any particular match, using Equations (6) and (7). In this subsection, we present some illustrative examples of how the data on the first-innings score and match result can affect the estimated distribution of the conditions applying in that match.

Figure 4 and Table 4 show the posterior distributions of conditions for the two match outcomes where Team 1 scores 243 runs, which is the closest integer to the overall mean in the dataset. It also shows the prior distribution of conditions for comparison. There are two important things to note about these distributions. First, the conditional distributions provide more certainty about what the conditions are like in each game, as their variances are substantially lower than the prior distribution. Second, knowing the result of the game makes a substantial difference to the mean of the conditional distribution. An average score resulting in a win shifts the conditional mean further from the prior mean than an average score resulting in a loss, as there is a smaller than 50% chance of an average score resulting in a win, due to the second-innings performance advantage.

Figure 4: Inferred conditions under different match results

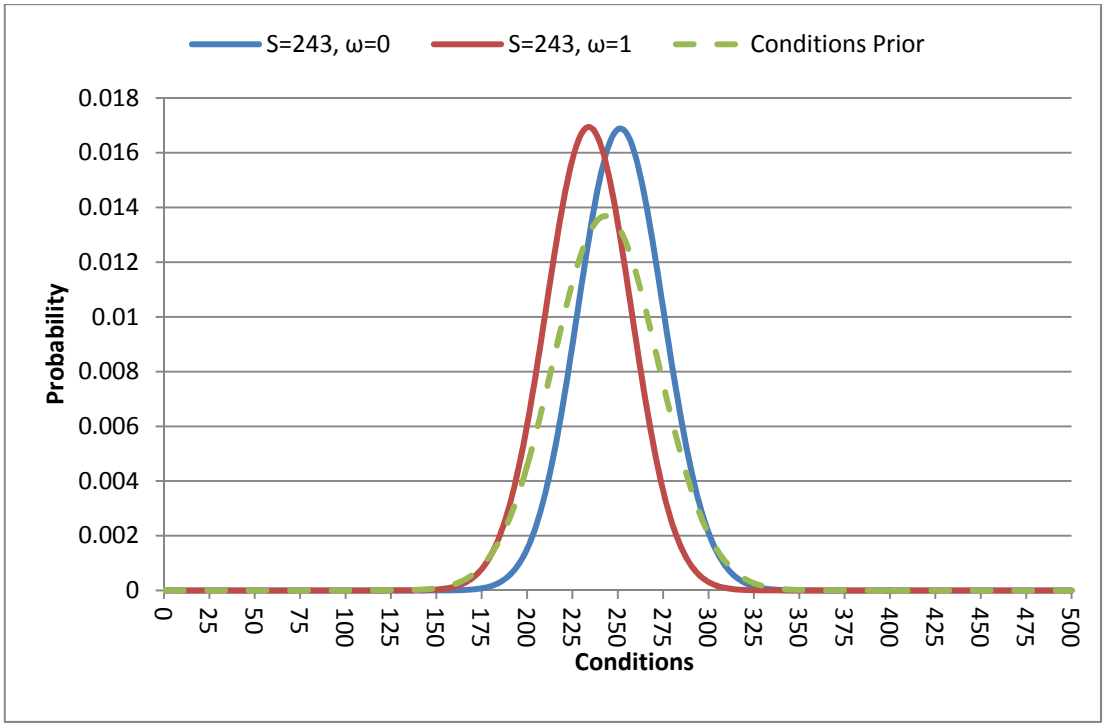


Table 4: Mean and Variance of inferred conditions under different match results.

Conditions Distribution	Mean	Variance
Prior Distribution	243.3	849.1
$S_1 = 243, \omega = 0$	251.7	558.8
$S_1 = 243, \omega = 1$	233.7	555.1

Figure 5 and Table 5 show the posterior distributions of conditions for a match with a particularly low first-innings score of 200, and a particularly high score of 300, both of which resulted in losses for Team 1. The conditional mean shifts much further away from the prior mean when 300 were scored as for Team 1 to lose when they have scored a very high score is a surprising result. The variance is also lower in this situation, implying a greater level of certainty about the conditions.

Figure 5: Inferred conditions under different first-innings scores.

