PAIRWISE RELATIONS
IN PRINCIPLE AND IN PRACTICE

Ian Allister Hamilton

A thesis submitted for the degree of Doctor of Philosophy in Statistics (Research)

January 2023

Contents

List of Figures

List of Tables

Acknowledgements

There are three people who are due special thanks on completing this thesis. First and foremost, I would like to thank my wife, Amber, whose patience and encouragement saw me through. Second, I would like to thank my supervisor, Professor David Firth. I am grateful for his guidance and wisdom (statistical and otherwise) and for sticking with me. Third, I would like to thank my second supervisor, Dr Nick Tawn. His enthusiasm was invigorating, and his conversations with David were enlightening. I would also like to thank my personal tutor, Dr Elke Thonnes, whose advice was always sage. Additionally, I have a large number of people who have provided useful feedback on earlier versions of some of this work. For Chapter 1, they include Patrick Zietkiewicz, Conor Hughes, Philip Sterzinger, and Asma Saleh. For Chapters 2 and 3, they include Professor David Papineau, Dr Theron Pummer, Dr Ben Ferguson, Dr Peter Wilson, Andrew Rubner, Dr Tom Parr, Dr Seth Bordner, and several anonymous reviewers. For feedback on chapter 3, I would like to particularly thank Dr Aaron Harper, whose thoughts and feedback contributed substantially to the work here. I was funded by EPSRC grants A.STAA.1617.IXH and A.STAA.1819.IXH.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. The work contained within is original, except as acknowledged, and has not been submitted previously for a degree at any university. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made.

Abstract

Pairwise relations are ubiquitous. They occur in any population where individual items are identifiable and interact somehow. Based on these pairwise interactions, it is often of interest to discern a rating — a value reflecting the degree to which an item has some quality, a ranking — an ordering with respect to some quality, or to identify communities of items and the nature of those. In this thesis, I am primarily interested in examples where we might be guided in this practice by something other than purely pragmatic concerns of computational efficiency, predictive ability or familiarity, but rather by more principled motivations.

The thesis comprises five chapters, each an independent reflection related to this subject. The first three chapters deal with aspects of those principled motivations. Chapter 1 looks at motivations for a particular well-known statistical rating model, the Bradley-Terry model. Chapter 2 takes a fundamental question in the philosophy of ranking, looking at the nature of the relation that ranking exercises often seek to model by addressing the philosophical controversy over the transitivity of the 'better than' relation. Chapter 3 presents an argument, based in sports philosophy, for principles that should guide the selection of ranking models in the context of competitive sport. In Chapter 4, I turn to an example of the practice of ranking based on pairwise comparison, investigating the statistical measures used to assess the reliability of rating exercises in Comparative Judgement, an educational assessment practice. In Chapter 5, I consider an example of community detection, identifying the relative importance of different factors in the propensity for school rugby union fixtures to exist and thus giving clues as to the nature and extent of the 'old boy' network.

Introduction

Pairwise relations appear in many contexts. They may occur organically, such as in food webs, the internet, or protein, economic or social networks, or be stimulated for some purpose, such as in sports tournaments or as a means of assessment. Sometimes, the relationship is hierarchical, most obviously in the form of a contest, for example in sport or in competitive animal behaviours. Sometimes, the relationship is associative, maybe even purely reciprocal, for example Facebook friendship or transport links between cities. Often, the pairwise relationship has both a hierarchical and an associative component and we may be interested in either or both of these features. For example, in an academic citation network, defined by the pairwise relation of one paper citing another, we may be interested in the hierarchical — the influence of particular papers, authors, journals or topic fields. Alternatively we may wish to investigate the associative relations — determining communities of academic endeavour based on the citation network. This thesis deals mostly with hierarchical relations, and the endeavour of rating or ranking. 'Rating' is taken to mean the assignment of a value to an item on some common scale, with 'ranking' being the ordering of the items, often, though not always, based on a rating, both with respect to some quality of interest. I also look at a question of association.

Faced with the task of ranking a set of items based on their pairwise relations, there are a cornucopia of options. For example, leaning on the taxonomy offered by De Bacco et al. (2018), there are: spectral methods, such as PageRank (Page et al., 1999), Eigenvector centrality (Bonacich, 1987), or Rank centrality (Negahban et al., 2017); ordinal ranking mathods, free from ratings, such as Minimum Violation (Slater, 1961; Ali et al., 1986), SerialRank (Fogel et al., 2014), and SyncRank (Cucuringu, 2016); axiomatic approaches from the Social Choice literature such as Fair Bets (Daniels, 1969; Slutzki and Volij, 2005) or Generalized Row Sum (Chebotarev, 1994); methods motivated from particular fields such as Colley and Massey matrix methods from sport (Colley, 2002; Massey, 1997), trophic levels from ecology (Lindeman, 1942), or Trueskill from online gaming (Herbrich et al., 2006); statistical models such as the Bradley-Terry (Zermelo, 1929; Bradley and Terry, 1952) and Thurstone-

Mosteller (Thurstone, 1927b; Mosteller, 1951) models; and exchange-based models such as Elo ratings (Elo, 1978). Besides these, there are often context-specific or tweaked or rediscovered versions of the above. How then to choose an approach?

Often, that decision may be driven by pragmatic concerns. Sometimes the size of the data set or the speed requirement of the response forces computationally cheap approaches to be preferred. Sometimes predictive ability will be a determining factor. For example, if one is paid by having a user click on a link then it is of commercial interest to offer links that are most likely to be clicked on (prediction) as quickly as possible (computational speed). Sometimes, familiarity of an approach might be a concern for the modeller. It may be attractive to use an approach with which they are comfortable and about which they know the pitfalls. In this thesis, I am primarily interested in examples where we might be guided by something other than these pragmatic concerns, rather by more principled motivations.

In Chapter 1, I consider this question through an investigation of the Bradley-Terry model. The work seeks to bring together a diverse set of motivations for its use. Some of these motivations will be context-specific, such as those related to penalty shoot-outs or decision-making constructs, meeting the conditions of particular situations and assumptions. Others will be much more broadly applicable, such as those that appeal to constrained simplicity, entropy or likelihood maximisation or the Bradley-Terry model's status as the unique model whereby the number of wins for each item is a sufficient statistic. This chapter includes well-known, lesser-known and novel motivations, as well as linkages between them. In doing this, I lay the foundation for the Bradley-Terry model's later use in a practical situation in Chapter 4.

In Chapter 2, a fundamental controversy from the Philosophy literature related to ranking is addressed — the transitivity of 'better than'. If A is better than B and B is better than C, is A better than C? Transitivity is a definitional assumption to the exercise of ranking, but, as I argue, it may not be true. In particular, I discuss the difficulty of reconciling a pairwise notion of 'better than' with a population-level notion of 'better than'. I make an argument that there may be situations where the pairwise unit is so fundamental to the context that 'better than' should be understood primarily on a pairwise basis and therefore may be intransitive. The main argument is presented in the clarifying context of competitive sport, most relevant to this thesis. But it is also shown how the argument may be extended to more directly address the philosophical controversy in a moral realm.

In Chapter 3, the question of what principles ought to guide ranking is considered from a philosophical perspective and in relation to a context with popular appeal, that of sports ranking in unbalanced league tournaments. In these tournaments, the

schedules of participating teams may vary substantially in the strength of opposition, the number of matches played, and the proportion of matches played at home. Perhaps the most high-profile unbalanced league ranking exercise in the world, the annual selection of NCAA basketball teams for participation in the March Madness tournament, is taken as an example. The aim of this chapter is not to argue for any single ranking method, but to attempt to make a rigorous argument for what criteria a ranking approach in that setting should seek to meet. The argument is based on the norms of sports ranking, well-established across many sports, including basketball, in the USA and worldwide. A set of seven principles are argued for that may be used to guide the selection of a ranking method. The current approach in NCAA basketball is found to be deficient with respect to these principles and recommendations for change are made.

With Chapter 4, we move on to an example of rating based on pairwise comparison as it is practised. Comparative Judgement (CJ) is the exercise of rating a population of items by ranking subsets of the population. In the majority of cases, and in the example dealt with here, that subset consists of two items and therefore a pairwise comparison. The use of CJ in education has been growing, with it recently being employed in primary and secondary schools and universities in the U.K. It has appealing features. For example, advocates of the approach argue that it: allows more holistic assessment, mitigating 'teaching to the test' and allowing the evaluation of tasks not well-suited to rubrics; greatly reduces the impact of judge inconsistency; allows for robust progress and cohort comparisons against meaningfully age-graded scales; and, in the context of peer assessment, can turn the provision of marks into a formative learning experience.

However, the efficiency of the approach — the amount of assessor time taken to produce assessments of a sufficient reliability — has been questioned. A natural way to improve efficiency would be to attempt to optimise the scheduling of the pairwise comparisons. But such efforts have been frustrated by the inflationary impact they have on the measures used to evaluate the reliability of an assessment. Chapter 4 is grounded in that CJ literature and its aim is to explain why these measures may be misleading and what might be done about it. It highlights elements, such as the importance of the estimation method and the nature of bias in this context, that seem to be under-appreciated within that literature, and discusses the practical and conceptual limitations of current measures.

Chapter 5 pivots to considering an example of associative relationships. It takes data from school sports fixtures to investigate the nature of what leads to fixtures occurring in schoolboy rugby union and therefore perhaps to the nature of institutional links, and thus what might be termed the 'old boy network'. It demonstrates

how a variety of statistical methods related to a latent space approach can be used to give a consistent but nuanced interpretation of the situation.

The five chapters that constitute this thesis may be read as five independent reflections on various aspects of the nature of pairwise relations and a quantitative understanding of them. Nevertheless, there are links across them that are highlighted and it is to be hoped that they reflect a cohesive perspective in advocating for a broad consideration on the quantitative approaches taken in the investigation of pairwise relations, and that in each instance they provide novel and useful insights.

Chapter 1

The many routes to the ubiquitous Bradley-Terry model

Abstract

The rating of items based on pairwise comparisons has been a topic of statistical investigation for many decades. Numerous approaches have been proposed. One of the best known is the Bradley-Terry model. This chapter seeks to assemble and explain a variety of motivations for its use. Some are based on principles or on maximising an objective function; others are derived from well-known statistical models, or stylised game scenarios. They include both examples well-known in the literature as well as what are believed to be novel presentations.

1.1 Introduction

The first conference that the author attended as a PhD student was an American sports statistics conference. He presented a poster related to the Bradley-Terry model. As a retrodictive model on rugby union in a sea of American sports predictions it felt a little out of place. But a kind attendee took pity on him and decided to engage him with a question. She asked: "Why would I choose Bradley-Terry rather than the Thurstone model?" (by which he took her to mean what is more commonly referred to as the Thurstone-Mosteller model). He flummered a vague response involving analytic niceness and simplicity — he suspects Occam's razor even got a mention. She looked suitably unconvinced. It is to be hoped that this chapter represents a more ordered response to the conference interlocutor and an aggrega-

tion of, as David (1988, p.13) puts it in his canonical survey of pairwise comaprison methods, "the many routes to the ubiquitous Bradley-Terry model."

The main original contribution of the work is in aggregating the motivations for the Bradley-Terry model. This is an updating of and addition to Bradley (1976), and the relevant parts of David (1988), with the presentations of Sections 1.3, 1.5, 1.6, and 1.7 not discussed in those works. In doing so, it takes in a diverse scope of motivating ideas including likelihood and entropy maximisation, psychological choice and sensation models, a prominent Markov chain Monte Carlo method, other well-known rating models such as PageRank and the RPI of American college sports, sudden-death play-offs, pub pool norms and the British playground game of conkers. It is hoped that such an aggregation serves to demonstrate the broad appeal of the Bradley-Terry model in many settings.

The chapter also offers a number of novelties including: the explicit discussion of the Bradley-Terry model in the context of an exponential family of distributions, which provides a uniting theme to a number of the more notable motivations; a formalisation of perhaps the most intuitive motivation for the model, by proposing an explicit measure for the simplicity of a model in the pairwise comparison scenario and showing that, under plausible constraints, Bradley-Terry is the model that maximises this measure; and a demonstration of how the ideas behind the ranking method of Wei (1952) and Kendall (1955) and the heursitic of the Ratings Percentage Index (RPI) can be related to the Bradley-Terry model through Perron-Frobenius Theorem.

The scenario under consideration in this chapter is one where there is a desire to create a ranking of items based on the observation of a set of binary-outcome pairwise comparisons. One popular approach to creating rankings is to construct a uni-dimensional rating, and then order items by their ratings. The Bradley-Terry model achieves this by defining the probability of a preference for alternative i over alternative j in a pairwise comparison as

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j},$$

where π_i is a positive-valued parameter that may be interpreted as a rating of alternative i, with a higher rating indicating greater 'strength' or 'worth'. An equivalent characterisation is to consider it as a member of the class of generalised linear models (McCullagh and Nelder, 1989) with

$$F(p_{ij}) = \lambda_i - \lambda_j,$$

where λ_i is a real-valued parameter indicating the strength of i, and F is taken as the logit function. The Thurstone-Mosteller model (Thurstone, 1927a; Mosteller,

1951), about which the interlocutor asked, is derived from taking F to be the probit function instead. In practice, as Stern (1992) notes, the models are often empirically very similar.

The Bradley-Terry model has formed the basis for many models in many contexts over time including those for journal citations (Stigler, 1994), college sports (Wobus, 2007), animal behavior (Stuart-Fox et al., 2006), risk analysis (Merrick et al., 2002), wine tasting (Oberfeld et al., 2009), university ranking (Dittrich et al., 1998), font selection (O'Donovan et al., 2014), exam marking (Pollitt, 2012b), and of course in chess, which was the subject of the original work by Zermelo (1929), as well as being the subject of the popular closely-related ranking method proposed by Elo (1978), which is widely known and is still in use in the sport today.

Originally documented by Zermelo (1929), the Bradley-Terry model took the name by which it came to be commonly known when Bradley and Terry (1952) independently rediscovered it. Following the work of Thurstone (1927a,b,c) and Zermelo (1929), paired comparison methods saw little development for the best part of a quarter of a century until they became an active area of investigation in the 1950s and 60s. Much of this work took place in the context of the psychological literature, with Luce's Choice Axiom (Luce, 1959) a particularly notable contribution, leading to the model sometimes being referred to as the Bradley-Terry-Luce (BTL) model. A number of these works showed how the Bradley-Terry model could be derived based on highly plausible axioms or desirable model features (Good, 1955; Luce, 1959; Bühlmann and Huber, 1963; Luce and Suppes, 1965). Towards the end of this period, Thompson and Singh (1967) demonstrated that a consideration of extreme value distributions within a discriminal process leads to the Bradley-Terry model, and Daniels (1969), in a highly original paper, noted the links between the Bradley-Terry model and what might now be recognised as an undamped PageRank (Page et al., 1999).

For further details of the development of the model up to this point David (1988) provides a thorough account of the paired comparison literature more generally, Davidson and Farquhar (1976) provides an interesting snapshot of the literature related to the Bradley-Terry model at the end of this period, and Glickman (2013) is a highly readable account of the history, particularly as it pertains to the contribution of Zermelo.

The next significant contributions to motivating the Bradley-Terry model came from Henery (1986) and Joe (1988) in identifying the model as the result of maximising an objective function subject to a suitable constraint. The later work (Joe, 1988) seems to have been unaware of Henery (1986), but provides a more complete presentation. As well as considering the Bradley-Terry model as a maximum en-

tropy model and noting its relationship to an appropriate sufficient statistic, Joe (1988) also explicitly notes the link to a maximum likelihood derivation. A number of motivations in this chapter are based on game-style scenarios. Perhaps the most interesting paper related to this also comes from this period (Stern, 1990).

More recently Slutzki and Volij (2006), Negahban et al. (2012), Maystre and Grossglauser (2015) and Selby (2020) provide more detailed accounts of the link between the Bradley-Terry model and the limiting distribution of a Markov Chain, and thereby to an undamped PageRank. The Social Choice literature provides an interesting perspective, of which Slutzki and Volij (2006) is perhaps the most notable example in the present context. Much of the relevant work investigates axiomatisations of ranking methods, with Chebotarev and Shamis (1998) providing a thorough summary of considered conditions. An early and influential contribution of this type was due to Rubinstein (1980), which in axiomatising the number of wins as a rating in the very specific case of a round-robin tournament with a single round could be interpreted as a motivation for Bradley-Terry under those conditions (further details of Rubinstein's axiomatisation are discussed in Chapter 3).

The chapter proceeds by dividing the motivations up into six types: axiomatic motivations; objective function maximisation; discriminal processes; standard models; game scenarios; and quasi-symmetry and consistent estimators. These categorisations are somewhat arbitrary, and linkages exist across them which will be highlighted, but for the present purpose they provide a useful means to order the work. It begins with Section 1.2, the discussion of axiomatic approaches, which takes as a starting point features that one might reasonably desire of a pairwise comparison model. A number are very closely linked and might even be thought of as restatements of the same idea, but the intuitions behind them differ sufficiently, as evidenced by their separate appearances in the literature, such that they are presented separately here.

In Section 1.3, the selection of a rating model is cast in the familiar framework of a constrained optimisation. This also leads to a discussion of the Bradley-Terry model in the context of an exponential family of distributions, which provides a natural link to the axiomatic approaches of Section 1.2. Section 1.4 takes the context of Thurstone's discriminal processes, and discusses the distributions that lead to a Bradley-Terry model under this set-up, and how they might be motivated. In Section 1.5, it is noted how the Bradley-Terry model is apparent in other well-known statistical models, as a conditional form of Rasch, hazard and network models. In Section 1.6, some examples are introduced that derive from realistic game scenarios picking up on the highly intuitive nature of the model. In Section 1.7, the quasi-symmetry model is discussed, and is used to show how the often intuitive approaches

that underlie a number of other popular rating methods can be related to Bradley-Terry and produce consistent estimators for the Bradley-Terry strength parameters. This also leads to noting the link to Barker's algorithm, a popular Markov chain Monte Carlo method. In each subsection, the reference given in the title is that of the earliest work linking the approach explicitly to the Bradley-Terry model, and the subsections are ordered chronologically by these. The sections are ordered with statistical interest and chronology in mind. In the final section some short concluding remarks are made.

Throughout the chapter, p_{ij} will be the probability of i beating j or for a preference for i over j, where $i, j \in T$ and T is of size n. The $n \times n$ data matrix $C = [c_{ij}]$ will be the 'competition' matrix of preferences or wins, such that c_{ij} is the number of preferences for i over j. $M = C + C^T$ is defined as the symmetric matrix where m_{ij} is the number of comparisons, or 'matches' in sports parlance, between i and j. C is taken to be irreducible, that is, as described by Ford Jr (1957, p.29): "[I]n every possible partition of the objects into two non-empty subsets, some object in the second set has been preferred at least once to some object in the first set." This ensures that strength estimates are finite. It is not assumed that there are the same number of comparisons between any two items, nor indeed that the number of comparisons between any two items is non-zero. Where appropriate, the language of sports — contests, scores, teams — is used to aid in providing clear interpretability, though the motivations may often be analogised outside this context.

1.2 Axiomatic motivations

It is sometimes possible to fix properties that we would desire of a model and use them to derive a unique model. In this section we consider such properties that lead to the Bradley-Terry model.

1.2.1 Transitivity of odds (Good, 1955)

Consider four teams i, j, k, l. Suppose that the probability that j beats k is greater than the probability that j beats l,

$$p_{jk} > p_{jl},$$

then it is intuitive to think that the probability that i beats k will be greater than the probability that i beats l,

$$p_{ik} > p_{il}.$$

A simple way to enforce this would be by insisting on the transitivity of odds as Good (1955) proposes, that is

$$\frac{p_{ij}}{p_{ji}} \times \frac{p_{jk}}{p_{kj}} = \frac{p_{ik}}{p_{ki}}.$$

Alternatively one might think of the same condition in the manner that Luce and Suppes (1965) refers to it as the *product rule*, where for any triple (i, j, k) the probability of the intransitive cycle i beats j, j beats k, k beats i is the same as that of the intransitive cycle i beats k, k beats j, j beats i, expressed

$$p_{ij}p_{jk}p_{ki} = p_{ik}p_{kj}p_{ji} \quad \text{for all triplets } (i, j, k).$$

Strang et al. (2020) characterise this as an 'arbitrage free' condition and it is also known as Kolmogorov's criterion (Kolmogorov, 1936; Kelly, 1979).

Jech (1983, p. 249) provides a sketch justification for the principle that we adapt here. Suppose we wish to estimate the odds of an item i beating an item k. We estimate the odds of i beating k by the ratio of the number of times i beats k to the number of times k beats i. However, suppose we can compare objects i and k only indirectly by comparing i with j and j with k. We do this by determining that if i beats j and j beats k then we consider that i has beaten k. If i loses to j and j loses to k then we consider that k has beaten i. For other result combinations (i beats j and k beats j, or j beats i and j beats k) we reserve judgement. We repeat this operation M times and allow M to be very large

$$\frac{p_{ik}}{p_{ki}} = \lim_{M\to\infty} \frac{(\text{number of times } i \text{ beats } k)}{(\text{number of times } k \text{ beats } i)} = \frac{Mp_{ij}p_{jk}}{M(1-p_{ij})(1-p_{jk})} = \frac{p_{ij}}{p_{ji}}\frac{p_{jk}}{p_{kj}}$$

Jech (1983, p.246) claims that this leads to the "one and only one correct way of comparing the records of teams in an incomplete tournament", which seems a little bold, but it is nevertheless useful for understanding the appealing intuition behind the property.

Returning to how it leads to the Bradley-Terry model, it may alternatively be expressed as

$$\log \frac{p_{ij}}{p_{ji}} + \log \frac{p_{jk}}{p_{kj}} = \log \frac{p_{ik}}{p_{ki}}.$$

Letting $p_{ij}/p_{ji} = \exp(\tau(\theta_i, \theta_j))$, where θ_i can be thought of as a parameter summarising the strength of i, then

$$\tau(\theta_i, \theta_j) + \tau(\theta_j, \theta_k) = \tau(\theta_i, \theta_k).$$

Then setting $\theta_j = \theta_i$, it may be noted that $\tau(\theta_i, \theta_i) = 0$ for all i. By setting $\theta_k = \theta_i$ it may be noted that τ is an antisymmetric function. Further by differentiating with respect to θ_i it may be noted that the partial derivative of $\tau(\theta_i, \theta_j)$ with respect to θ_i is independent of θ_j, so that $\tau(\theta_i, \theta_j)$ is some function of θ_i alone plus some function of θ_j alone, and since τ is antisymmetric it must be of the form $t(\theta_i) - t(\theta_j)$.

Now $t(\theta_i)$ may be taken as an increasing continuous function of θ_i, and $\lambda_i = t(\theta_i)$ can be used as a parameter for the strength of i also, so that

$$\frac{p_{ij}}{p_{ji}} = \exp(\lambda_i - \lambda_j) \quad \text{for all } i, j,$$

giving the Bradley-Terry model.

1.2.2 Luce's Choice Axiom (Luce, 1959)

Let $p_S(i)$ be the probability that item i is chosen from a set $S \subseteq T$, then a complete system of choice probabilities satisfies Luce's Choice Axiom if and only if for every i and for $S \subseteq T$

$$p_S(i) = \frac{p_T(i)}{\sum_{k \in S} p_T(k)} \quad .$$

The choice axiom is a version of the decision theory axiom of the independence of irrelevant alternatives, the idea that a choice from S is independent of the other choices available in T. Luce (1959) introduces it with the assertion that many choice situations are characterised by a multistage process, whereby a subset of the total choice set is selected, from which further subsets are selected iteratively, until a single choice is made from one of these subsets. While it is noted that the final result is likely to depend on these intermediate categorisations for complex choices and a multistage process, for a simple decision and a two stage process, it is argued that the two-stage choice, reflected by the product $p_S(i) \sum_{k \in S} p_T(k)$, does not depend on S, and by setting $S = T$ it is apparent that this must be $p_T(i)$. The Choice Axiom has also been motivated by appealing to the decomposition of a full ranking model (Block and Marschak, 1960, Theorem 3.6), to invariance under uniform expansion of the choice set (Yellot, 1977), and under specific assumptions in a consideration of the utility of gambling (Luce et al., 2008).

A complete system satisfies the Choice Axiom if and only if there exist a set of numbers $\pi_1, \pi_2, \ldots \pi_n$ such that for every i and $S \subseteq T$

$$p_S(i) = \frac{\pi_i}{\sum_{k \in S} \pi_k} \quad .$$

In order to see this, let

$$\pi_i = \kappa p_T(i), \quad \kappa > 0,$$

then

$$\begin{aligned} p_S(i) &= \frac{p_T(i)}{\sum_{k\in S} p_T(k)} \\ &= \frac{\kappa p_T(i)}{\sum_{k\in S} \kappa p_T(k)} \\ &= \frac{\pi_i}{\sum_{k\in S} \pi_k}. \end{aligned}$$

π_i is unique up to a multiplicative constant since suppose there is another π_i' satisfying this condition, then

$$\pi_i = \kappa p_T(i) = \frac{\kappa \pi_i'}{\sum_{k\in T} \pi_k'},$$

and setting $\kappa' = \kappa / \sum_{k\in T} \pi_k'$ then $\pi = \kappa' \pi_i'$

Taking S to be the two member set $\{i, j\}$ gives the Bradley-Terry model.

1.2.3 Reciprocity (Block and Marschak, 1960)

What might be thought of as an alternative expression of the Choice Axiom is noted in Block and Marschak (1960, p.103). The idea is that the odds of i beating j should be equivalent to the ratio of strength parameters of i and j.

$$\frac{p_{ij}}{p_{ji}} = \frac{\pi_i}{\pi_j} \quad \text{for all } i, j \quad .$$

Of course this condition can be framed in other familiar equivalent terms, either as detailed balance, more typically expressed as

$$p_{ij}\pi_j = p_{ji}\pi_i \quad \text{for all } i, j,$$

or that the irreducible, positive recurrent, aperiodic Markov chain for which $P = [p_{ij}]$ is the transition matrix is reversible, which itself is the case if and only if the transitivity condition of Section 1.2.1 holds (Kelly, 1979). This condition leads immediately to

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j} \quad .$$

1.2.4 Points as a sufficient statistic (Bühlmann and Huber, 1963)

Suppose $w_i = \sum_j c_{ij}$ are the wins gained by team i and that the vector $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)^T$ is a sufficient statistic such that the likelihood is dependent on C only through $\boldsymbol{w}$.

Consider the comparison matrix $C = [c_{ij}]$ with c_{kl}, c_{lm}, c_{mk} non-zero, for the triplet (k, l, m) where without loss of generality $k < l < m$. Now consider an alternative C' with $c'_{kl} = c_{kl} - 1$, $c'_{lm} = c_{lm} - 1$, $c'_{mk} = c_{mk} - 1$ and $c'_{lk} = c_{lk} + 1$, $c'_{ml} = c_{ml} + 1$, $c'_{km} = c_{km} + 1$, and all else the same. Then the score vectors are identical and so if score is a sufficient statistic then the likelihoods must also be identical. The likelihood is

$$\prod_{i<j} \binom{m_{ij}}{c_{ij}} p_{ij}^{c_{ij}} (1 - p_{ij})^{m_{ij} - c_{ij}},$$

so that the log-likelihood, up to a constant term, is

$$\sum_{i<j} c_{ij} \log \left(\frac{p_{ij}}{1 - p_{ij}} \right) + m_{ij} \log(1 - p_{ij}).$$

Setting these equal for C and C', we get that

$$(c_{kl} - c'_{kl}) \log \frac{p_{kl}}{p_{lk}} + (c_{lm} - c'_{lm}) \log \frac{p_{lm}}{p_{ml}} + (c_{mk} - c'_{mk}) \log \frac{p_{mk}}{p_{km}} = 0,$$

and so

$$\log \frac{p_{kl}}{p_{lk}} + \log \frac{p_{lm}}{p_{ml}} + \log \frac{p_{mk}}{p_{km}} = 0,$$

by the specifications of $c'_{kl}, c'_{lm}, c'_{mk}$, giving the Bradley-Terry model following the same argument as in Section 1.2.1.

1.3 Objective function maximisation

It is a common procedure in quantitative analysis to identify an appropriate objective function and seek to maximise that function under certain plausible constraints. Indeed the familiarity of such procedures makes these motivations perhaps some of the most persuasive in the use of the Bradley-Terry model. In this section, four such maximisations are presented. There is also discussion of the model in the context of an exponential family of distributions, which provides a link between the entropy and likelihood maximising motivations of this section and the motivations presented in Section 1.2.

1.3.1 Maximum entropy with retrodictive criterion (Henery, 1986; Joe, 1988)

Consider an objective function $S(p)$, which is a function of the probabilities p_{ij}. We wish to maximise this objective function subject to some identified criterion. The proposed constraint is that of the 'retrodictive criterion', that the observed number of wins for each team is equal to the expected number of wins given the matches played. That is

$$\sum_{j \neq i} c_{ij} = \sum_{j \neq i} m_{ij} p_{ij} \qquad \text{for all teams } i.$$

A justification for this criterion was pithily expressed by Stob (1984): "What sort of a claim is it that a team solely on the basis of the results should have expected to win more games than they did?"[1]

Turning to the objective function, the approach of maximising entropy is common in statistical physics. Entropy is a measure of the uncertainty of a random variable. By maximising it, roughly speaking, the assumptions in the model are minimised. Jaynes (1957) influentially advocated for the choice of entropy in a broader range of statistical settings, building on the ideas from information theory of Shannon (1948). Good et al. (1963) provides further discussion noting "[t]he mere fact that the principle of maximum entropy generates classical statistical mechanics, as a null hypothesis, would be sufficient reason for examining its implications in mathematical statistics." Luce (1959), on the other hand, casts doubt on its applicability to choice contexts. In this setting, the entropy is defined as

$$S(p) = -\sum_{i \neq j} m_{ij} p_{ij} \log p_{ij} = -\sum_{j<i} m_{ij} (p_{ij} \log p_{ij} + (1 - p_{ij}) \log(1 - p_{ij})).$$

We consider maximising the entropy subject to the retrodictive criterion using the method of Lagrange multipliers

$$\mathcal{L}(p, \boldsymbol{\eta}) = S(p) - \sum_{i=1}^{n} \eta_i \Bigg(\sum_{j=1, j \neq i}^{n} (m_{ij} p_{ij} - c_{ij}) \Bigg),$$

and setting $\frac{\partial \mathcal{L}}{\partial p_{ij}} = 0$ for all p_{ij} in the normal way gives that

$$\frac{\partial S(p)}{\partial p_{ij}} = \frac{\partial}{\partial p_{ij}} \sum_{r=1}^{n} \eta_r \Bigg(\sum_{s=1, s \neq r}^{n} (m_{rs} p_{rs} - c_{rs}) \Bigg) \quad \text{for all } i, j.$$

[1] As will be discussed further in Chapter 4, this could be seen as failing to appreciate the bias present from finite observations. Nevertheless, it reflects the intuitive appeal of the condition.

So that for all i, j such that $m_{ij} \neq 0$,

$$-\log p_{ij} + \log(1 - p_{ij}) = \eta_i - \eta_j,$$

or equivalently

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j},$$

where $\pi_i = \exp(-\eta_i)$, and it can readily be checked by differentiating $S(p)$ that this is a maximum.

1.3.2 Maximum likelihood estimation with retrodictive criterion (Joe, 1988)

Suppose the probability of i being preferred to j is given by

$$p_{ij} = f(\lambda_i, \lambda_j),$$

where λ_i and λ_j are real-valued parameters describing the strength of items i and j, and $f : \mathbb{R} \times \mathbb{R} \to [0, 1]$. Then the likelihood function is given by

$$L(\boldsymbol{\lambda}) = \prod_{i<j} \binom{m_{ij}}{c_{ij}} p_{ij}^{c_{ij}} (1 - p_{ij})^{m_{ij} - c_{ij}} = \prod_{i<j} \binom{m_{ij}}{c_{ij}} p_{ij}^{c_{ij}} p_{ji}^{c_{ji}},$$

and the log-likelihood function, ignoring the constant term, is

$$l(\boldsymbol{\lambda}) = \sum_{i<j} c_{ij} \log(p_{ij}) + c_{ji} \log(p_{ji}).$$

The log-likelihood may be maximised under the constraint of the retrodictive criterion, that the number of wins for each team is equal to the expected number of wins given the matches played,

$$\sum_j c_{ij} = \sum_j m_{ij} p_{ij} \qquad \text{for all teams } i.$$

At an extreme point of the log-likelihood, for all k,

$$0 = \frac{\partial}{\partial \lambda_k} l(\boldsymbol{\lambda}) = \sum_j c_{kj} \frac{\partial}{\partial \lambda_k} \log(p_{kj}) + c_{jk} \frac{\partial}{\partial \lambda_k} \log(p_{jk}).$$

Considering the constraint we note that

$$\begin{aligned} 0 = \sum_j c_{kj} - m_{kj}p_{kj} &= \sum_j c_{kj} - (c_{kj} + c_{jk})p_{kj} \\ &= \sum_j c_{kj}(1 - p_{kj}) - c_{jk}p_{kj} \\ &= \sum_j c_{kj}(1 - p_{kj}) - c_{jk}(1 - p_{jk}), \end{aligned}$$

and so there is an extreme point where

$$\begin{aligned} \frac{\partial}{\partial \lambda_k} \log(p_{kj}) &= (1 - p_{kj}) \\ \frac{\partial}{\partial \lambda_k} \log(p_{jk}) &= -(1 - p_{jk}), \end{aligned}$$

which gives

$$\begin{aligned} \frac{\partial p_{kj}}{\partial \lambda_k} &= p_{kj}(1 - p_{kj}) \\ \frac{\partial p_{jk}}{\partial \lambda_k} &= -p_{jk}(1 - p_{jk}). \end{aligned}$$

Solving these separable differential equations for p_{ij} gives

$$\begin{aligned} p_{ij} &= \frac{1}{1 + e^{-(\lambda_i - \lambda_j)}} \\ &= \frac{\pi_i}{\pi_i + \pi_j} \end{aligned}$$

where $\pi_i = e^{\lambda_i}$, and as before this is a maximum since the log-likelihood is strictly concave. So that the Bradley-Terry model is the likelihood maximising model.