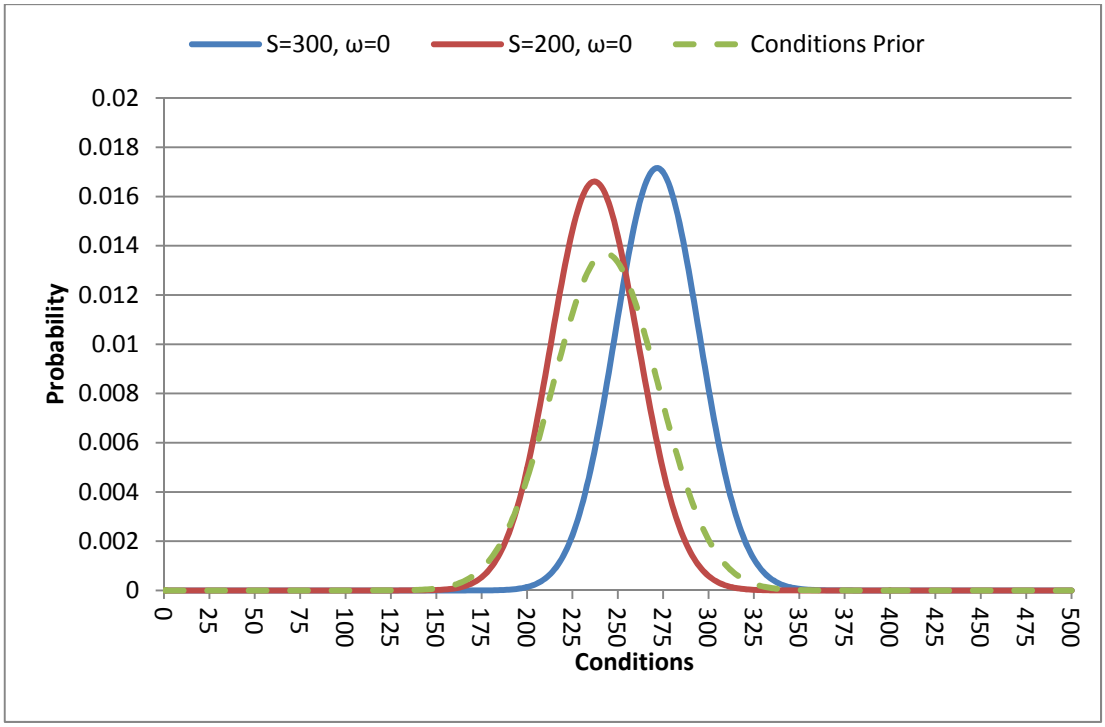


Table 5: Mean and Variance of inferred conditions under different scores.

Conditions Distribution	Mean	Variance
Prior Distribution	243.3	849.1
$S_1 = 200, \omega = 0$	237.5	577.3
$S_1 = 300, \omega = 0$	271.9	541.1

More generally, we plot the means and variances of the inferred conditions distributions for each score and result of the game in Figures 6 and 7, respectively. As expected, the mean of the conditions distribution is higher in games lost by Team 1 than in games won by Team 1, for a given first-innings total. We also note that the further away from the overall mean the first-innings score is, the larger the impact of one result compared with the other on the conditions distributions. Figure 7 shows that we have a higher level of certainty about the value of conditions when the result observed is the less likely one, given the first-innings score.

Figure 6: Inferred conditions means.

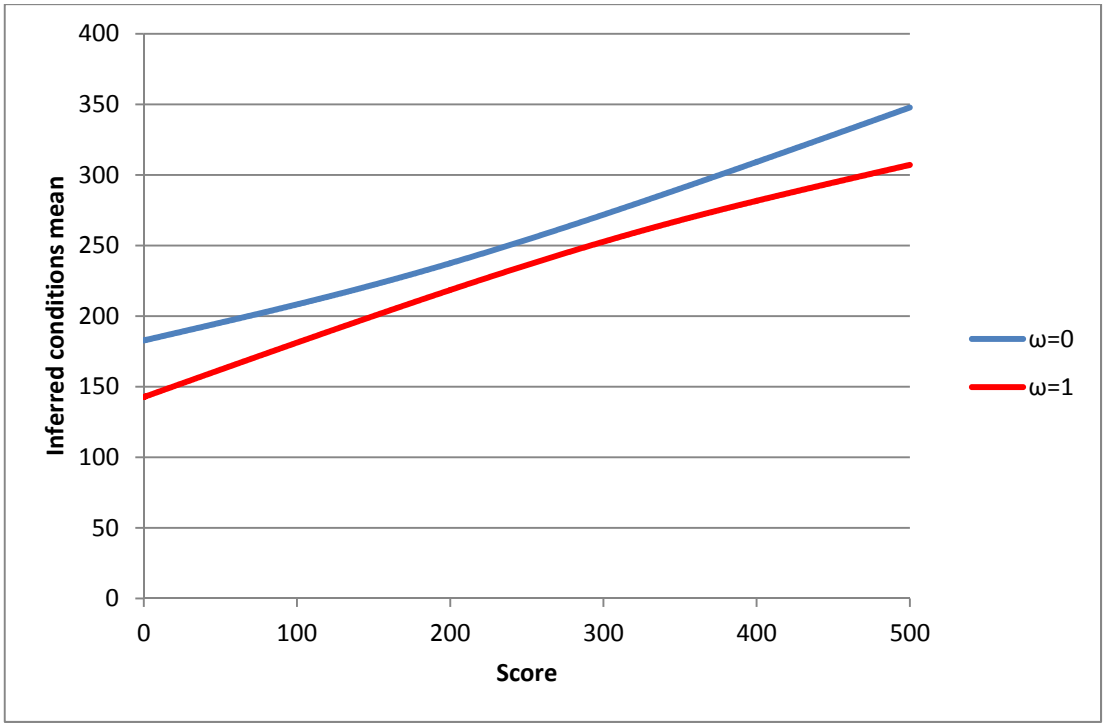
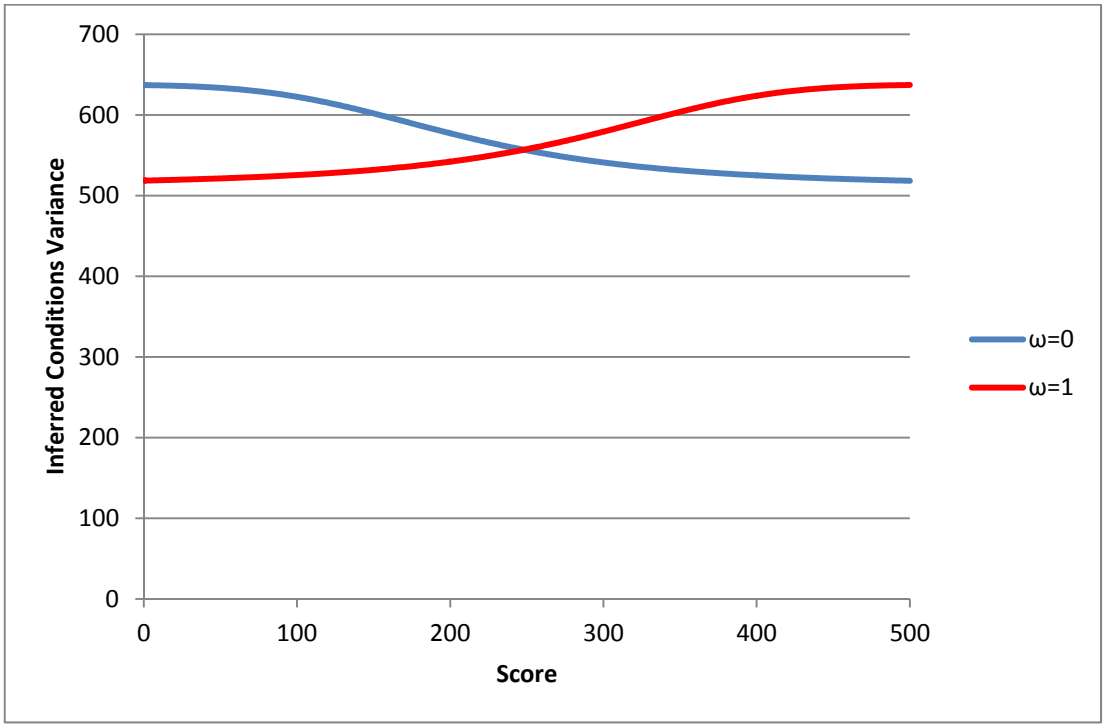


Figure 7: Inferred conditions variances.



6. Diagnostic Assessment of the Posterior Distributions.

As we noted in the introduction, estimates of the conditions applying in a match are useful as they provide a means that a researcher can control for a potentially important confounding variable in empirical work. Our approach does not produce a specific number for each match, but rather a distribution. The ideal way to use this information in empirical work is to sample from this distribution to create an expanded dataset. In this case, it would be useful if the posterior distributions were easily characterised by their mean and variance so that the computer memory requirements required to handle a database with a large number of matches is not excessive.

Testing the Posterior Distributions for Normality.

We choose three situations from our analysis in the previous section in to examine for normality. If the distributions are perfectly normal then they should have skewness and excess kurtosis equal to zero. Additionally, if we take *Z*-scores of the cumulative probability at each value of conditions, these *Z*-scores should be perfectly linear and therefore a linear regression through these *Z*-scores should have an *R*-square value equal to one. We show this information in Table 6.

Table 6: Normality checks for selected conditions distributions.

Conditions Distribution	Skewness	Kurtosis	R^2 of Z-score OLS
$S_1 = 200, \omega = 0$	0.0291	0.0034	0.999899
$S_1 = 243, \omega = 1$	-0.0230	0.0060	0.999929
$S_1 = 300, \omega = 0$	0.0164	0.0052	0.999962

Table 4.8 shows that for our three selected situations, the conditional distributions are very slightly skewed, but are hardly discernable from a normal distribution, with the *R*-squared of the OLS regression of Z-scores on score being so close to one. More generally, we plot the skewness and excess kurtosis for all scores from zero to 500, in Figures 8 and 9, respectively. We see that the conditions distributions are positively skewed when Team 1 loses and negatively skewed when Team 1 wins, with the skewness distributions themselves having the opposite skewness. The kurtosis distributions are more complicated; however, we

see that regardless of the game result the excess kurtosis tends to be positive in the scores around the overall mean score of 243.3, where most scores would actually occur.

Figure 8: Skewness of conditions distributions.