Consideration of the maximum likelihood and maximum entropy motivations in the context of an exponential family of distributions (Pitman, 1936; Koopman, 1936; Darmois, 1935) provides a link to the motivations of Section 1.2. Following Geyer (2020), a statistical model is an exponential family of distributions if it has a log-likelihood of the form

$$l(\theta) = \langle y, \theta \rangle - k(\theta),$$

where y is a vector-valued canonical statistic; θ is a vector-valued canonical parameter; $\langle ., . \rangle$ represents an inner product; and k is a real-valued function, the cumulant

function, which is defined such that $\nabla k(\theta) = \mathbb{E}_\theta(Y)$. In seeking a maximum likelihood estimate, the derivative is taken and set equal to zero

$$0 = \nabla l(\theta) = y - \nabla k(\theta) = y - \mathbb{E}_\theta(Y),$$

by the definition of the cumulant function within an exponential family.

In the model discussed here the likelihood is

$$\prod_{i<j} \binom{m_{ij}}{c_{ij}} p_{ij}^{c_{ij}} (1 - p_{ij})^{m_{ij} - c_{ij}},$$

so that the log-likelihood, up to a constant term, may be taken to be

$$\frac{1}{2} \sum_{i,j} c_{ij} \log \left(\frac{p_{ij}}{1 - p_{ij}} \right) + m_{ij} \log(1 - p_{ij}),$$

and may be rewritten in the form

$$l(\theta) = \frac{1}{2} \sum_{i,j} c_{ij} \theta_{ij} - m_{ij} \log(1 + e^{\theta_{ij}}),$$

where θ is the canonical parameter, a vector of length $n(n-1)$ corresponding to the directed pairwise comparisons, and with $\theta_{ij} = \log(p_{ij}/(1 - p_{ij}))$; the canonical statistic vector y takes scaled outcomes $c_{ij}/2$ as its elements; and the cumulant function is $k(\theta) = \sum_{i,j} m_{ij} \log(1 + e^{\theta_{ij}})/2$.

What Geyer et al. (2007) refer to as an *affine canonical submodel* may be parametrised through the linear transformation

$$\theta = a + X\beta,$$

where a is an offset vector, X is a design matrix, and β is the canonical parameter for the submodel, giving a log-likelihood of

$$l(\beta) = \langle X^T y, \beta \rangle - k_{SUB}(\beta),$$

where $k_{SUB}(\beta) = k(a + X\beta)$, so that this defines a new exponential family with canonical statistic vector $X^T y$, canonical parameter vector β, and cumulant function k_{SUB}.

In the context of the Bradley-Terry model one may take $a = 0, \beta = \lambda$, where λ is the vector of log-strengths $\lambda_i = \log \pi_i$, and X to be the design matrix with the columns representing the n participants, and the rows representing the $n(n-1)$

directed pairwise comparisons. The entry in the row corresponding to a preference for i over j has 1 in column i, -1 in column j and zero elsewhere. This gives a log-likelihood

$$\begin{aligned} l(\lambda) &= \frac{1}{2}\sum_{i,j}(c_{ij} - c_{ji})\lambda_i - \frac{1}{2}\sum_{i,j} m_{ij}\log(1 + e^{\lambda_i - \lambda_j}) \\ &= \frac{1}{2}\sum_{i,j}(2c_{ij} - m_{ij})\lambda_i - \frac{1}{2}\sum_{i,j} m_{ij}\log(1 + e^{\lambda_i - \lambda_j}) \\ &= \sum_{i,j} c_{ij}\lambda_i - \frac{1}{2}\sum_{i,j} m_{ij}(\lambda_i + \log(1 + e^{\lambda_i - \lambda_j})). \end{aligned}$$

Using the same notation as before, where a vector of wins $\boldsymbol{w}$ is defined by $w_i = \sum_j c_{ij}$, then

$$l(\lambda) = \sum_i w_i\lambda_i - \frac{1}{2}\sum_{i,j} m_{ij}(\lambda_i + \log(1 + e^{\lambda_i - \lambda_j})),$$

defining an exponential family where the score is the vector-valued canonical statistic and log-strength the vector-valued canonical parameter. It is a feature of an exponential family of distributions that 'observed equals expected', or more precisely that the observed value of the canonical statistic vector equals its expected value under the MLE distribution, that is to say

$$y = \mathbb{E}_{\hat{\theta}}(Y) = \nabla k(\hat{\theta}),$$

which under this affine canonical submodel translates to

$$\begin{aligned} w_k &= \frac{1}{2}\sum_j m_{kj}\left(1 + \frac{e^{\lambda_k - \lambda_j}}{1 + e^{\lambda_k - \lambda_j}}\right) - \frac{1}{2}\sum_i m_{ik}\frac{e^{\lambda_i - \lambda_k}}{1 + e^{\lambda_i - \lambda_k}} \\ &= \sum_j m_{kj}\frac{e^{\lambda_k}}{e^{\lambda_k} + e^{\lambda_j}} \quad \text{for all } k, \end{aligned}$$

noting that $p_{kj} = e^{\lambda_k}/(e^{\lambda_k} + e^{\lambda_j})$ gives what was referred to as the retrodictive criterion in Sections 1.3.1 and 1.3.2.

The motivations based on score as a sufficient statistic, maximum entropy and maximum likelihood of Sections 1.2.4, 1.3.1, and 1.3.2 may thus be seen as an example of a general fact about exponential families. If one starts with a canonical statistic, then the corresponding affine submodel, if it exists, will be uniquely determined

and it will be the maximum entropy and maximum likelihood model subject to the 'observed equals expected' constraint on the canonical statistic. As shown in Section 1.2.4, the requirement to take score as a sufficient statistic leads directly to the same statistical condition as the other axiomatic motivations presented in Section 1.2. A consideration of the Bradley-Terry model as an exponential family of distributions therefore gives a synthesis to the axiomatic and objective function motivations.

1.3.3 Simplicity 1

Often when selecting a model, transparency and interpretability are desirable features. This may be especially so in contexts where the outcomes affect a wide group of stakeholders. These sort of contexts are not uncommon in pairwise comparison with the methods being used to perform activities like ranking sports teams (Firth, 2022) or in educational assessment (Pollitt, 2012b). Therefore, there may be a legitimate desire for simpler, more intuitive models. It is thus appealing to consider how one might select a model with the goal of maximising simplicity.

Suppose one wished to determine a ranking by defining a probability for the preference for i over j related only to positive real-valued strength parameters π_i and π_j respectively,

$$p_{ij} = f(\pi_i, \pi_j).$$

A reasonable set of criteria for this function would be:

1. $f : \mathbb{R}^+ \times \mathbb{R}^+ \to [0, 1]$,
2. $f(\pi_i, \pi_j) = \frac{1}{2}$ when $\pi_i = \pi_j$,
3. $\lim_{\pi_i \to 0, \pi_j \text{ fixed}} f(\pi_i, \pi_j) = 0$,
4. $\lim_{\pi_j \to 0, \pi_i \text{ fixed}} f(\pi_i, \pi_j) = 1$,
5. $\lim_{\pi_i \to \infty, \pi_j \text{ fixed}} f(\pi_i, \pi_j) = 1$,
6. $\lim_{\pi_j \to \infty, \pi_i \text{ fixed}} f(\pi_i, \pi_j) = 0$.

where $\mathbb{R}^+$ is taken to be the set of positive real numbers not including zero.

Assume that the simplest set of functions are those that may be defined solely using the four basic operators $(+, -, \times, \div)$, and that any measure of the simplicity of a function is a strictly decreasing function of the number of these operators used. So that maximising simplicity is equivalent to minimising the number of basic operators. Bracketing anywhere, used in the conventional sense, to identify a functional subclause, is allowed without increasing or reducing simplicity. Constants are

also allowed in place of parameters without increasing or reducing simplicity. In the language of Computer Science, this is therefore defining simplicity by the minimum number of floating point operations (flops).

No f with exactly zero or one operator can meet criterion 5 other than $f(\pi_i, \pi_j) = 1$ or equivalents (for example, $f(\pi_i, \pi_j) = \pi_i/\pi_i$), which violates criteria 2, 3 and 6. Likewise, considering a function with two operators and again considering criterion 5, then it must be that the operator $\div$ is employed as otherwise the limit of criterion 5 would be infinite in absolute value other than in cases which are equivalent to a constant (for example, $f(\pi_i, \pi_j) = \pi_i + (1 - \pi_i)$) or where π_i is not included, but if π_i is not included then criteria 2 and 3 will be in contradiction. So if there is a solution with exactly two operators then it must be of the form $f(\pi_i, \pi_j) = g(\pi_i, \pi_j) \div h(\pi_i, \pi_j)$ where either g or h is equal to either one of the parameters or to a constant in order that only two operators are used, and the other must be a single operator function involving $+$ or $-$ in order to meet criterion 5 without being equivalent to a constant (for example, $f(\pi_i, \pi_j) = \pi_i \div (c \times \pi_i)$). From criterion 3 it must be that $g(\pi_i, \pi_j) = \pi_i$ and then from criterion 5, h must take π_i as one of its terms. Criterion 6 implies that the other term in h is π_j and criterion 2 then implies that $h(\pi_i, \pi_j) = \pi_i + \pi_j$. This gives $f(\pi_i, \pi_j) = \pi_i \div (\pi_i + \pi_j)$, which meets all the required criteria. It may be noted that not all the criteria were required for its unique derivation, and that other subsets of the criteria may be used to derive the same result. That is to say that

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j}$$

will be the unique simplicity maximiser under a number of different subsets of the plausible criteria.

1.3.4 Simplicity 2

Given positive-valued strength parameters π_i and π_j for i and j respectively, one may want to consider a model where the probability of i being preferred to j is a function f of the ratio $x_{ij} = \pi_i/\pi_j$,

$$p_{ij} = f(x_{ij}).$$

A reasonable set of criteria for this function would then be:

1. $f : \mathbb{R}^+ \to [0, 1]$,
2. $f(1) = \frac{1}{2}$,

3. $\lim_{x\to 0} f(x) = 0$,

4. $\lim_{x\to\infty} f(x) = 1$,

Proceeding in a similar fashion to the previous section, the only function including exactly zero or one flop that meets criterion 4 is $f(x) = 1$ (or equivalents, for example, $f(x) = x \div x$), but this violates criteria 2 and 3. Considering a function with two operators and again considering criterion 4, then it must be that the operator $\div$ is employed as otherwise the limit would be infinite in absolute value other than in cases which are equivalent to a constant (for example, $f(x) = x + (1 - x)$). So if there is a solution with exactly two operators then it must be of the form $f(x) = g(x) \div h(x)$ where either $g(x) = x$ or $h(x) = x$ or $g(x) =$ constant or $h(x) =$ constant in order that only two operators are used, and the other must be a single-operator function involving $+$ or $-$ in order to meet criterion 4. Criterion 3 implies that $g(x) = x$, and criterion 2 then tells us that $h(x) = 1 + x$. Thus

$$f(x) = \frac{x}{1+x},$$

giving

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j}.$$

1.4 Discriminal processes

Consider a scenario where the strength of each of two entities in a given pairwise interaction is observed with error. Denote the observed strength of i as b_i with 'true' strength λ_i, so that $b_i = \lambda_i + \epsilon_i$, where ϵ_i is an error term. Item i is preferred to item j if and only if $b_i > b_j$. This is the model of Thurstone's 'discriminal processes' (Thurstone, 1927a). Taking the error to be Gaussian, as Thurstone himself did, leads to what is commonly known as the Thurstone-Mosteller model (Thurstone, 1927a; Mosteller, 1951), but the set up may also be used to motivate the Bradley-Terry model by considering alternative distributions for b_i.

1.4.1 Exponential Distribution (Holman and Marley as cited by Luce and Suppes (1965, p.338))

Suppose b_i follows an exponential distribution

$$\mathbb{P}(b_i \leq x) = F_i(x) = 1 - \mathrm{e}^{-\frac{x}{\pi_i}}, \quad x \in \mathbb{R}^+.$$

Then the probability that i is preferred to j in a pairwise comparison is

$$\begin{aligned}
p_{ij} &= \int_0^\infty F_j(x) F_i'(x) \mathrm{d}x \\
&= \int_0^\infty \left(1 - \mathrm{e}^{\frac{x}{\pi_j}}\right) \frac{1}{\pi_i} \mathrm{e}^{-\frac{x}{\pi_i}} \mathrm{d}x \\
&= 1 - \frac{1}{\pi_i \left(\frac{1}{\pi_i} + \frac{1}{\pi_j}\right)} \int_0^\infty \left(\frac{1}{\pi_i} + \frac{1}{\pi_j}\right) \mathrm{e}^{-\left(\frac{1}{\pi_i} + \frac{1}{\pi_j}\right)x} \mathrm{d}x \\
&= 1 - \frac{\pi_j}{\pi_i + \pi_j} \\
&= \frac{\pi_i}{\pi_i + \pi_j} \quad .
\end{aligned}$$

1.4.2 Extreme value distributions (Bradley, 1965; Thompson and Singh, 1967)

Thompson and Singh (1967) provide a rationale for a broader class of distributions that lead to a Bradley-Terry model under a discriminal process. Based on ideas from Psychology, sensations are hypothesised to be a result of a large number of stimuli. These stimuli are modeled as having independent identical distributions $G(x)$. One might then consider the distribution of the resultant sensation.

Two intuitive possibilities would be to model the distribution of the sensation $F(x)$ either as the average of those stimuli or the maximum of those stimuli. Taking the average gives a normal distribution for $F(x)$ and leads to a Thurstone-Mosteller comparison model. Taking the maximum of the stimuli, gives, by extreme value theorem (Fisher and Tippett, 1928; Gnedenko, 1943; Gumbel, 1958), one of three distributions for $F(x)$ — Gumbel, Weibull, or Frechet — depending on the underlying stimuli distribution $G(x)$, and leads to a Bradley-Terry comparison model. The Gumbel is the most notable of these, being the sensation distribution for stimuli distributions such as the normal, lognormal, logistic, and exponential.

While Thompson and Singh (1967) provided a clear motivation for considering such models and do not assume that the underlying stimuli distributions need have the same location parameters for i and j, Lehmann (1953) had previously considered a family of distributions in the context of the power of rank tests of the form $F_{X_i}(x; \pi_i) = G^{\pi_i}(x)$, where $G(x)$ is itself a distribution function. Bradley (1965) discussed this family of distributions with respect to the Bradley-Terry model. As Bradley (1976) notes, if $G(x)$ is a distribution function, and X_i is the random variable

relating to a sensation i, with distribution function

$$\mathbb{P}(X_i \leqslant x) = G^{\pi_i}(x),$$

where $\pi_i > 0$, then comparing sensations i and j,

$$p_{ij} = \mathbb{P}(X_i > X_j) = \int_{x_i > x_j} dG^{\pi_i}(x_i) dG^{\pi_j}(x_j) = \frac{\pi_i}{\pi_i + \pi_j}, \quad i \neq j.$$

Gumbel distribution (Thompson and Singh, 1967)

Suppose b_i follows a Gumbel distribution with mean λ_i. Then

$$\Pr(b_i \leq x) = F_i(x) = \exp\left(-\pi_i \mathrm{e}^{-\alpha x}\right) \text{ for } x \in \mathbb{R} \text{ and parameter } \alpha > 0,$$

where $\pi_i = \mathrm{e}^{\alpha\lambda_i - \gamma}$, with γ the Euler-Mascheroni constant. Then the probability that i is preferred to j in a pairwise comparison is

$$\begin{aligned} p_{ij} &= \int_{-\infty}^{\infty} F_j(x) F_i'(x) \mathrm{d}x \\ &= \int_{-\infty}^{\infty} \exp\left(-\pi_j \mathrm{e}^{-\alpha x}\right) \alpha \pi_i \exp\left(-\alpha x - \pi_i \mathrm{e}^{-\alpha x}\right) \mathrm{d}x \\ &= \frac{\pi_i}{\pi_i + \pi_j} \int_{-\infty}^{\infty} \alpha(\pi_i + \pi_j) \exp\left(-\alpha x - (\pi_i + \pi_j)\mathrm{e}^{-\alpha x}\right) \mathrm{d}x \\ &= \frac{\pi_i}{\pi_i + \pi_j} \quad . \end{aligned}$$

Weibull distribution (Thompson and Singh, 1967)

Suppose b_i follows a Weibull distribution

$$\mathbb{P}(b_i \leq x) = F_i(x) = 1 - \exp\left(-(x/\lambda_i)^{\alpha}\right) \text{ for } x \in \mathbb{R}^+ \text{ and parameter } \alpha > 0.$$

Then the probability that i is preferred to j in a pairwise comparison is

$$\begin{aligned} p_{ij} &= \int_0^\infty F_j(x)F_i'(x)\mathrm{d}x \\ &= \int_0^\infty [1-\exp\left(-(x/\lambda_j)^\alpha\right)]\frac{\alpha}{\lambda_i}(x/\lambda_i)^{\alpha-1}\exp\left(-(x/\lambda_i)^\alpha\right)\mathrm{d}x \\ &= 1-\int_0^\infty \frac{\alpha}{\lambda_i}(x/\lambda_i)^{\alpha-1}\exp\left(-(x/\lambda_j)^\alpha\right)-(x/\lambda_i)^\alpha)\mathrm{d}x \\ &= 1-\frac{\lambda_j^\alpha}{\lambda_i^\alpha+\lambda_j^\alpha}\int_0^\infty \frac{\alpha}{\lambda_i\lambda_j}(x/\lambda_i\lambda_j)^{\alpha-1}(\lambda_i^\alpha+\lambda_j^\alpha)\exp\left(-(x/\lambda_i\lambda_j)^\alpha(\lambda_i^\alpha+\lambda_j^\alpha)\right)\mathrm{d}x \\ &= \frac{\pi_i}{\pi_i+\pi_j} \quad , \end{aligned}$$

where $\pi_i = \lambda_i^\alpha$.