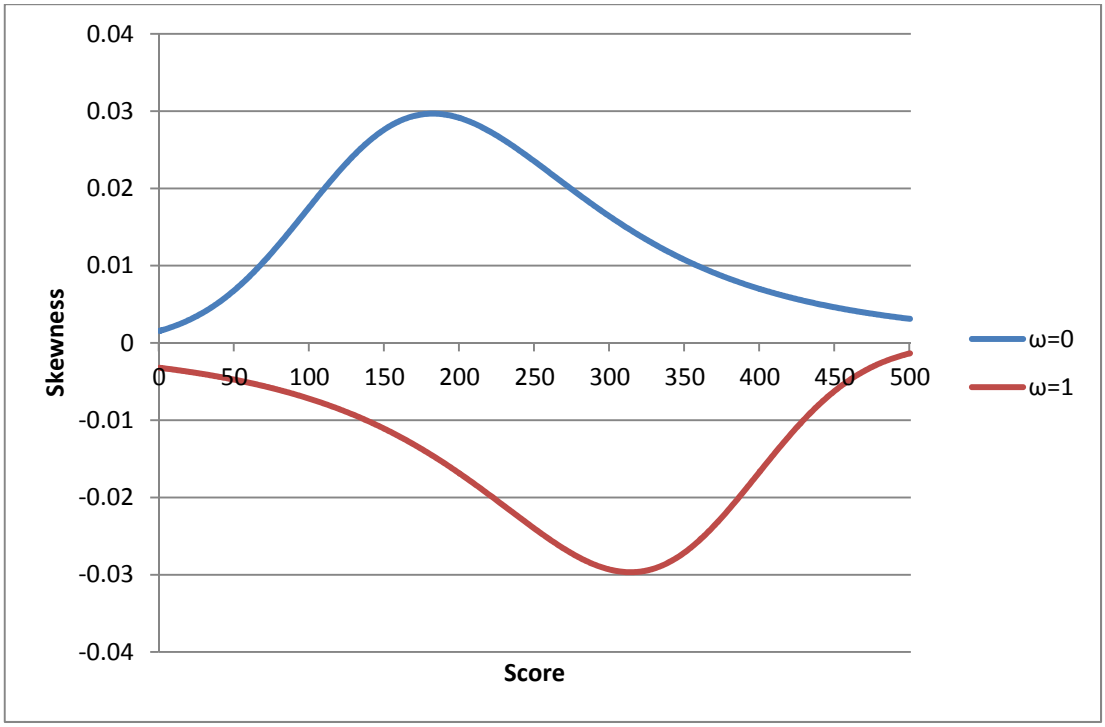
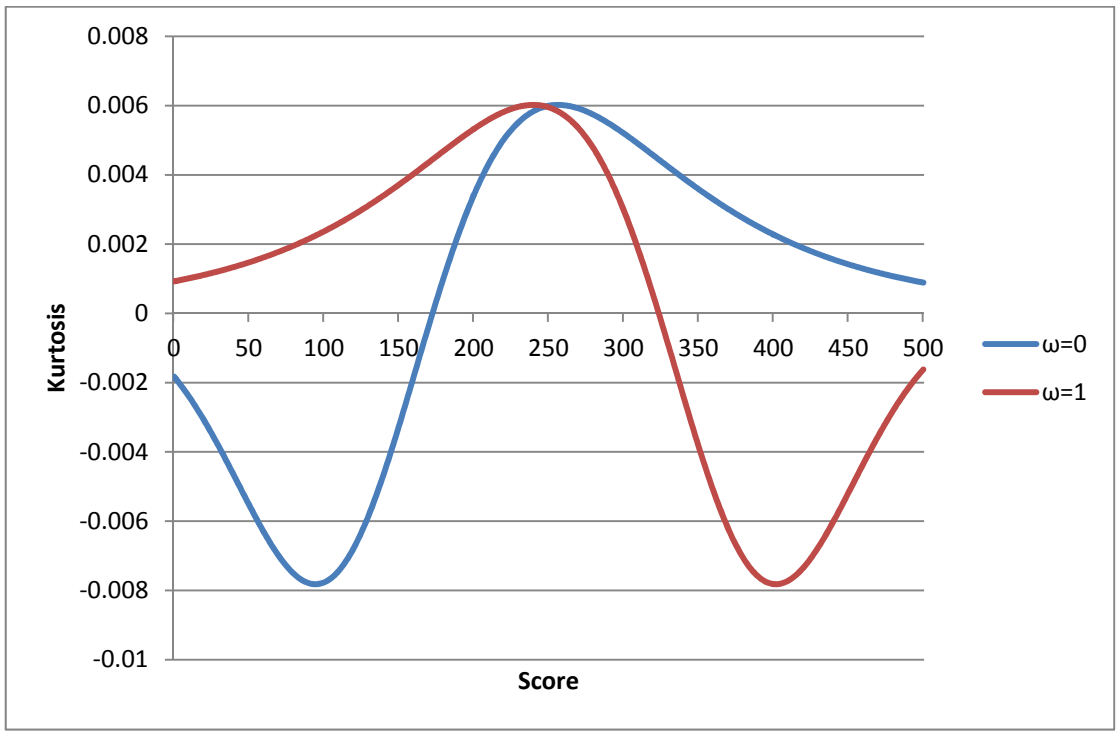
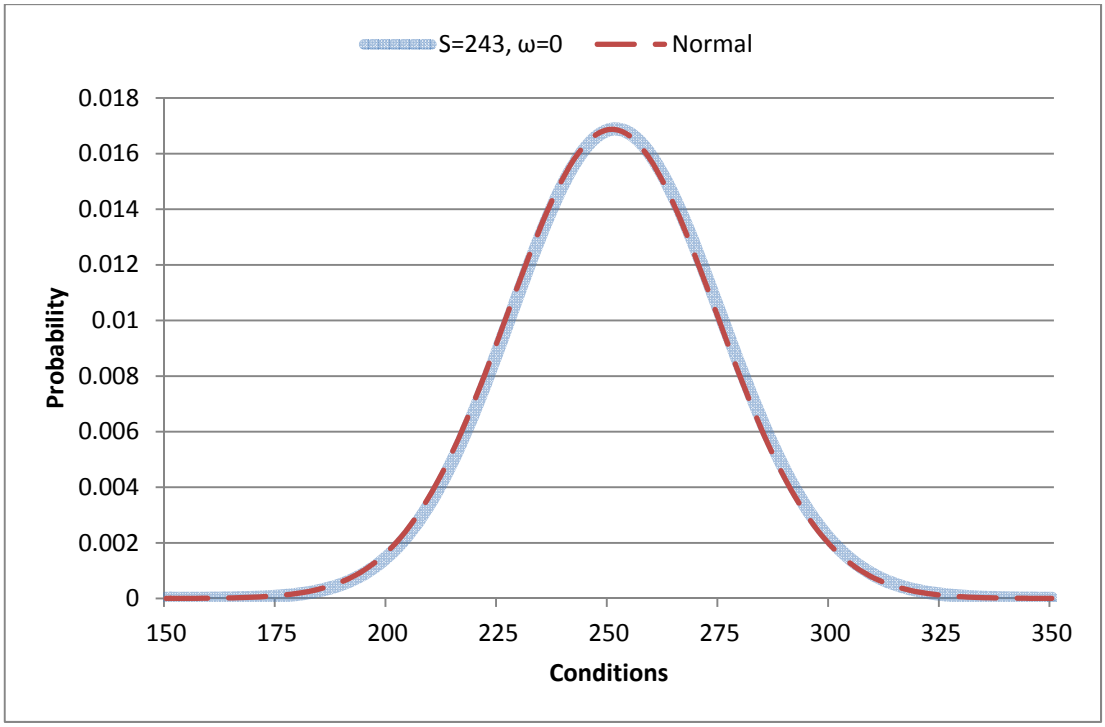


Figure 9: Kurtosis of conditions distributions.



Despite the systematic skewness and kurtosis shown in Figures 8 and 9, the numbers involved are very small. We demonstrate in Figure 10 that assuming normality causes few problems by plotting one of our conditions distributions along with a normal distribution with the same mean and variance. We choose the situation where Team 1 scores 243 and loses the match, as this is a situation resulting in a relatively high combination of skewness and excess kurtosis and therefore should provide an approximate upper bound of the negative impact of assuming normality. The graph shows that we should not be concerned about assuming normality and the cost of this slight simplifying assumption is likely to be trivial in comparison to the benefits provided by the simulation of a larger number of values for conditions in subsequent analyses. To confirm this, we perform the same normality test that we performed on the first-innings score distribution. That is, we simulate 1000 values of conditions from our posterior distribution and sort the data in ascending order of drawn conditions. If the data is normally distributed the i^{th} score should be equal to the inverse normal of i/1,000 for the simulated mean and variance of conditions. Eliminating the five lowest and five highest observations in order to defend against outliers, the mean-absolute deviation of drawn conditions from what would be expected under a normal distribution is 0.7 runs. We conclude from this that, as a practical matter, the posterior distributions can be assumed to be normal in any empirical analysis employing them.

Figure 10: Implied conditions distribution with normality approximation



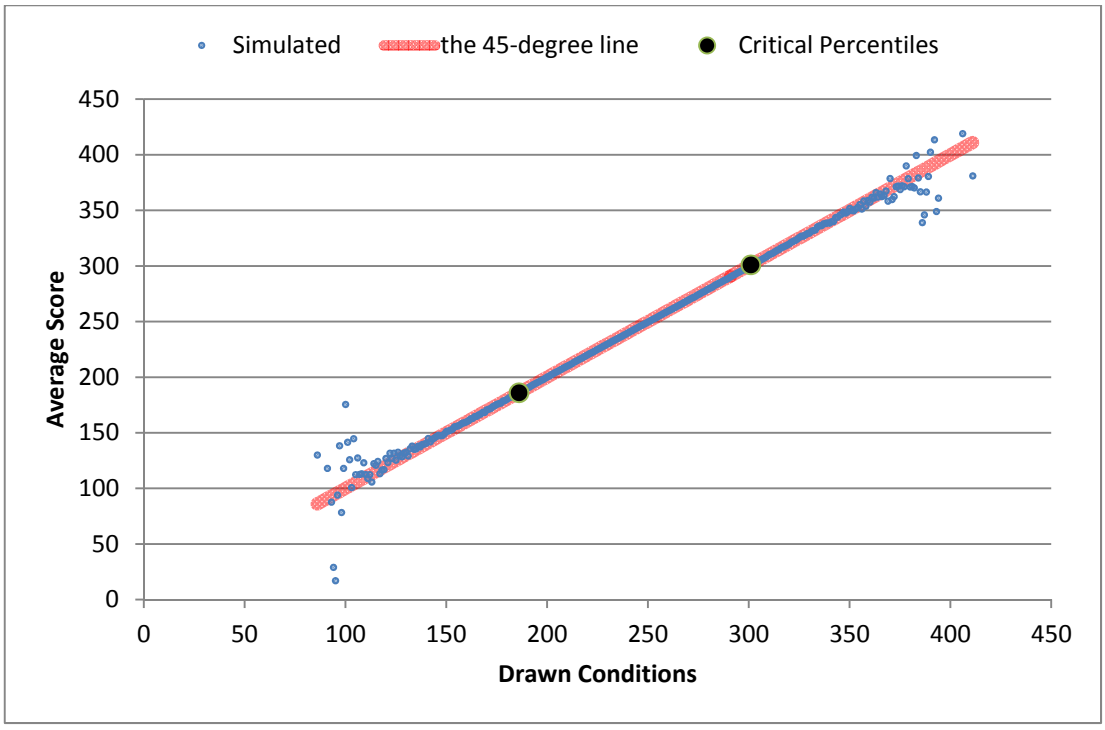
Assessing the fit of the conditional distributions to the data