Fréchet distribution (Thompson and Singh, 1967)

Suppose b_i follows a Frechet distribution

$$\mathbb{P}(b_i \leq x) = F_i(x) = \exp\left(-\pi_i x^{-\alpha}\right) \text{for } x \in \mathbb{R}^+ \text{ and parameter } \alpha > 0.$$

Then the probability that i is preferred to j in a pairwise comparison is

$$\begin{aligned} p_{ij} &= \int_0^\infty F_j(x)F_i'(x)\mathrm{d}x \\ &= \int_0^\infty \exp\left(-\pi_j x^{-\alpha}\right)\frac{\pi_i\alpha}{x^{\alpha+1}}\exp\left(-\pi_i x^{-\alpha}\right)\mathrm{d}x \\ &= \frac{\pi_i}{\pi_i+\pi_j}\int_0^\infty \alpha\frac{\pi_i+\pi_j}{x^{\alpha+1}}\exp\left(-(\pi_i+\pi_j)x^{-\alpha}\right)\mathrm{d}x \\ &= \frac{\pi_i}{\pi_i+\pi_j} \quad . \end{aligned}$$

1.5 Standard models

A number of models familiar to statisticians may be related to the Bradley-Terry model by considering conditional forms. Here we discuss three well-known models where that is the case.

1.5.1 Rasch model (Andrich, 1978)

Let X_{vi} be a binary random variable, representing the outcome of a test v taken by candidate i, where $X_{vi} = 1$ represents passing the test, and $X_{vi} = 0$ denotes failure. Under the Rasch simple logistic model (Rasch, 1960, 1961) the probability of the outcome $X_{vi} = 1$ is taken to be

$$\mathbb{P}(X_{vi} = 1) = \frac{e^{\lambda_i - \delta_v}}{1 + e^{\lambda_i - \delta_v}},$$

where λ_i represents the ability of candidate i and δ_v the difficulty of test v.

There are two conceptualisations by which we might derive the Bradley-Terry model from this. First, as Andrich (1978) notes, if we take

$$p_{ij} = \mathbb{P}(i \text{ passes a test } v \mid \text{exactly one of } i \text{ and } j \text{ pass the test } v),$$

then since

$$\mathbb{P}(X_{vi} = 1, X_{vj} = 0) = \frac{e^{\lambda_i - \delta_v}}{(1 + e^{\lambda_i - \delta_v})(1 + e^{\lambda_j - \delta_v})},$$

and

$$\mathbb{P}(X_{vi} + X_{vj} = 1) = \frac{e^{\lambda_i - \delta_v} + e^{\lambda_j - \delta_v}}{(1 + e^{\lambda_i - \delta_v})(1 + e^{\lambda_j - \delta_v})}$$

then conditional on being able to discern that one of the test-takers has performed better based on the binary test outcome and taking their test outcomes to be independent conditional on their abilities and the test difficulty then the probability that i has beaten j is

$$p_{ij} = \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}} = \frac{\pi_i}{\pi_i + \pi_j},$$

where $\pi_i = e^{\lambda_i}$.

Second, we might more directly consider that in comparing i with j we are setting a test for i of difficulty equal to the strength of the comparator λ_j.

1.5.2 Cox proprtional hazards model (Su and Zhou, 2006)

Consider a proportional hazards model (Cox, 1972) on random variables T_i with

$$h_i(t) = h(t)\pi_i$$

then

$$\begin{aligned}
\mathbb{P}(T_i < T_j) &= \int_0^\infty F_{T_i}(t) f_{T_j}(t)\, dt \\
&= \int_0^\infty \left(1 - \exp\{-\int_0^t h(x)\pi_i\, dx\}\right) h(t)\pi_j \exp\{-\int_0^t h(x)\pi_j\, dx\}\, dt \\
&= 1 - \int_0^\infty h(t)\pi_j \exp\{-(\pi_i + \pi_j)\int_0^t h(x)\, dx\}\, dt \\
&= 1 - \frac{\pi_j}{\pi_i + \pi_j} \\
&= \frac{\pi_i}{\pi_i + \pi_j}
\end{aligned}$$

Further, as Su and Zhou (2006) note, if a stratified proportional hazards model is used such that each stratum represents a different match with

$$h_i(t) = h_{s_{ij}}(t)\pi_i,$$

where s_{ij} is the stratum for a match between i and j then the contribution to the partial likelihood from the random variables T_i and T_j with the event $\{T_i < T_j\}$ is $\pi_i/(\pi_i + \pi_j)$.

1.5.3 Network models

Consider a binary directed network Y, with an edge $i \to j$ taking the value y_{ij}. A common class of models in network analysis takes a conditional independence approach, assuming that the value of any directed edge is independent of all other edge values given an appropriate set of parameters. In a generalised form for the current purposes it can be expressed as

$$\begin{aligned}
\mu_{ij} &= \mathbb{P}(y_{ij} = 1) \\
\text{logit}(\mu_{ij}; \delta_i, \gamma_j, f_{ij}) &= \delta_i + \gamma_j + f_{ij},
\end{aligned}$$

where δ_i and γ_j, sometimes referred to as *sociality* and *attractivity* parameters (Krivitsky et al., 2009), reflect the heterogeneity of out-degree and in-degree respectively, and $f_{ij} = f(i, j)$ is a symmetric function capturing the propensity for an edge in either direction to exist. For example, Hoff et al. (2002) takes $f(i, j)$ to be the Euclidean distance between points associated with i and j in a latent space but note that $f(i, j)$ could be any distance measure satisfying the triangle inequality

$f(i,j) \leqslant f(i,k) + f(k,j)$. Often models also incorporate a term of the form $\beta^T x_{ij}$ within $f(i,j)$, where x_{ij} is a vector of pair specific characteristics, to capture known homophilies.

Applying the conditional independence assumption

$$\mathbb{P}(y_{ij} = 1, y_{ji} = 0; \delta_i, \delta_j, \gamma_i, \gamma_j, f_{ij}) = \frac{e^{\delta_i+\gamma_j+f_{ij}}}{(1+e^{\delta_i+\gamma_j+f_{ij}})(1+e^{\delta_j+\gamma_i+f_{ij}})},$$

and

$$\begin{aligned}
\mathbb{P}(y_{ij} = 1 \mid y_{ij} + y_{ji} = 1; \delta_i, \delta_j, \gamma_i, \gamma_j, f_{ij}) &= \frac{e^{\delta_i+\gamma_j+f_{ij}}}{e^{\delta_i+\gamma_j+f_{ij}} + e^{\delta_j+\gamma_i+f_{ij}}} \\
&= \frac{e^{\delta_i-\gamma_i}}{e^{\delta_i-\gamma_i} + e^{\delta_j-\gamma_j}} \\
&= \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}} \\
&= \frac{\pi_i}{\pi_i + \pi_j},
\end{aligned}$$

where $\pi_i = e^{\lambda_i}$ and $\lambda_i = \delta_i - \gamma_i$. If Y is considered as a tournament matrix with a directed edge $i \to j$ indicating i beats j, then *sociality* is a team's propensity for winning and *attractivity* the propensity for losing so that assessing the strength of a team as the difference between these is readily intuitive.

1.6 Game scenarios

The Bradley-Terry model has frequently been associated with an analysis of sport. So it is perhaps not surprising that there are a number of game scenarios in which the model may be very naturally motivated. Some of these are presented here.

1.6.1 Poisson scoring (Audley, 1960; Stern, 1990)

Consider two teams i and j who score according to independent Poisson random variables X_i and X_j with rate parameters π_i and π_j respectively. The winner is the

first team to score. Then

$$\begin{aligned} p_{ij} &= \mathbb{P}(X_i = 1 \mid X_i + X_j = 1) \\ &= \frac{\mathbb{P}(X_i = 1)\mathbb{P}(X_j = 0)}{\mathbb{P}(X_i + X_j = 1)} \\ &= \frac{e^{-\pi_i}\pi_i e^{-\pi_j}}{e^{-(\pi_i+\pi_j)}(\pi_i + \pi_j)} \\ &= \frac{\pi_i}{\pi_i + \pi_j} \quad . \end{aligned}$$

Audley (1960) presents an argument for this based in the psychological literature, considering the probability of one response occurring before another, where the probability of a response occurring in any given small time interval is determined by a response-specific parameter. While the argument is presented in terms of discrete time, it notes that the continuous alternative would be to consider Poisson distributions. Stern (1990) notes that the context may be widened to that of two gamma random variables with the same shape parameter and different scale parameters, showing that taking a shape parameter of one returns the Bradley-Terry model, whereas allowing it to tend to infinity sees the model tend to the Thurstone-Mosteller model. The idea might also be considered in the context of the discriminal process on exponential distributions of Section 1.4.1, since the interarrival time of a homogeneous Poisson process with rate parameter λ has an exponential distribution with a mean $1/\lambda$. More directly it is simply an expression of the standard equivalence between a multinomial distribution, in this case Bernoulli, and independent Poisson distributions conditional on their total, sometimes referred to as the "Poisson trick" (Baker, 1994).

1.6.2 Sudden death (Stirzaker, 1999; Vojnović, 2015)

Consider two teams i and j involved in a 'sudden death' shoot-out. They play a game where in each round they succeed with independent probabilities p_i and p_j respectively. The winner is the team who first has more successes than the other team. Let $(i \succ j)_n$ be the event that i wins the 'sudden death' contest in round n.

Then

$$\begin{aligned}
p_{ij} &= \sum_{n=1}^{\infty} \mathbb{P}[(i \succ j)_n] \\
&= \sum_{n=1}^{\infty} \sum_{k=0}^{n-1} p_i(1-p_j) \binom{n-1}{k} (p_i p_j)^k ((1-p_i)(1-p_j))^{n-k-1} \\
&= p_i(1-p_j) \sum_{m=0}^{\infty} \sum_{k=0}^{m} \binom{m}{k} (p_i p_j)^k ((1-p_i)(1-p_j))^{m-k} \\
&= p_i(1-p_j) \sum_{m=0}^{\infty} (p_i p_j + (1-p_i)(1-p_j))^m \\
&= p_i(1-p_j) \sum_{m=0}^{\infty} (2p_i p_j - p_i - p_j + 1)^m \\
&= \frac{p_i(1-p_j)}{p_i + p_j - 2p_i p_j} \\
&= \frac{p_i(1-p_j)}{p_i(1-p_j) + p_j(1-p_i)} \\
&= \frac{\frac{p_i}{1-p_i}}{\frac{p_i}{1-p_i} + \frac{p_j}{1-p_j}} \\
&= \frac{\pi_i}{\pi_i + \pi_j},
\end{aligned}$$

where $\pi_i = \frac{p_i}{1-p_i}$.

Further suppose there is an alternative sudden death contest but now the winner is the team that is the first to have r more successes than the opposition. Let A_i be the event that i wins and A^{r+k} be the event that a result, either i or j winning, occurs

after the winning team has scored exactly $r+k$ times, then defining $q_i = p_i/(1-p_i)$,

$$\begin{aligned} p_{ij} = \mathbb{P}(A_i) &= \sum_{k=0}^{\infty} \mathbb{P}(A_i|A^{r+k})\mathbb{P}(A^{r+k}) \\ &= \sum_{k=0}^{\infty} \frac{q_i^{r+k} q_j^k}{q_i^{r+k} q_j^k + q_i^k q_j^{r+k}} P(A^{r+k}) \\ &= \frac{q_i^r}{q_i^r + q_j^r} \sum_{k=0}^{\infty} P(A^{r+k}) \\ &= \frac{q_i^r}{q_i^r + q_j^r} \\ &= \frac{\pi_i}{\pi_i + \pi_j}, \end{aligned}$$

where $\pi_i = q_i^r$.

1.6.3 Accumulated win ratio (Vojnović, 2015)

Take a sequence of matches between two players, i and j, where the probability that team i wins is proportional to the accumulated number of wins in previous matches. Suppose that the probability that i wins the first match is $\pi_i/(\pi_i + \pi_j)$. Then consider the probability that i will win the nth match. The claim is that this is $\pi_i/(\pi_i + \pi_j)$. We proceed to show this by induction. Define notation $(i \succ j)_n$ as meaning i beats j in match n then

$$\mathbb{P}[(i \succ j)_1] = \frac{\pi_i}{\pi_i + \pi_j}.$$

Now assume that

$$\mathbb{P}[(i \succ j)_k] = \frac{\pi_i}{\pi_i + \pi_j}.$$

Then proceeding by induction

$$\begin{aligned} \mathbb{P}[(i \succ j)_{k+1}] &= \mathbb{P}[(i \succ j)_{k+1} \mid (i \succ j)_k]\mathbb{P}[(i \succ j)_k] \\ &\quad + \mathbb{P}[(i \succ j)_{k+1} \mid (j \succ i)_k]\mathbb{P}[(j \succ i)_k] \\ &= \frac{\pi_i + 1}{\pi_i + 1 + \pi_j} \frac{\pi_i}{\pi_i + \pi_j} + \frac{\pi_i}{\pi_i + 1 + \pi_j} \frac{\pi_j}{\pi_i + \pi_j} \\ &= \frac{\pi_i(\pi_i + 1 + \pi_j)}{(\pi_i + 1 + \pi_j)(\pi_i + \pi_j)} \\ &= \frac{\pi_i}{\pi_i + \pi_j} \quad . \end{aligned}$$

1.6.4 Continuous time state transition (Brown, 2018)

Consider a match where the winner is the team winning at the end of a defined period of play. We choose to model the continuous state of 'winning' by a continuous time Markov chain on a binary state space $I = \{i \text{ winning}, j \text{ winning}\}$, Let the rate at which there is a switch from the state 'i winning' to the state 'j winning' be denoted by π_j, and the rate at which the switch from the state 'j winning' to the state 'i winning' be denoted by π_i. Then the intensity matrix is

$$\boldsymbol{Q} = \begin{pmatrix} -\pi_j & \pi_j \\ \pi_i & -\pi_i, \end{pmatrix}$$

and the equilibrium distribution vector of this process $\boldsymbol{p}$ is such that

$$\boldsymbol{pQ} = \boldsymbol{0},$$

and in this case is given by the probability vector $\boldsymbol{p} = \left(\frac{\pi_i}{\pi_i+\pi_j}, \frac{\pi_j}{\pi_i+\pi_j}\right)$.

Assuming that we are likely to see a large number of state changes during the course of the match, or that the probability of the initial state being 'i winning' is approximately $\pi_i/\pi_i + \pi_j$ then the probability that i beats j may be approximated by

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j}.$$

1.7 Quasi-symmetry and consistent estimators

The quasi-symmetry model was proposed by Caussinus (1965). A matrix C is quasi-symmetric if it can be decomposed such that

$$c_{ij} = \alpha_i \beta_j \gamma_{ij},$$

where $\gamma_{ij} = \gamma_{ji}$. The form of this can be simplified by taking $a_i = \alpha_i/\beta_i$ and $s_{ij} = \beta_i \beta_j \gamma_{ij}$, so that

$$c_{ij} = a_i s_{ij},$$

or in matrix form

$$C = AS,$$

where A is a diagonal matrix and S is symmetric. Informally one might think of the symmetric matrix representing the intensity of interactions, and the diagonal matrix as the relative ratings. Asymptotically by the Law of Large Numbers under

a Bradley-Terry data generating process we would expect the results matrix to be quasi-symmetric, since

$$\mathbb{E}[c_{ij}] = p_{ij}m_{ij} = \frac{\pi_i}{\pi_i + \pi_j}m_{ij} = a_{ii}s_{ij},$$

where $s_{ij} = m_{ij}/(\pi_i + \pi_j) = s_{ji}$ and $\pi_i = a_{ii}$. So, rating methods that accord with Bradley-Terry in the case of a quasi-symmetric results matrix are consistent estimators for the Bradley-Terry model given a Bradley-Terry data generating process, and thus motivations for those rating methods are of interest in the context of this chapter. This is especially so as it provides a link to a number of other, sometimes familiar, rating methods.