Theoretically, matches played in conditions with a particular value should result in an average first-innings score of that value. We test our results by employing several Monte-Carlo simulations. There are two motivations behind this analysis. It is important to confirm that our method of calculating conditional distributions for conditions and simulating from these distributions for each given score and result actually works. Additionally, it would be useful to know if our data set has any abnormalities that might lead to the average score for each value of conditions not being approximately equal to that value of conditions. This could occur, for example, if an unusual percentage of games had been won by either team around any particular score. This information could help explain any strange results in subsequent analyses using the conditions variable.

We test the mechanics of our method by randomly drawing one value from the distribution of χ and two values from the distribution of ρ. We add the first draw of ρ to χ in order to determine a first-innings score, S_1, which we round to the nearest integer. If the first draw of ρ is greater than the sum of the second draw plus the performance advantage, this is a win to the team batting first, otherwise it is a loss. We generate 10,000 scores and results by repeating these steps. We then can apply the appropriate posterior distribution for

conditions to each game and we draw 5,000 conditions values from this distribution, again rounding to the nearest integer. This gives us a generated data set with 50,000,000 observations of score and drawn conditions and we can subsequently determine the average score achieved for each (rounded) value of drawn conditions. We plot the results in Figure 11 below, showing the 2.5th and 97.5th percentiles of the overall conditions distributions to show the range of conditions that are most likely to be experienced.

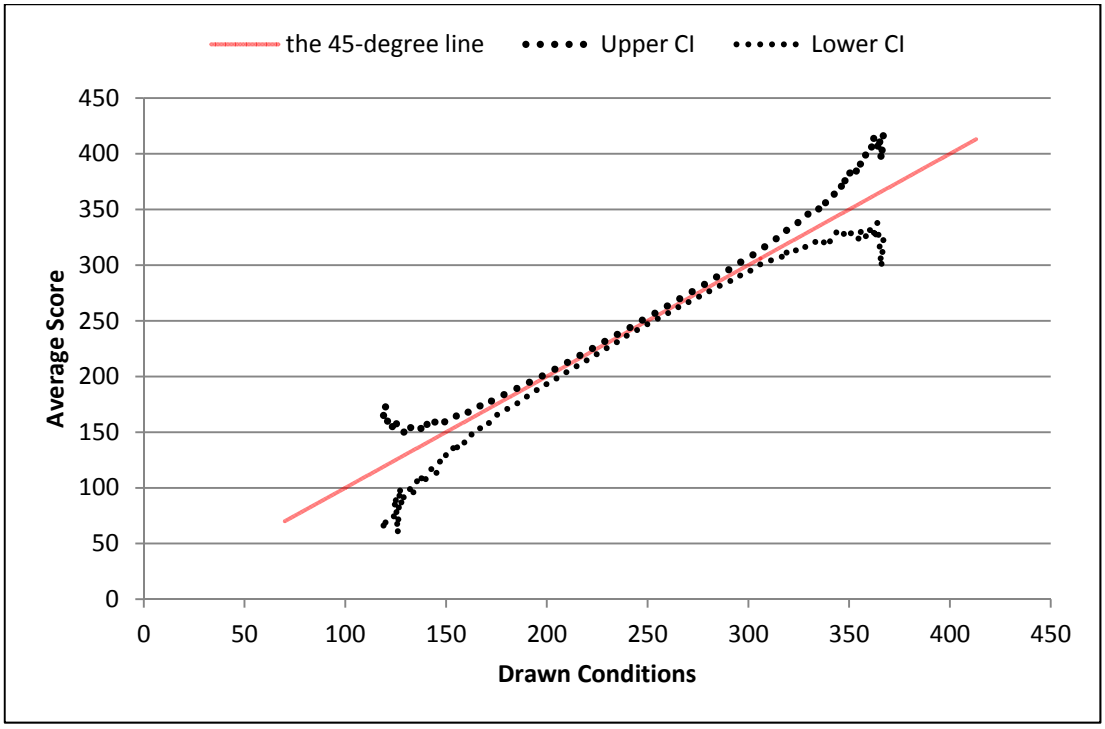
Figure 11: Average Score in generated data set.



It is clear that the average first-innings score in a given set of conditions closely approximates the value of those conditions. We have, to this point, simply confirmed that our method works in theoretical games and we need to check the relationship between inferred conditions and average first-innings score in our data set of matches. Before doing so, we need to think about the amount of deviation from the 45-degree line that would be acceptable, given our sample size. In order to do this, we randomly sample 784 of the 10,000 scores and results previously generated, along with the 5000 draws of conditions for each of those games, and we calculate the average first-innings score for each rounded value of drawn conditions. We repeat this process 100 times, thus generating 100 samples of 784 simulated matches, and generate a 95% confidence interval for the average first-innings score given a particular value of drawn conditions. These confidence intervals are shown in Figure 12. Note

that we exclude from the confidence interval lines where we did not observe at least one draw of a particular value of conditions in all 100 iterations; that is, where in 784 games and 5,000 drawn conditions for each game, we did not observe the particular value of rounded conditions even once.

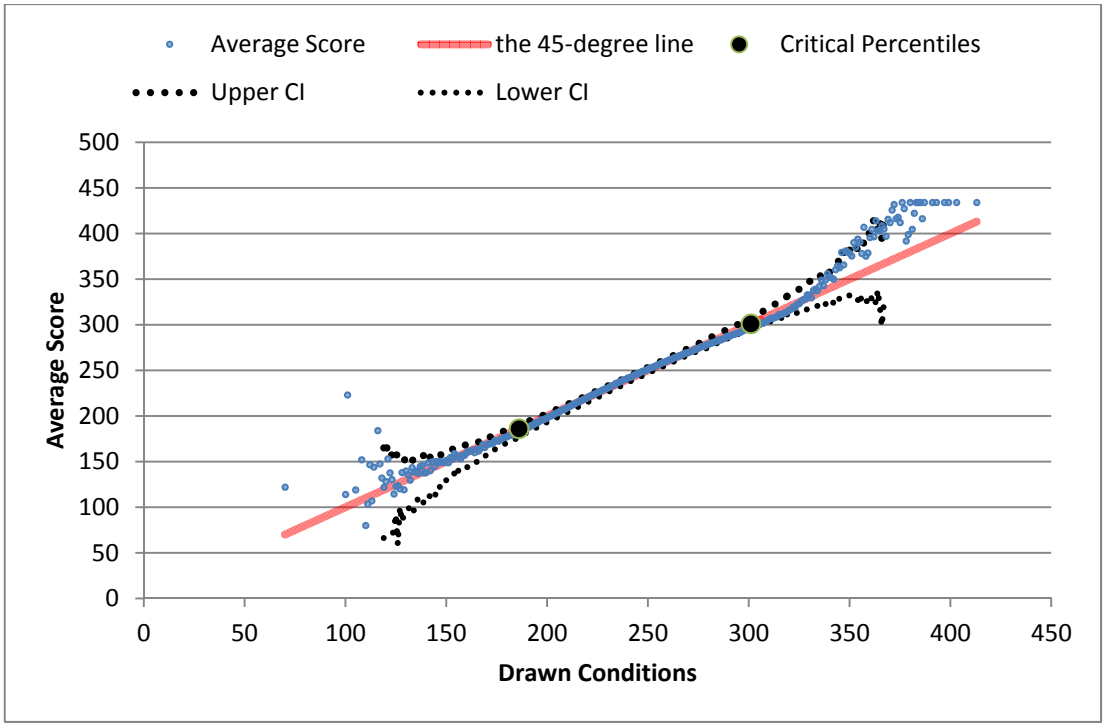
Figure 12: Confidence intervals for a sample size of 784.



In order to assess the fit of our drawn conditions to the theoretical 45-degree line, we take the actual observed first-innings score and result from our 784 games and apply the mean and variance for the conditions distribution implied by each score and result. As in the previous simulation, we generate 5000 values for conditions from the conditional distribution for each match.

Figure 13 shows the average first-innings score for each value of conditions. We see that again the draws from the conditions distributions do a good job of predicting what the average first-innings score will be, particularly within the range in which 95% of conditions fall. The high draws of conditions result in an average score close to the upper bound of the confidence interval over the range that the confidence interval is estimated; this is likely to be due to a particularly unusual game where Australia scored an extremely large total of 434 against South Africa and remarkably lost the match.

Figure 13: Average Score in observed data set



7. Discussion.

By assuming a functional form for a model of first-innings score, determining the contribution to the total score variance of each component in the model and applying Bayes' Rule, we have obtained information pertaining to a critical but unobservable variable. This information is in the form of a distribution that is conditional on the first-innings score and the result of the game. We believe this approach can result in a large improvement to empirical analysis of data from ODI cricket matches, relative to the current situation in which conditions is a missing variable, with highly problematic implications for the inferences made from statistical analysis.