1.7.1 PageRank (Daniels, 1969)

Daniels (1969) appears to have been the first to document the link between the Bradley-Terry model and what might now be recognised as an undamped PageRank (Page et al., 1999). PageRank has come to be widely known as it formed the basis for the original Google search algorithm. An intuitive explanation for the way it functions is the so-called 'random surfer' model. It envisages a surfer, who is randomly assigned to a node in a directed network. The random surfer then moves randomly with equal probability to one of the other nodes to which there is a directed edge from the node where they are currently. This process continues with the proportion of time spent on each node representing the PageRank for that node. Thus the PageRank rating vector is the eigenvector to the transition matrix with eigenvalue 1.

In the notation of this chapter, taking the comparison matrix to define the relevant directed network, with c_{ij} the weight of the directed edge from j to i, the PageRank rating vector $\boldsymbol{\alpha}_{PR}$ is thus the stationary distribution for the column-normalised comparison matrix

$$\boldsymbol{\alpha}_{PR} = CD^{-1}\boldsymbol{\alpha}_{PR},$$

where D is the diagonal matrix of column sums, $d_{jj} = \sum_k c_{kj}$.

While this rating is perhaps best known from its link to PageRank it had been previously identified as the 'total influence' metric in Pinski and Narin (1976) in the context of bibliometrics. It has been independently axiomatised in Altman and Tennenholtz (2005) and in Slutzki and Volij (2006). More prosaically, such a measure might be motivated in the context of sports competition by the idea of a 'glory-seeker' fan, or as Langville and Meyer (2012, p. 68) terms it the 'fair weather' fan. Consider a fan who begins by selecting a team to support at random. At each step they

transfer their allegiance to one of the teams that has beaten the team they previously supported. This decision is made at random in proportion to the number of their defeats that were against each team. Each team is then rated by the proportion of time that the glory-seeker has spent supporting them.

Consider now $\boldsymbol{\pi} = D^{-1}\boldsymbol{\alpha}_{PR}$. Then

$$\boldsymbol{\pi} = D^{-1}\boldsymbol{\alpha}_{PR} = D^{-1}CD^{-1}\boldsymbol{\alpha}_{PR} = D^{-1}C\boldsymbol{\pi},$$

so that $\boldsymbol{\pi}$ is an eigenvector for $\hat{C} = D^{-1}C$. In the context of sports, $\boldsymbol{\pi}$ might be thought of as an adjusted undamped PageRank so that teams that have played more matches are not overweighted. Alternatively the adjustment might be motivated by considering, for example, a five team round-robin tournament where A beats B, C and D; B beats C, D and E; C beats D and E; D beats E; and E beats A, as represented in Table 1.1.

	A	B	C	D	E	Score
A	0	1	1	1	0	3
B	0	0	1	1	1	3
C	0	0	0	1	1	2
D	0	0	0	0	1	1
E	1	0	0	0	0	1

Table 1.1: Five team round-robin tournament

Undamped PageRank would rate A and E joint first, intuitively because every time the glory-seeker selects team A, they will subsequently select team E, whereas standard round-robin ranking would rate A as joint first and E as joint last.

A vector $\boldsymbol{\pi}$ is an eigenvector for $\hat{C} = D^{-1}C$ with an eigenvalue of 1 if and only if

$$\sum_j c_{ij}\pi_j = d_{ii}\pi_i \quad \text{for all } i,$$

but if $C = AS$ is quasi-symmetric such that A is a diagonal matrix and S is symmetric then choosing $\pi_i = a_{ii}$ yields

$$\sum_j c_{ij}\pi_j = \sum_j a_{ii}s_{ij}a_{jj} = a_{ii}\sum_j s_{ji}a_{jj} = \pi_i\sum_j c_{ji} = d_{ii}\pi_i \quad \text{for all } i,$$

so that the adjusted PageRank $\boldsymbol{\pi} = D^{-1}\boldsymbol{\alpha}_{PR}$ is the diagonal component of a quasi-symmetric matrix. Equivalently it is the Bradley-Terry rating vector in the special

case of a quasi-symmetric comparison matrix C and thus a consistent estimator for the Bradley-Terry rating vector given a Bradley-Terry data-generating process.

This rating method was proposed as the 'influence weight' measure by Pinski and Narin (1976), where c_{ij} within the comparison matrix represents a citation in journal j of an article in journal i. It was motivated by noting that journals are likely to be of different sizes and that one may be interested in determining influence independent of size. The proposal was therefore to normalise the citations received by i by the citations given by i. More recently, the 'Rank Centrality' algorithm of Negahban et al. (2012) proposes the same estimator applied to ratio matrices, and it is also equivalent to the 'Luce Spectral Ranking' of Maystre and Grossglauser (2015) in the $k = 2$ case. A more explicit discussion of the link was provided by Selby (2020).

1.7.2 Fair bets (Daniels, 1969)

Daniels (1969) introduces an idea referred to as 'fair scores'. It was elaborated on and cast in the perhaps more intuitive language of 'bets' by Moon and Pullman (1970). Both provide interesting discussions of more general approaches. More recently, Slutzki and Volij (2006) provides an excellent summary of the approach, providing two axiomatisations for it, a presentation of a more informal motivation due to Laslier (1997), the link to undamped PageRank, and a discussion as to why the axiomatisations may lead us to believe that the 'fair bets' method is more appropriate for sports tournaments, while the undamped PageRank is more suitable for citation networks.

The first of the axiomatisations shows that the 'fair bets' model is the unique ranking derived under the three simultaneous requirements of uniformity, inverse proportionality to losses, and neutrality. Uniformity here requires that if a tournament outcome is balanced in the sense that every competitor has the same number of wins and losses then the competitors must be ranked equally. Inverse proportionality to losses requires that if one begins with a balanced tournament outcome, and then a single competitor's losses are multiplied by a constant then its rating will be divided by the same constant relative to the other competitors. Neutrality requires that if one begins with a balanced tournament outcome and some new matches are added between two teams where they share the wins equally then competitors will remain equally ranked.

The second of the axiomatisations requires two axioms, consistency between a ranking and its reduced forms and reciprocity. Reciprocity here requires that, in a two-player tournament, the ratio of the two competitors' ratings is equal to the ratio of their wins in matches between them, assuming that there are a non-zero

number of matches between them. The reduced form condition considers a reduced tournament without a team k, with the comparison matrix modified to, in effect, reallocate results involving k so that the comparison matrix is redefined as

$$c_{ij} = \begin{cases} 0 & i = j \\ c_{ij} + \frac{c_{ik}c_{kj}}{\sum_t c_{tk}} & \text{otherwise.} \end{cases}$$

The axiom requires that the relative ratings of two teams in any reduced tournament are equal to their ratio in the full tournament. Consistency requirements of this type are a common feature of axiomatic approaches to ranking (Thomson et al., 1996).

Alternatively, inkeeping with the original presentation of Daniels (1969), suppose one retrospectively wishes to assign a betting scheme to a tournament, where the loser pays to the winner an amount on the result of each match. This is subject to two conditions. First, that the amount that is paid to the winner by the loser is a value dependent solely on the strength of the loser. So that if i beats j then i will receive an amount α_j^{FB} from j. Second, that the betting scheme is fair. Here 'fair' is taken to mean that the wagered amounts will have led to the result that betting on any team throughout the tournament will have a net gain of zero. Then one has the condition that, for all i,

$$\sum_{j;j\neq i} c_{ij}\alpha_j^{\mathrm{FB}} = \sum_{j;j\neq i} c_{ji}\alpha_i^{\mathrm{FB}},$$

where $\boldsymbol{\alpha}^{\mathrm{FB}}$ may be taken as a rating vector for the participants, with the intuition being that one would be prepared to wager more on a strong team.

If $C = AS$ is quasi-symmetric then we have for all i

$$\sum_{j\neq i} a_{ii}s_{ij}\alpha_j^{\mathrm{FB}} = \sum_{j\neq i} a_{jj}s_{ji}\alpha_i^{\mathrm{FB}},$$

so that

$$\sum_{j\neq i} s_{ij}(a_{ii}\alpha_j^{\mathrm{FB}} - a_{jj}\alpha_i^{\mathrm{FB}}) = 0.$$

Thus, $\alpha_i^{\mathrm{FB}} = a_{ii} = \pi_i$, and the Fair Bets rating is a consistent estimator for the Bradley-Terry rating vector given a Bradley-Terry generating process.

1.7.3 Wei-Kendall

The rating method introduced in Wei (1952) and Kendall (1955) relies on an iterative application of the comparison matrix. The motivation for such a procedure might be

seen by taking the tournament example from Section 1.7.1. One might argue that ranking D and E equally is unfair as E's single victory occurred against a top-ranked team A, whereas D gained its only victory against bottom-ranked E. An approach to address this suggested by Wei (1952) is to weight each victory by the rating of the defeated team. The notion of inheriting the wins of a defeated opponent to inform a rating is somewhat intuitive. For example, the idea is present in the predominant rating system of the British playground game of conkers (Barrow, 2014). Under the Wei-Kendall method we would begin with a rating vector defined by the sum of wins

$$_{\mathbf{1}}\boldsymbol{\alpha}_{WK} = C\boldsymbol{e} = \{3, 3, 2, 1, 1\}^T,$$

where $\boldsymbol{e}$ is a $n \times 1$ vector of 1s. Then we assign to each team the sum of the first iteration ratings of each team they have beaten

$$_{\mathbf{2}}\boldsymbol{\alpha}_{WK} = C_{\mathbf{1}}\boldsymbol{\alpha}_{WK} = C^2\boldsymbol{e} = \{6, 4, 2, 1, 3\}^T.$$

This second iteration measure is sometimes used in chess for tie-breaking, where it is known as the Sonneborn-Berger score (Hooper and Whyld, 1996). But then one might reason that the victories should instead have been weighted by this updated rating. Proceeding in this way for the next five iterations we have Wei-Kendall rating vectors

$$\begin{aligned}
_{\mathbf{3}}\boldsymbol{\alpha}_{WK} &= \{7, 6, 4, 3, 6\}^T, \\
_{\mathbf{4}}\boldsymbol{\alpha}_{WK} &= \{13, 13, 9, 6, 7\}^T, \\
_{\mathbf{5}}\boldsymbol{\alpha}_{WK} &= \{28, 22, 13, 7, 13\}^T, \\
_{\mathbf{6}}\boldsymbol{\alpha}_{WK} &= \{42, 33, 20, 13, 28\}^T, \\
_{\mathbf{7}}\boldsymbol{\alpha}_{WK} &= \{66, 61, 41, 28, 42\}^T.
\end{aligned}$$

Note that E continues to be ranked higher than D and C.

Generalising, one may define a series of rating vectors

$$_{\boldsymbol{k}}\boldsymbol{\alpha}_{WK} = C^k\boldsymbol{e}.$$

It is then natural to consider the limit, but this is clearly not convergent. However, as Moon (1968) notes, since the matrix C is irreducible then by the Perron-Frobenius theorem (Frobenius, 1912) the rating vector defined by

$$\boldsymbol{\alpha}_{WK} = \lim_{k\to\infty} \left(\frac{C}{\rho}\right)^k \boldsymbol{e},$$

where ρ is the dominant eigenvalue of C, is convergent and this normalised limit may be thought of as a rating vector. In the case considered above this gives

$$\boldsymbol{\alpha}_{WK} = \{1.63, 1.38, 0.87, 0.55, 0.95\}^T.$$

The same argument can be applied to give a consistent estimator of the Bradley-Terry rating vector in the case of a Bradley-Terry data-generating process. In both cases, the idea is that we start with an intuitive rating vector and then reweight wins based on their quality as reflected by the ratings. In the case of the Wei-Kendall method, the initial rating is based on the number of wins. Here, the first rating is based on the win-loss ratio of each team, $\hat{C}\boldsymbol{e} = D^{-1}C\boldsymbol{e}$. As before, one may argue iteratively that the value of the wins within that calculation should not be assumed equal and should instead be weighted by their rating. Proceeding in this manner, we define a rating vector

$$\boldsymbol{\pi} = \lim_{k\to\infty} \hat{C}^k \boldsymbol{e}.$$

Since the scaled matrix $\hat{C}$ has unit dominant eigenvalue, then by Perron-Frobenius Theorem the limit is convergent and $\boldsymbol{\pi}$ is equal to the leading eigenvector of $\hat{C}$. If additionally $\hat{C}$ is quasi-symmetric, which it will be if C is quasi-symmetric, then this leading eigenvector will be the vector of Bradley-Terry ratings. Thus by applying the same reasoning used to motivate the Wei-Kendall method, but starting with an alternative plausible initial rating vector, we derive a consistent estimator for the Bradley-Terry rating vector given a Bradley-Terry data-generating process.

1.7.4 Ratings Percentage Index

A rating measure that until recently was prevalent in college sports in North America is the Ratings Percentage Index (RPI). It is commonly defined as

$$\begin{aligned}\text{RPI} = {} & 25\% \times \text{Win Percentage} \\ & + 50\% \times \text{Opposition's Win Percentage} \\ & + 25\% \times \text{Opposition's Opposition's Win Percentage.}\end{aligned}$$

In the notation of this article, recalling that M is the matrix of the number of matches, let the matrix $\hat{M} = [\hat{m}_{ij}]$ with $\hat{m}_{ij} = m_{ij}/\sum_j m_{ij}$, so that $\hat{m}_{ij}$ is the proportion of i's matches that are against team j. Define the win percentage vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ where $x_i = \sum_j c_{ij}/\sum_j m_{ij}$, then the RPI rating vector $\mathbf{RPI} = (\text{RPI}_1, \text{RPI}_2, \ldots, \text{RPI}_n)^T$ may be defined as

$$\mathbf{RPI} = 0.25\boldsymbol{x} + 0.5\hat{M}\boldsymbol{x} + 0.25\hat{M}^2\boldsymbol{x}$$

An argument very much like the one in the previous section may be followed to motivate this, that we must consider the strength of opposition in aggregating wins and that we can do this iteratively. In the RPI it is assumed that the previous iterations carry information that should be included in the overall rating and that three such applications is sufficient.

The choice of win percentage as the initial rating vector and of the proportion of matches as the relevant weighting factor when taking account of the strength of opposition is not unintuitive, but not exclusively so. For example, one might instead take each team's win-loss ratio as the initial rating and account for the strength of opposition by weighting wins, rather than matches, in line with those ratings. The 0.25/0.5/0.25 weighting is arbitrary and indeed has been criticised as overweighting the strength of a team's opposition and for producing perverse incentives (Baker, 2014). In the absence of any clear reason to do otherwise, an equal weighting might instead be applied. This would give an initial rating vector

$$\boldsymbol{\alpha_1} = \hat{C}\boldsymbol{e},$$

and considering down to an opposition's opposition's strength as in RPI

$$\boldsymbol{\alpha_3} = \frac{1}{3}\hat{C}^2\boldsymbol{\alpha_1} + \frac{1}{3}\hat{C}\boldsymbol{\alpha_1} + \frac{1}{3}\boldsymbol{\alpha_1} = \frac{1}{3}(\hat{C}^3 + \hat{C}^2 + \hat{C})\boldsymbol{e}.$$

Clearly there is no particular reason to stop after recursively considering two levels of opposition antecedents and so one might more generally consider

$$\boldsymbol{\pi} = \lim_{r\to\infty} \frac{1}{r} \sum_{k=1}^{r} \hat{C}^k \boldsymbol{e}.$$

This is the row sum vector of the Cesaro average for $\hat{C}$ and so

$$\boldsymbol{\pi} = \lim_{k\to\infty} \hat{C}^k \boldsymbol{e},$$

giving the same result as under the Wei-Kendall procedure applied to $\hat{C}$. And so we have that an RPI-style rating applied to win-loss ratios also gives a consistent estimator for the Bradley-Terry rating vector given a Bradley-Terry data-generating process.

1.7.5 "Winner stays on" - Barker's algorithm

It is a convention in some settings, for example pub pool tables, to play on the basis of "winner stays on", where the winner of any match continues to play the

next competitor. While rarely part of an official ranking system, it is intuitive that players who spend more games as "reigning champion" might be considered stronger. If we assume that games conform to a Bradley-Terry data-generating process and that new opponents are selected randomly then the probability of the "reigning champion" status transferring form j to i given a game between them is

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j}.$$

If we consider that the "winner stays on" scheme continues indefinitely with opponents selected uniformly at random and an irreducible set of results obtained, then, by the ergodic theorem, the long-run proportion of games that each player has been reigning champion is proportional to the Bradley-Terry strengths, $\boldsymbol{\pi}$. The "proportion of time as reigning champion" rating within a "winner stays on scheme" where opponents are selected uniformly at random is thus a consistent estimator for the Bradley-Terry model.

A very similar idea underlies Barker's algorithm (Barker, 1965) in the context of discrete space Markov chain Monte Carlo simulation. In that context, we have a distribution known up to a scaling and we wish to draw a sample from it. At each iteration a proposal $\theta \to \phi$ is generated and accepted with Barker acceptance probability

$$\alpha(\theta, \phi) = \frac{\pi(\phi)q(\phi, \theta)}{\pi(\phi)q(\phi, \theta) + \pi(\theta)q(\theta, \phi)},$$

where $\pi(.)$ is the target density and $q(\theta, .)$ is the density of the proposed transition from θ. The transition density q is selected to be symmetric, $q(\theta, \phi) = q(\phi, \theta)$, giving the resultant acceptance probability as $\pi(\phi)/(\pi(\phi) + \pi(\theta))$. By the ergodic theorem, this generates a Markov chain with long-run occupation times of states proportional to π.

1.8 Concluding Remarks

Special status is accorded to models and phenomena that become apparent from a diversity of seemingly unrelated scenarios. It is partly in this spirit that this chapter is written. Undoubtedly some of these motivations carry more weight than others. It would seem, for example, that being the unique solution to maximising entropy subject to the retrodictive criterion may be a relevant motivation in more scenarios than being a readily hypothesised model for a sudden death contest on a difference of r points. Nevertheless, the number and diversity of motivations is suggestive of

the wide applicability and attractiveness of the model, and lays the basis for its referencing and use later in this thesis.

The very first, and also the oldest explicit, motivation addressed in this thesis was that of Good (1955) and the transitivity of odds. Transitivity is a key assumption that modellers make when ranking. Indeed, it is definitional to ranking. However, it is not clear that transitivity does obtain in the type of relations that ranking seeks to model. This is the topic of the next chapter.

Chapter 2

The transitivity of 'better than' in competitive sport and elsewhere

Abstract

An assumption in performing ranking is that the comparative being expressed through that ranking is transitive — if A is better than B and B is better than C, then A is better than C. The philosophical claim that the 'better than' relation is transitive is widespread but disputed. For statisticians, the question of the transitivity of 'better than' has import, as if 'better than' relations are not transitive, then it calls into question the exercise of ranking itself. Further, philosophical consideration of the relative nature of 'good' and 'better than' may offer insight on comparative approaches and guidance on the design of ranking measures.

In this chapter, I argue that in competitive sports of the type considered, 'expected to beat' is a plausible notion of 'better than'. An example is presented to demonstrate that this is an intransitive relation, and to explain how one might reconcile this with challenges to intransitive 'better than' notions proposed by philosophers, in particular those of monotonicity and semantics. The competitive sport context is pertinent to this thesis. But it also allows notions of 'better than' to be grounded in the readily interpretable idea of winning or losing, and for a presentation of the argument in a quantitative framework that obviates many objections that have been proposed to the well-known *spectrum arguments*. In a final section, I show how these arguments might be adapted to a wider moral realm, as considered by those philosophers. In doing so, I develop a novel class of betterness cycles, so-called *unambiguous arguments*, that take an unambiguously intransitive relation as their starting point. Stylistically, this chapter adopts some of the features of the

philosophical literature of which it is a part, with more liberal use of the first person singular and of footnotes.

2.1 Introduction

Many hold that the relation 'better than' is transitive, such that for any value bearers x, y, and z, if x is better than y, and y is better than z, then x is better than z (see, for example, Broome (1991, p.11-16), Broome (2004, p.50-63), Binmore (2008, p.26-28), Parfit (2011, p.128), Chang (2014, p.35)). The *spectrum arguments* developed by Rachels (1998) and Temkin (1987, 1996, 2014) have provoked an active discussion on the transitivity of 'better than' in the Philosophy community. Spectrum arguments present a sequence of states — outcomes, lives or experiences. These are typically characterised by two dimensions, often pain or pleasure in one dimension and time or numbers of lives in a second dimension. The states are ordered, such that each is worse (better) than the last, but the final state is better (worse) than the first. As way of illustration, Temkin (2014) offers a particularly lucid example. He considers a series of alternative lives, A_1 to A_{n+1}. All lives are subject to a low-level background annoyance of 15 mosquito bites per month. He then supposes that all lives are the same except for an additional exposure of:

> A_1: two years of excruciating torture
>
> A_2: four years of torture whose intensity is almost, but not quite, as severe as that of A_1
>
> ...
>
> A_n: a very mild discomfort for a very long time
>
> A_{n+1}: one extra mosquito bite per month for a sufficiently long time that A_{n+1} is perceived worse than A_n

The claim is that intermediate states A_2, A_3, A_4 and so on may be chosen with lower amounts of pain and higher durations each time, such that for any $k = 1, \ldots, n$, A_{k+1} is worse than A_k, but that A_1 is worse than A_{n+1}.

Many responses to the spectrum arguments have considered aspects of incommensurability, incomparability, or vagueness (Qizilbash, 2005; Knapp, 2007; Handfield, 2016; Handfield and Rabinowicz, 2018; Thomas, 2021), or have focused on the structure of the spectrum arguments, comparing them to Zeno's or sorites paradoxes, or questioning the intuitions of the pairwise judgements (Voorhoeve and Binmore,

2006; Binmore, 2008; Voorhoeve, 2008, 2013; Pummer, 2018). The aim of this chapter is to investigate the nature of the transitivity of 'better than', avoiding these considerations, by using examples where both the attributes of alternatives and the comparisons of those alternatives are clearly defined, quantitative and independent of any similarity-based reasoning.

Competitive sport provides the setting for the main working example. The use of competitive sport allows for an appeal to a strong intuition around the context and its features, and those may reasonably be represented and compared quantitatively. The term 'competitive sport' in this chapter is restricted to the situation where matches consist of two competitors. This is the predominant structure in team sports. While others have noted the domain of competitive sport as it pertains to the transitivity of 'better than' (Sugden, 1985; Broome, 2004; Temkin, 2014; Bordner, 2016), I hope to demonstrate that a more extensive consideration can be illuminating.[1] My starting point is to consider various relations and examine their claims to being a 'better than' notion, including those claims relating to transitivity.

I discuss two arguments in support of the transitivity of 'better than' at greater length. First, I present an argument based on a monotonicity principle, roughly if Q is some property and x is Q and y is Qer than x, then y is Q. Monotonicity principles are not refuted, but their appropriate application is clarified and hence it is demonstrated why we might not conclude that they must lead to the rejection of intransitivity in the relation 'better than'. Second, I present an argument based on semantics. Broome (2004, p.50-51) notes, "Some authors write as though the transitivity of betterness is an issue in ethics. It is not; it is an issue in semantics." Roughly, the argument states that 'A is more Q than B' means that the degree to which A has property Q is greater than the degree to which B has property Q. Since 'greater than' is a transitive relation then 'more Q than' is transitive. This applies to all comparatives of the form 'more Q than' or 'Qer than', with 'better than' being an irregular synonym of 'more good than'. In addressing this argument, I make the case for 'better than' as a precursor for goodness, rather than the transitive-implying understanding of goodness as the precursor for 'better than'.

With regard to terminology, I use 'better than' to refer to what some might term 'all things considered better than', as in general the distinction plays no role in the

[1]Broome (2004, p.52) dismisses the sports example as a potential counterexample to the transitivity of 'better than': "The very fact that the relation 'can regularly beat' may be intransitive amongst football teams should make you realize it is not equivalent to 'is better than'." Thus, he recognises the potential intransitivity of a relation such as 'can regularly beat' in the sports scenario but assumes that this in itself must disqualify it from any claim to being a 'better than' relation. This is question-begging.

arguments made. Where the distinction has relevance, in the discussion of natural language meaning, it will be expressed explicitly. In the moral realm, the 'better than' notion is that of the *reason-implying* sense described in Temkin (2014, p.13): "[r]oughly, on this use, outcome A is better than outcome B, all things considered, if one would have more reason to prefer A to be realized than B, from an impartial perspective." With 'all things considered' taken also from Temkin (2014, p.15): "to say that X is better than Y *all things considered* is simply to contend that there is most reason to prefer X to Y from an impartial perspective after accurately taking into account *all* of the factors that are relevant and significant for comparing such outcomes from that perspective." These definitions seem to be translatable in some meaningfully analogous way to a competitive sports context. Thus, I take A to be better than B in a competitive sports context if one would have more reason to believe A to be better than B from an impartial perspective after accurately taking into account all of the factors that are relevant and significant for comparing such teams from that perspective. 'Intransitive' will be taken to mean that for a relation R, aRb and bRc do not imply aRc.

The chapter proceeds in Section 2.2 with a discussion of some of the relations that we might understand to constitute 'better than' in the context of competitive sport, noting that the relation 'expected to beat' has appealing features in this context, but that it is intransitive. A working example is introduced in order to discuss some of the challenges that an intransitive notion faces. Two of the most significant are addressed in the subsequent sections. In Section 2.3, monotonicity is discussed by highlighting the apparent paradox created by holding simultaneously a general monotonicity principle and an intransitive notion of 'better than'. In Section 2.4, the semantic justification for the transitivity of 'better than' is discussed. In Section 2.5, I summarise the arguments up to this point and note that even under more complex notions of 'better than' intransitivity is likely to obtain. In Section 2.6, the arguments considered in Sections 2.3 and 2.4 are used to motivate an example in the moral realm. This aims to demonstrate that intransitivity is not a property exclusive to 'better than' in a competitive sports context. In the final section some short concluding remarks are provided.

2.2 'Better than' in competitive sport

The aim of a competitor within a competitive sport is to win. When situations arise that compromise this, stakeholders object to the circumstance because it transgresses

this principle.[2] Since this is the defining aim of a competitor (the lusory goal as defined by Suits (1978) and discussed further in Chapter 3), then an acceptable 'better than' relation in a competitive sports context needs to be consistent with the principle that winning is better than losing. For example, under this stipulation I may not claim that Arsenal are a better soccer team than Manchester City because they play more attractively (unless I also claim that the attractiveness of the soccer is monotonically related to winning). I may claim that Arsenal are, or were, better than Manchester City because they have beaten them, because they have a better record against equivalent opposition, or because they would be expected to beat them, since all of these notions respect the primacy of winning over losing. As such, multiple definitions of 'better than' accord with this principle. We will proceed by examining the claims that some of these have on being a 'better than' notion for competitive sport.

Let us start by taking the relation 'has beaten'. It is clearly not necessarily transitive. It is common in sport for team A to have beaten team B, team B to have beaten team C, and team C to have beaten team A. Setting this aside for now, one may make at least four other objections to using this as a notion of 'better than'. First, this outcome may be due to some arbitrariness in the particular instantiation. Upsets are a common feature in sport. This is so much a part of sport that the refrain "may the best team win" is commonplace on the eve of matches. Second, this is an outcome under a particular set of external conditions. The weather and officiating decisions, for example, may have an impact on a match and are beyond the attributes of the teams. Third, it is possible that A has beaten B, and B has beaten A. Most would agree that 'better than' is asymmetric. Fourth, 'has beaten' represents an instantiation from a particular point in the past. It is not clear to what time period the 'better than' relation may then apply. If A beats B but then goes on a bad run of form, while team B goes on to win the championship, it does not seem credible to maintain that A is better than B. One might insist that the 'better than' relation applies solely to the period over which the match took place, but then it is a very limited notion, and it seems we are able to make comparisons outside of that period.

An alternative relation would be the predominant relation used in league sports,

[2]An example of this comes from the 2018 soccer World Cup. In the final game in Group G, England were to face Belgium, with both teams having qualified for the knock-out stages already. Based on results in other groups, it seemed to be the case that losing offered a clearer route to the final. England and Belgium together changed 17 of their 22 starters from their previous group match, which they had both won, suggesting that winning was not a priority. There was much discussion and some discontent at this state of affairs.

namely 'has aggregated more wins against all other competitors'. The ubiquity with which this notion is applied might support its case. But there are several goals of a league tournament, including the creation of a ranking of competitors, making that ranking methodology transparent to stakeholders, and identifying the best team. There is no a priori reason to believe all (or any) of these are possible simultaneously. Definitionally, to produce a ranking one requires a transitive relation. So even if there were an agreed upon notion of 'better than', if it were intransitive, then it could not be employed directly for this purpose. So, we must instead look at the attributes of the relation. Let us consider the first two objections raised previously, those of random instantiations and external conditions. While the impacts of these are averaged over the outcome of all matches, they are reduced but not eliminated. Third, the consideration of time becomes yet more unclear with an aggregation of outcomes from different time points, so that we cannot even relate 'better than' to a particular point in the past. What relative weight should a result from a year ago take compared to a result from a week ago? Fourth, it may not be that we have a complete set of matches on which to compare teams, for example at the mid-point of a season, or in the case of school or college sports.

This may then encourage us to consider the relation 'expected to aggregate more wins against all other competitors'. In referencing expected rather than actual instantiations, we are considering this claim over all possible conditions and randomness in line with the probability that they occur, and it may be referenced to any particular point in the past, present or future by using variants 'was/is/would be expected to aggregate more wins against all other competitors', and so objections based on the randomness and circumstances of instantiations and the lack of temporal clarity are avoided. However, whether A is better than B then depends on the other teams in the comparison set. It would seem desirable that if A and B have not changed in their qualities then whether A is better than B ought not to change. But, if the relative betterness of A and B is dependent on other teams in the comparison set then their relative betterness can change due to changes in the other teams. In this sense, it is not independent of irrelevant alternatives.

It might be that there is a natural comparison set within which any particular pair should be considered and consists of the same teams over time. For example, if comparing Manchester City and Arsenal, then perhaps one might take the teams in the Premier League to be the most relevant comparison set. But those teams will be continuously evolving as their players develop or grow old or as their manager tries new tactics. So there is still the possibility for the betterness of A with respect to B to change without A or B changing in their qualities. Alternatively, one might take the relevant comparison set to be the set of all possible soccer teams. As Temkin (2014)

notes in discussing his *Sports Analogy*, the number of calculations required would be very high, making it of limited practical use. But setting aside this epistemic objection, how does one define a 'possible soccer team'? For example, if it is any group of eleven people who consider themselves to be a team, any group who are considered to be a team by others, or all possible groupings of eleven people who consider themselves to be soccer players then these will change over time and thus the relative betterness of A and B may change without their qualities changing. It seems likely that under any plausible identifiable comparison set the other teams are likely to change continuously over time.

Instead, we might understand 'all possible teams' to span a space of the various relevant qualities of a team. It is far from clear whether such orthogonal qualities could be defined, but assuming one could, then these qualities would seem to be continuous. If two teams have two different levels of fitness say, then there would seem to be an uncountably infinite number of intermediate levels of fitness, and so the space of 'all possible teams' also has an uncountably infinite number of teams. This would not matter if we could come to a conclusion by comparing teams based on a finite sample from the space. But there seems no way to do this since the interaction of relative styles and strengths in competitive sport means that even if we know that the probability that A beats X is greater than the probability that B beats X, we may note generally conclude that this will also be true against an alternative opponent Y. So 'better than' becomes undefined, which is inconsistent with our ability to use the term.

Another objection to 'expected to aggregate more wins against all other competitors' is one related to the discussion of Smead (2019). Given a set of match probabilities there are many potential aggregations of these that may be used in order to provide a ranking on which the assessment of 'better than' would be based. Ranking by a simple sum of wins is a convention in sports, but there may be reasons for that based on the competing aims of tournaments, including notably transparency, other than it being an accepted notion of 'better than'. Instead, perhaps (expected) wins against higher ranked opposition should count for more. Even 'expected to aggregate more wins against all other competitors' might be ambiguous. Is the comparison being made based on the number of expected wins or the expected number of wins? For the purposes of the following arguments, 'expected to aggregate more wins against all other competitors' will continue to be used and will be taken to mean a comparison based on the expected number of wins. Later in the chapter, I take the term 'aggregate better than' to refer to the family of plausible 'better than' relations that aggregate the probabilistic match outcomes.

A 'better than' notion that might seem to avoid these objections would be to

determine a single 'strength' value for each team based on their salient qualities and then determine that A is better than B if and only if it has a higher strength. Such a relation is independent of arbitrary instantiations and circumstances, definable at a particular point in time based on the qualities at that point in time, and appears to be independent of irrelevant alternatives. It is also transitive. This is analogous to what Temkin (2014) calls an *Internal Aspects View*, roughly that A is better than B if and only if the extent to which A is good, as determined solely on the basis of A's internal features is greater than the extent to which B is good, as determined solely on the basis of B's internal features. The *Internal Aspects View* has wide appeal as it accords with many comparatives with which we are very familiar.