Of course, our identification strategy rests on a number of maintained assumptions about the normality and independence of the underlying distributions that we feel are justified on *a priori* grounds based on our knowledge of the game of cricket, and are not inconsistent with the available data.

There are, however, two assumptions we have made that could be relaxed in future work. First, the identification of the means and variances of the underlying conditions and performance distributions took as given the mean and variance of the first-innings scores in

our dataset, and the mean and variance of the implied distribution estimated by a probit regression of result on first-innings score. These sample means and variances, however, are of course subject to sampling error. It would be possible to take that error into account when constructing the posterior distributions of conditions. With a large dataset of 784 games, however, the effect of mis-estimation of the means and variances is likely to be small and would not justify the additional complexity of an estimation strategy taking the sampling error into account.

The second area where a useful extension is possible is in allowing for predictable differences in ability between teams. We expect that doing so would result in attributing a greater fraction of the variance in first-innings scores to variation in batting conditions. The intuition for this is as follows. Some of the variation in ability across teams at any particular time is correlated across both batting and bowling/fielding. That is if the expected performance of Team 1 against Team 2 is positive, then Team 2's expected performance against Team 1 would be negative (in other words, the distributions G_1 and G_2 would not be independent. This implies that if there were no variance in conditions, the cumulative density of first-innings scores would show a *higher* variance than the implied distribution from regressing the probability of Team 1 winning on first-innings score. Since our method of inferring the variance in conditions relies on mapping the excess variance in the implied second-innings distribution into the variance of conditions, this non-independence of G_1 and G_2 would lead to our underestimating the variance of conditions.

Making this adjustment remains for later work, as it would require building up a dynamic dataset of team ability. For now, we simply note that most existing empirical work implicitly assumes a zero variance due to conditions, so even if we have underestimated the true conditions variance, our approach does represent a step in the right direction.

References.

Brooker, S. (2011), "An Economic Analysis of Ability, Strategy, and Fairness in ODI cricket", unpublished Ph.D. thesis, University of Canterbury.

Carter, M and Guthrie, G (2004), "Cricket Interruptus: Fairness and Iincentive in Interrupted Cricket Matches", *Journal of the Operational Research Society* **55:** 822-829.

Clarke, Stephen R. (1988), "Dynamic Programming in One-Day Cricket – Optimal Scoring Rates", *Journal of the Operational Research Society* **39(4):** 331-337.

Duckworth, F.C. and Lewis, A.J. (1998), "A Fair Method for Resetting the Target in Interrupted One-Day Cricket Matches", *Journal of the Operational Research Society* **49(3):** 220-227.

Duckworth, F.C. and Lewis, A.J. (2005), "Comment on Carter M and Guthrie G (2004). Cricket Interruptus: Fairness and Incentive in Limited Overs Cricket Matches", *Journal of the Operational Research Society* **56:** 1333-1337.

Preston, Ian and Thomas, Jonathan (2002), "Rain Rules for Limited Overs Cricket and Probabilities of Victory", *Journal of the Royal Statistical Society: Series D (The Statistician)* **51(2):** 189-202.

Appendix : The Necessary Basics of the Game of Cricket

Cricket is a sport played between two teams of 11 players on a large, approximately circular field with a 22-yard-long strip of pressed clay, soil and grass known as a "pitch" in the centre. One team will initially be the bowlers and the other team will be the batsmen. All 11 members of the bowling team are on the field while only two members of the batting team are on the field at any one time. The basic idea of the game is relatively simple. A bowler bowls a ball from one end of the pitch by releasing it with a straight arm action in the direction of the batsman. The ball will usually bounce once before reaching the batsman. The two main goals of a batsman are to score "runs" and avoid getting "out". A run is scored each time a batsman, having hit the ball with his bat, running to swap ends of the pitch with the other batsman. Alternatively, a batsman may score an automatic four or six runs by hitting the ball so far that it leaves the playing field. These automatic runs are known as "boundaries", with four being scored if the ball bounces before leaving the playing field and six otherwise. If a batsman is "out" then his turn at batting is over and he must leave the field to be replaced by a team mate.

The batting side may continue batting until ten of the 11 members of their side are out, then the two teams switch roles. A team's turn at batting is called an innings and each team will have either one or two innings depending on the type of game. In general, the team that scores the highest number of runs wins the game.

There are three main versions of the game. In test cricket, the traditional form of the game, each team bats for two innings and a match lasts a maximum of five days, with the match being declared a draw if it is not finished in this time. One Day International (ODI) cricket allows each team to bat for one innings but with a limit of 300 balls per innings. The innings finishes when ten batsmen are out or the 300 balls are up. As the name suggests, this type of game is all over in a day, running for approximately eight hours. Twenty20 cricket is the newest form of the game and is similar to ODI cricket except that the limit is 120 balls per innings and the game takes approximately three hours. In this paper, we consider only ODI cricket.