In the present context, the strength value must accord with a propensity to win, based on the aims of competitive sport. This entails that given a set of salient qualities for each team and the relevant set of competitors, it will be possible to define a single strength value based on the qualities that will accord with some plausible notion of propensity to win. Clearly it cannot accord with an intransitive notion of propensity to win, but even taking a plausible transitive notion it faces difficulties. Under the highly plausible assumption that the qualities of one team interact with those of an opposition in the determination of a match outcome then it is not possible to define a single metric based on the qualities of each team that can account for all possible comparison sets and maintain an accordance with a propensity to win. If we allow the function by which we convert the qualities into a single strength value to alter with comparison sets then it ceases to be independent of irrelevant alternatives. So either it does not accord with a propensity to win, or it is not independent of irrelevant alternatives. There may also be a challenge in aggregating meaningful qualities that exist on orthogonal scales. These arguments will be elucidated further with the example in Section 2.3.

A suggestion that avoids this challenge but maintains transitivity and independence of irrelevant alternatives would be to claim A is better than B if and only if it is superior with respect to each of the relevant qualities. But suppose we consider fitness to be a relevant quality in the determination of the betterness of a soccer team. Under this notion, if we were to compare Manchester City with a team made up of the fittest eleven people in the world who had never played soccer (who we take to be fitter than Manchester City) then we would be unable to conclude that Manchester City were a better team, even if they were expected to beat them (comprehensively) and were expected to perform better against all opposition.

So let us consider the relation 'expected to beat', where A is expected to beat B if and only if the probability that A beats B is greater than the probability that B beats A (for ease of explication I assume binary win/loss outcomes). None of the

objections mentioned so far pertains to this — it is not subject to arbitrary outcomes or circumstances, the relevant time may be expressed precisely, it is asymmetric, it is dependent solely on the qualities of the two teams being compared, and it does not depend on a definition of a comparison set. It is also, importantly for some, not at odds with the natural language meaning in that when told "A is (all things considered) better than B" many will interpret this as "A would be expected to beat B". However, 'expected to beat' is intransitive.

Two significant objections to an intransitive 'better than', those of monotonicity and semantics, will be discussed in more detail in the following sections. In order to facilitate those discussions and to clarify the sense in which 'expected to beat' may be intransitive the following example is introduced.

The Intransitive American football teams[3]

Take three American football teams A, B and C. Let us suppose that their important qualities may be summarised by their offensive and defensive ability in their running and passing game, and that points are expected to be scored in a monotonically increasing way with the difference in the strength of a team's chosen offense and the opposition's corresponding defense. We also assume that the coaches know their opposition's qualities, presumably having reviewed their past matches, so they will choose a running or passing offense dependent on where they have greatest advantage (and will flip a coin if they are equal). Let us suppose that their qualities are summarisable as in Table 2.1.

Consider a game between A and B. As the coach of team A, you wish to select your most effective offense. You note that team B's run defense is rated at -2 and so with your run offense, rated at 0, you have a net advantage of 2 $(= 0 - (-2))$ if playing a run offense. Your pass offense is rated at 0 and your opposition's pass defense is

[3]American football serves as a useful example here because the nature of what goes on is more readily discretised than in other sports. Many team sports share the binary of being in possession of the ball (offense) or not (defense). But in American Football, a team with the ball will attack in one of two ways — by giving it to a player who attempts to run round or through the opposition (run offense) or by throwing the ball to a player downfield (pass offense). Teams will have varying strengths in their ability to execute run or pass offensive and defensive operations. The game is broken down into discrete units of play, where the team in possession, and specifically often their coach, will decide what kind of offense (run or pass) to execute. The defensive team have to react accordingly. The example is a simplification as, in practice, all teams will use a mixture of pass and run offense in order to keep the defensive team guessing from one play to the next. But it is a reasonable characterisation as the proportions of run or pass offensive plays will be selected based on their relative strengths.

	Run		Pass	
Team	Offense	Defense	Offense	Defense
A	0	0	0	0
B	$+1$	-2	0	$+1$
C	-1	-1	$+2$	0

Table 2.1: Qualities of teams A, B and C

rated at $+1$ and so you have a net advantage of -1 $(= 0 - (+1))$ if playing a pass offense. You therefore choose a run offense as it brings you greater net advantage. Team B's coach makes a similar calculation and finds that a run offense gives a net advantage of $+1$ $(= +1 - 0)$ and a pass offense a net advantage of 0 $(= 0 - 0)$. They therefore also choose a run offense. However, Team A's net advantage from their selected offense of $+2$ is greater than team B's net advantage from their selected offense of $+1$. Team A is therefore expected to beat team B. Following the same reasoning we would expect team B to beat team C, and team C to beat team A. The relation is therefore intransitive.

The quality values should be understood as cardinal values. A uniform positive linear transformation of these quality values does not change the conclusion, and for some it may aid interpretability. For example, if a constant between two and eight is added to every number then they might be understood as a grading on a scale of 0 to 10. The specific numbers used here are chosen to highlight notions of neutrality and as the minimum absolute integer value version. The values may be thought of as comparable in that they all have a direct impact on expected score, which is a consistent scale.

Before moving on to the substantial objections based on monotonicity and semantics, I address two other challenges. The first is that this is merely a rock-paper-scissors set-up, and no-one would understand it to be correct to apply the 'better than' relation to rock-paper-scissors. But the set-up here is different. In rock-paper-scissors whether player 1 will beat player 2 is of primary interest, and each player has a choice from all three items. In contrast, in the *Intransitive American football teams* example the pair for comparison is set and the expected outcome of that comparison is of primary interest. The *Intransitive American football teams* example is also distinct in that the cycle arises naturally from plausible qualities of the items under consideration and the nature of the comparisons rather than being directly stipulated. The second challenge is that this example constitutes a change of criteria, with the expected match outcomes being determined by a comparison of run qualities in the case of A vs B and B vs C (depending on C's coin flip) but a

combination of pass and run in the case of C vs A. The argument goes that if the relevant criteria are being changed then it is not the same relation. But all qualities are considered in all comparisons and in the same clearly defined and plausible way each time, consistent with the primacy of winning in competitive sport. The result of the nature of the comparison is that different qualities have different force depending on the items being compared, and that is a feature that I claim to be true of the 'better than' relation in these circumstances.

2.3 Monotonicity

A prominent argument offered in favour of the transitivity of 'better than' is monotonicity. Here the objection will be framed in the form of a paradox. The particular example presented is derived from the account of Nebel (2018). Consider a monotonicity principle defined as follows.

> **The strong monotonicity principle**
>
> For any property P, with the opposite property Q, if x is not P (i.e. it is Q or neutral) and y is Qer than x, then y is Q.

Suppose we take 'better than' to be 'expected to beat' and define a neutral team A with qualities r_o, r_d, p_o, p_d for run offense and defense, and pass offense and defense respectively. Then we may determine a team B with respective qualities r_o+1, r_d-2, p_o, p_d+1, and a team C with qualities r_o-1, r_d-1, p_o+2, p_d. We will then have the cyclic triad as before, but by the *strong monotonicity principle*, B is a bad team since it is worse than a neutral team, and C is a good team since it is better than a neutral team. But the bad team is better than the good team, since team B would be expected to beat team C, which violates the *strong monotonicity principle*. The apparent paradox relies on three assumptions:

1. we may have an intransitive 'better than' relation,
2. we may define a neutral alternative,
3. the *strong monotonicity principle* holds.

For many, the resolution is to refute the possibility of an intransitive 'better than'. In the last section, we saw that there may be independent reasons to understand 'better than' to be intransitive, and so here we consider more carefully the other assumptions.

In considering defining the neutral alternative, there is nothing particularly special about neutrality on the good-bad spectrum. Neutrality is used here as there exists readily available language to describe the situation: names for the point on the spectrum ('neutral') and for the two sides it separates ('good' and 'bad'). But if one could define a point that lies at the transition from 'bad' to 'very bad' then one could make the same argument and conclude that 'very bad' is better than 'bad'. We might even note that the *Combined Spectrum Arguments* that Nebel (2018) introduces are not necessary. A spectrum argument of the canonical type with one of the outcomes identified as a particular point on the good-bad spectrum would lead to the same conclusion given the same reasoning process. So, the point to be refuted is that one may identify any particular point on the good-bad spectrum.

There does seem to be some general meaning to what a 'good' team or a 'bad' team would be with regard to its qualities. A 'good' team would be one with high scores across its qualities, and a 'bad' team one with low scores across its qualities. This does not imply an ability to summarise the quality of a team in a single value related to the qualities however, and the qualities have value only in so far as they increase the propensity to win, in line with the aim of competitive sport. It may be appealing for example to consider that a team with mean values for each of the individual qualities would be a team of neutral quality. But consider the example presented in Table 2.2. Here the four teams in the comparison set each have the same mean of 0 across their qualities, and each of the qualities has the same mean of 0 across the four team comparison set. However, team D would be expected to lose all their matches. It therefore seems unsatisfactory to consider them a team of neutral quality.

	Run		Pass	
Team	offense	defense	offense	defense
D	0	0	0	0
E	-1	-1	$+2$	0
F	-1	-1	$+2$	0
G	$+2$	$+2$	-4	0

Table 2.2: Qualities of teams D-G

Perhaps the mean is just not the right statistic. But this will in fact be the case with any single value summary of the qualities. For example, it is possible for a team H to be worse than a team I based on all standard statistics summarising their qualities e.g. mean, median, mode, maximum, minimum, but still be expected to beat I and to be considered better than I in the two team comparison set under any

'better than' notion that accords with the stipulation that winning is better than losing. For example, as summarised in Table 2.3. Of course, one might determine that the appropriate strength value is not any of these standard statistics but instead equal to the run offense value, but then if teams J and K, with qualities as in Table 2.4, are added to the comparison set then H would be rated better than J and K under the strength value measure despite being expected to lose to and accumulate fewer wins than both. This demonstrates the sense in which an *Internal Aspects View* cannot accord with propensity to win given the highly plausible contention that a team's strengths and weaknesses will interact with those of their opposition to produce a win for one team or the other.

	Run		Pass	
Team	offense	defense	offense	defense
H	$+2$	-1	-9	-1
I	0	0	0	$+10$

Table 2.3: Qualities of teams H and I

	Run		Pass	
Team	offense	defense	offense	defense
J	0	$+2$	0	$+10$
K	0	$+2$	0	$+10$

Table 2.4: Qualities of teams J and K

In fact, the nature of the comparison function makes some of these statistics somewhat arbitrary. The pass and run qualities are defined up to a constant. If we add an arbitrary constant to all teams' pass values, both offense and defense, and a different arbitrary constant to all run values, both offense and defense, then the comparisons will not change. This does not negate that the qualities exist and interact to affect the match outcome, nor that they may be represented quantitatively. However it demonstrates another objection to the *Internal Aspects View*, that the aggregation of qualities into a single strength value may be arbitrary if the qualities exist on orthogonal scales.

These examples suggest the difficulties in defining a neutral team, or even a good or bad team, based directly on their qualities, even having considered them relative to the appropriate comparison set. They suggest that instead a neutral team ought to be defined with respect to an aggregation of their pairwise comparisons with

the other members of the appropriate comparison set. In this setting one plausible choice for the qualities of a neutral team would be those qualities that imply the team would be expected to win as many matches as they lose. This is consistent with the stipulation that the determination of 'good' or 'bad' must be made with respect to the propensity to win. Given a comparison set, it may thus be possible to define a neutral team, one that would be expected to win as many matches as it would lose. So there are grounds for not rejecting the assumption of being able to define a 'neutral' team.

However, the discussion gives a lead as to why we might believe that the *strong monotonicity principle* is misapplied in the characterisation of the paradox. We first note that the existence of a 'good-bad' spectrum presupposes a transitive order. One resolution is indeed to insist on a transitive 'better than' that accords with these notions of 'good', 'neutral' and 'bad'. An alternative resolution would note that the 'good' and 'bad' here might be better specified as 'aggregate good' and 'aggregate bad' since they are divided by a notion of 'neutral' that is reliant on an aggregation of the pairwise comparisons. In contrast the 'better than' might be understood as a 'pairwise better than'. To say that a team that is 'aggregate bad' is 'pairwise better than' a team that is 'aggregate good' is not paradoxical. In the language of the *strong monotonicity principle*, if x is Q and y is P, where P and Q are opposite properties, we are not saying that y is Qer than x. We are instead saying that y is Rer than x. Under this view, the *strong monotonicity principle* is not violated. So, if there are independent reasons for holding that 'better than' is an intransitive notion then the *strong monotonicity principle* does not itself preclude this.

However, if we are to understand 'better than' to be an intransitive 'pairwise better than', then the argument I have presented here requires that, in this context, one concedes that the most salient understandings of 'good', 'bad' and 'neutral' are not related by the most salient understanding of 'better than'. This leads us into the area of semantics.

2.4 Semantics

Broome (2004) argues forcefully that the transitivity of 'better than' is a semantic truth. He understands 'better than' to be the comparative of the monadic predicate 'good', with this being a semantic fact. He argues for this transitive stipulation on the basis that cyclical relations do not have the semantic structure 'Qer than' or 'more Q than', and for relations that do have this structure there is always an associated value of 'Qness' that is being compared. For example, if considering people sitting in a circle then 'to the left of' is a cyclical relation. But there is not an associated

meaning to 'person A is lefter than person B', because there is no meaning to 'person A is left'.

Broome (2004) considers possible exceptions to this, including 'later than'. One may say '1 a.m. is later than midnight' and '2 a.m. is later than 1 a.m.' and so forth until we have that 'midnight is later than 11 p.m.' completing the cycle. But then 'later than' does not seem to be related to lateness. Perhaps we might understand 11 p.m. to be late and 6 a.m. to be early. Then we can start at a time that is late, continue via times that are later than the previous one, and end at a time that is early. Broome (2004) seeks to resolve this by distinguishing between the historical lateness that is behind our intuition for the comparative 'later than' and the contextual lateness of 'late'/'early'. Based on a monotonicity principle, he asserts that there is therefore some cut-off point on the 24-hour clock, before which is 'late' and after which is 'early', and 'later than' acts in relation to these. He recognises this cut-off point as vague.

An alternative account is offered by the discussion of neutrality in the previous section. There I argued that 'neutral', and by extension any point on the good-bad spectrum, is an aggregate of pairwise comparisons, such that it is the pairwise comparison that is the precursor. Taking the relation 'later than' as a precursor explains why 'early' and 'late' are located where they are in contextual time, and also why they are vague. Due to the nature of human sleep cycles most of us have more experience of 6 a.m. being 'earlier than' than 'later than' in relevant comparisons with other times of the day, and of 11 p.m. being 'later than' rather than 'earlier than' in relevant comparisons. An aggregation of these comparisons leads to our identification of these times as 'early' and 'late' respectively. But there is no clear deterministic function for how these experiences ought to be aggregated; we have varying experiences over time, relatively little experience for sleep hours, and experiences vary across individuals, and so the resulting understanding is vague.[4]

These two accounts therefore agree that 'late' and 'early' exist in contextual time, and that they are vague. The account of Broome (2004) insists that it is a semantic truth that 'later than' is the relation that acts between items based on their lateness. The analogue to my account of 'better than' claims that the intuitive historical understanding of 'later than' is correct and provides the precursor to lateness. These

[4]Analogously my account readily explains why a sports team, whose intrinsic qualities do not change, may be viewed properly as 'good' in some comparison sets, and 'bad' in others. In a tournament with ten other teams all with the qualities of team B, then team A would be 'good', since they would be expected to win all their matches, but in a tournament with ten other teams all with the qualities of team C, then team A would be 'bad', since they would be expected to lose all their matches.

competing accounts both lead to semantically uncomfortable conclusions. Insisting on the precursor 'late' results in the claim that it would be incorrect to claim 'ten minutes after x is later than ten minutes before x', where x is our cut-off point between late and early. Insisting on the comparative precursor 'later than' leads to the possibility that 'x is later than y' where x is early and y is late.

Returning to 'better than', many would accept that goodness in competitive sport is derived from an aggregation of pairwise comparisons. As Broome (2004, p.52) notes "goodness of a football team is a complex matter involving the ability to do well against a variety of opponents." However, some might insist that these pairwise comparisons should properly be seen as something other than 'better than' relations. In arguing against the spectrum arguments, Voorhoeve (2013) argues that the pairwise comparisons may be seen as preference relations, and that these may be based on heuristics that are subject to psychological biases. It seems harder to dismiss the pairwise notion in competitive sport in the same way. There can be no claim that there is some comparable failure of judgement or misapplication of heuristics that have led to an incorrect judgement in the pairwise comparison. Furthermore, there are arguments for why a specifically pairwise understanding of 'better than' may be deserving of primacy, particularly in this setting.

First, "A is Qer than B" is a phrase that permits only the comparison of two items and gives no indication of a total comparison set other than the two items themselves. If only these two items are to be considered, that one is 'good' and the other 'bad' in the context of some wider comparison set may not be relevant for the determination of 'better than'. One might seek to challenge the claim that 'better than' is semantically pairwise by expanding it. For example, by saying "$\{A, B\}$ is Qer than $\{C, D\}$" or "A is Qer than B and C", but these are still pairwise. In "$\{A, B\}$ is Qer than $\{C, D\}$" the items are now sets rather than individual elements, but the relation still acts on exactly two items, and "A is Qer than B and C" is the conjunction of the two pairwise comparisons "A is Qer than B" and "A is Qer than C".

Second, if one accepts that goodness is based on an aggregation of pairwise comparisons that each consider all relevant qualities of the items being compared, then those pairwise comparisons are foundational. There would seem to be some sense in which a foundational comparison of overall relative merit has greater claim on a title of 'better than' than a comparison derived from an aggregation of foundational comparisons involving items outside of those being compared. One might point out that 'expected to aggregate more wins against all other competitors' when based on expected number of wins is not based on 'expected to beat' (though it would be if based on the number of expected wins). But I am arguing that probabilistic pairwise

comparisons are the foundational comparisons, with 'expected to beat' the appropriate way of interpreting betterness within a probabilistic pairwise comparison.

Third, in cases such as the *Intransitive American football teams* example the natural unit of comparison is uniquely pairwise. Matches are played between two teams, not three, or four, or five. This suggests that there is something essentially pairwise in the nature of 'better than' in this setting, by which I mean that the argument that whether A is better than B is independent of the nature of C has particular force.

Briefly addressing the semantic argument more directly, it supposes that the language we use is definitive for the nature of the comparative to which it is applied. But this is not always the case. For example, consider relative geographic position. 'North of' is a transitive relation, whereas 'west of' is cyclical, and yet they share a semantic structure. If one considers 'farther west than' to be a synonym to 'west of' then this may constitute a direct counterexample to Broome's claim that the semantic structure 'Qer than' is determinative of a transitive relation. Broome (1991) argues against an intransitive interpretation of 'more westerly than',[5] but the example of 'west of' and 'north of' might still cause us to be cautious of determining the transitivity of 'better than' based on semantic structure. It seems plausible that if a relation were predominantly transitive in our experience of it then we would adopt a form of language that was typically used for transitive relations. This may be the case for 'better than' without implying that it must be transitive in all circumstances.

2.5 Summary

Based on the requirement that any notion of 'better than' in competitive sport must accord with a propensity to win, two principal notions of 'better than' have been considered — 'expected to aggregate more wins against all other competitors' and 'expected to beat', characterised as 'aggregate better than' and 'pairwise better than' respectively. These share many features that mark them as preferable to alternatives. They are independent of particular instantiations or circumstances, have a clear meaning with respect to a particular point in time, and most importantly they accord with a propensity for winning. 'Expected to aggregate more wins against all other

[5]Of course, there is the more everyday sense in which we talk of a geopolitical 'West' and 'East'. These terms would seem to derive from societies being 'west of' or 'east of' other societies with which they came to interact, rather than any intrinsic property of 'westness' or 'eastness'. Indeed had the centre of mass of civilisation been located in the east Pacific ocean when these terms came to be coined rather than the Eurasian land mass, then it seems likely that we would now refer to North America and Europe as the 'East' and Asia as the 'West'.

competitors' accords with an understanding of 'good', 'bad' and 'neutral' in this context, but is not independent of irrelevant alternatives in the sense that whether A is better than B depends on the nature of C. It also faces the challenge of defining 'all other competitors', the epistemic challenge of computation against a large comparison set, and it is not clear that 'expected to aggregate more wins' is the uniquely best aggregation of the pairwise comparisons, or that a unique best aggregation exists. 'Expected to beat' accords with the natural unit of comparison and the pairwise semantic structure of 'better than' and is independent of irrelevant alternatives in the sense that whether A is better than B is independent of the nature of C, but it does not necessarily accord with a general understanding of 'good', 'bad' and 'neutral'.

For some, 'better than' must apply to the relative degree of goodness. I claim that 'better than' may be essentially pairwise, while 'good', 'bad', 'neutral' are inherently qualities that are held with respect to a wider comparison set. So if there are independent reasons for holding that 'better than' is intransitive then it is plausible that the most salient gradable quality and the most salient comparative are not compatible.

I have at times referred to the natural language meaning of 'better than'. The case made here has been for what we ought to understand by '(all things considered) better than' not what we do understand by '(all things considered) better than'. Nevertheless, it seems that we should not stray too far from usage or the term ceases to have the meaning we ascribe and instead is an alternative notion taking on a misleading label. It seems likely that the natural language meaning of '(all things considered) better than' in competitive sport incorporates both the notions discussed here and others besides. But if one grants that 'expected to beat' plays even some small role in the assessment of the natural language meaning of '(all things considered) better than' then the relation will inherit its intransitivity, since for any three teams equally rated on other notions but with a cyclic 'expected to beat' relation, the '(all things considered) better than' relation will be cyclic.

Importantly this also applies if we hold that 'better than' is some complex mix of notions but includes some element of 'expected to beat'. To paraphrase the characterisation of goodness due to Broome (2004, p.52), if we hold that betterness of one team compared to another "is a complex matter involving the ability to do well against a variety of opponents", then it seems plausible that the ability to do well against the direct comparator would be overweighted with respect to other comparisons and thus the relation would inherit intransitivity.

To the degree that the arguments in favour of an intransitive notion of 'better than' presented here are persuasive it is interesting to consider how they might have

relevance in the moral realm, which has been the focus of the Philosophy argumentation.

2.6 What of morality?

Consider a class of betterness cycle examples that take as their starting point the three part intuition that:

1. 'better than' is a complex relation,
2. complex relations are multi-dimensional,
3. multi-dimensional relations may be (and often are) intransitive.

This suggests that one might work backwards in this intuition by considering examples based on unambiguously intransitive relations in some non-moral space and relate them to a moral outcome in order to demonstrate intransitivity of 'better than' in the moral realm. As I noted in Section 2.4, these examples might have further force if there is an essentially pairwise comparison on which they are based. This strategy may be appealing, given the criticisms of spectrum arguments that generally question their intransitivity. Here there is certainly intransitivity, and the question is as to whether that intransitivity properly pertains to a 'better than' relation.

Consider the following example.

> **The racing evil-doer**
>
> Suppose there is an evil-doer who has identified a target 100 miles away and will take a vehicle to get there, and a good actor who wishes to thwart the evil-doer by taking another vehicle in order to get there first. There are only three types of vehicle. Their performance is known but unreliable. The time in hours that they will take to travel 100 miles is with equal probability of a third: 1, 6, or 8 hours for vehicle A; 2, 4 or 9 hours for vehicle B; and 3, 5 or 7 hours for vehicle C.
>
> Now suppose that there are only two vehicles available, but they are of as yet unknown type. The good actor is hurrying to the site of the two vehicles and is expected to arrive momentarily before the evil-doer. We must advise the good actor of the betterness relation of her choice set, so that when she arrives and is able to identify them, she is able to make a correct selection.

The probability that a vehicle of type A beats one of type B is 5/9, since with an equal 1/9 probability we have the time pairs $(1 < 2), (1 < 4), (1 < 9), (6 > 2), (6 > 4), (6 < 9), (8 > 2), (8 > 4), (8 < 9)$. Likewise, the probability that a type B beats a type C is 5/9, and that a type C beats a type A is 5/9; these probabilities being a direct result of the probabilistic speed of each.[6] I claim that with respect to the choice facing her, one may properly advise that A is better than B, B is better than C, and C is better than A, and that this betterness is of a moral type.

Some may object that this is simply rock-paper-scissors. But here the choice facing the good actor is not contingent on the action of anyone else, and the choice set is always pairwise. Alternatively, a transitivist might argue that these are three distinct 'better than' relations, pertaining to the three choice sets $\{A, B\}$, $\{B, C\}$, and $\{C, A\}$ respectively. This is akin to the argument that the *Intransitive American football teams* example included a criteria switch. There I argued that all criteria were included in all comparisons but that their force did change due to the nature of the comparison, but that this itself was not objectionable. Here it is not even clear what the supposed criteria switch would consist of. The relation being applied is entirely consistent in each case, that relating to the probability of arriving first, and each possible arrival time of one vehicle is considered against each possible arrival time of the other. This consistency across pairs is indicated by being able to use merely the phrase 'better than' in this context without clarification. Given the scenario, it seems clear what is meant by 'A is better than B', 'B is better than C', and 'C is better than A'.

Alternatively some might claim that while the relation is consistent, A, B and C are not. They might note that the consequence of choosing A when the alternative is B is different from the consequence of choosing A when the alternative is C, and this makes A different in the two cases. When the consequences are taken into account this might be thought of as like the choices {go to the park in car A, run someone over with car B}, {go to the park in car B, run someone over with car C}, {go to the park in car C, run someone over with car A} and no one would claim that this involves any interesting intransitivity. To see why we might object to this account, consider how this argument would apply in the competitive sport context. There it would claim that the *Intransitive American football teams* are not comparable because we would be taking the pairwise comparisons {A when it is expected to win, B when it is expected to lose}, {B when it is expected to win, C when it is expected to lose}, {C when it is expected to win, A when it is expected to lose}. Howsoever we understand the pairwise comparison and its relationship to 'better than' it seems

[6]Readers may recognise this example as relating to intransitive dice examples such as Efron's dice.

clear that the A in the comparison with B is the same as the one in the comparison with C, and that we can refer to it when discussing betterness. Likewise in the *racing evil-doer*, A is consistent across the choice sets — it is a vehicle that arrives at the target in 1, 6 or 8 hours with equal probability of a third. The expectation of arriving first/second is clearly a relevant property to the choice, but it is emergent from the pairwise comparison, not intrinsic to the item being chosen.

We might also consider what alternative 'better than' notion might apply instead. The most obvious alternative would be to claim that 'better than' is defined by the alternatives' expected performance against an unknown comparator. In which case all three are equal, since amongst the choice set $\{A, B, C\}$ they have equal probability of being first to reach the target against an unknown comparator. These alternative notions represent 'pairwise better than' and 'aggregate better than' in this context. As well as an evaluation of their properties, we might consider the competing notions of 'better than' in the iterative form of whether 'pairwise better than' is better than 'aggregate better than' or vice versa. I would contend that a plausible 'better than' notion is better than another if the choices it leads to are expected to be better. In the *racing evil-doer* example, if we take 'better than' to be a transitive relation then we may only advise that the alternatives are equally good. If the 'better than' relation is intransitive then we can advise the good actor that A is better than B, B is better than C, and C is better than A. She is therefore expected to make a choice leading to a better outcome when the intransitive 'better than' relation is employed. This is in contrast to many, perhaps most, situations in moral reasoning where employing an intransitive notion leads to significantly increased computational costs and perhaps even no decision, and making a decision itself has value, so that the transitivity of 'better than' may have value itself.

I have called arguments of this class *unambiguous arguments* because, in contrast to spectrum arguments, the intransitivity of the relation at the centre of the example is unambiguous. It is clearly not dealing with incommensurable, incomparable or vague alternatives, the intransitivity is not due to the application of a heuristic or a failure of intuition in the pairwise comparison, there is no ambiguity from the competing claims of *degree* and *kind*, and they are not open to challenge as Zeno's or sorites paradoxes. It is perhaps also informative in doubting our transitive intuitions that, despite being a mathematical fact, the intransitivity of the 'expected to arrive before' relation may be counter-intuitive for many.

2.7 Concluding Remarks

I have argued that 'better than' in the context of competitive sport is intransitive, either because it is defined by the relation 'expected to beat' or because it has some direct component of that relation in its nature. The sports context facilitated an explication for why we might challenge prominent arguments for the transitivity of 'better than' as an analytic or semantic truth. These challenges would seem to have relevance to wider considerations, and in particular led to the proposal of a family of betterness cycle examples, the so-called unambiguous arguments. Perhaps some would accept the plausibility of an intransitive 'better than' in the limited case where the fundamental unit of comparison is naturally pairwise and the qualities of the items being compared incontrovertibly interact in providing a direct comparison, but not otherwise. Such circumstances may be rare and this may be a contributory factor as to why some people's intuition in favour of a transitive 'better than' is so strong. But acceptance, even in this possibly limited context, provides a counterexample to the claim for the transitivity of 'better than' as a general truth and acknowledgement that the arguments for transitivity based on monotonicity principles and semantics are not insuperable. As well as this being important in its own right, it may have import for the way arguments based on other proposed betterness cycles, such as the spectrum arguments, are evaluated.

For the purposes of this thesis, the arguments presented in this chapter set a broad philosophical grounding to the exercise of ranking. In some sense, the arguments I have made here would seem to be antagonistic to that exercise. If 'better than' is an intransitive notion then it is impossible to represent the 'better than' relation in the uni-dimensional system that ranking, conventionally understood, imposes. However, I have argued that the most salient understanding of 'good' need not be consistent with the most salient understanding of 'better than'. Furthermore, I have made the claim that when a quality under consideration is intransitive then 'good' is derived from 'better than' and therefore that comparison is the appropriate means of understanding absolute quality. In Chapter 4, I will discuss why this might be the case from a behavioural perspective, but here the claim is of a more fundamental nature. Ranking can reasonably be applied as an ordering of items in terms of their absolute quality, even where that ordering may conflict with individual pairwise comparisons.

The notions of 'pairwise better than' and 'aggregated better than' discussed in this chapter were both framed in the context of expectations. In particular, the notion of 'aggregated better than' as represented by 'expected to aggregate more wins against all other competitors' is an idea to which we will return in Chapter 4,

when considering a model-independent parametrisation of quality. However, the use of expectations might imply the use of all data available in order to make the best possible predictions. This may not always be desirable. In the next chapter, we will discuss an example of ways in which an analyst may wish to constrain the ranking exercise to ensure accordance with values such as fairness and equity, as determined by the norms of the situation.

Chapter 3

Principled ranking and why the NCAA have got it wrong

Abstract

This chapter considers the question of ranking in incomplete league tournaments, of which the regular season of college basketball is perhaps the most notable example. The selection and seeding of teams for the annual NCAA Division I Men's and Women's basketball tournaments has taken on such a significance that it has its own day named after it, Selection Sunday. And yet the basis for these selections is highly disputed. Drawing on common practice in round-robin tournaments, a minimum set of principles that should guide quantitative analysts in performing generalized league ranking is proposed. In particular, the use of predictive measures and the inclusion of performance data other than wins is opposed. It is shown that the current NCAA method is highly deficient with respect to these principles, and it is argued that an approach consistent with these principles should be adopted. Further recommendations are made for future ranking.

3.1 Introduction

On 12th March 2021, two days before the final selection and seeding for the NCAA Division I (DI) men's basketball tournament, the Chair of the Selection Committee, Mitch Barnhart, gave an interview. In response to a question about the impact of COVID-19 season interruptions, he summarized the bases for selection and seeding, "It comes down to the foundational piece: who'd you play, where'd you play, and

what was the result." (NCAA, 2021) Yet it's not clear that the methods employed by the NCAA really do align with that statement. The intention in this chapter is to use the NCAA tournament as a case study for examination of the methods and principles involved in such ranking exercises, providing insight into the nature of tournaments, rankings, and the identification of athletic excellence.

In many kinds of league tournaments around the world, rankings are determined through ordering teams by total points, where points are awarded to teams dependent on the outcome of each game, or equivalently, by win percentage in binary outcome sports such as basketball. An underlying assumption is that the schedule strength of one team in the league is the same as, or sufficiently similar to, that of others. Outside of North America, the round-robin format predominates, in which every team plays every other team an equal number of times, often both home and away. North American leagues tend to be larger, precluding round-robin formats. To minimize differences in schedule strength, most North American leagues have a conference or divisional format, where teams are compared directly to others within their conference, who typically play each other more often and have comparable schedules.

However, sometimes equity is not possible. Recently, many leagues and tournaments were interrupted or modified by the COVID-19 pandemic. In other cases, including the regular season of NCAA DI basketball, unbalanced schedules derive from constraints of a logistical, historical, or commercial nature. There is a noticeable gap in the philosophy of sport literature concerning the ranking in these common tournament formats, what will be referred to as *generalized league tournaments*. For example, Pakaslahti (2019) explores the methods of tie-breaking in round-robins, but it is not clear how the arguments would be applicable to unbalanced tournaments. And Smead (2019) discusses the application of Arrovian social choice theory (Arrow, 1963) to sports, but social choice theory takes as its basic unit complete rankings of all comparators rather than the incomplete pairwise comparisons of games in an unbalanced tournament, leading to some principles being not readily translatable. Some scholars, like Bordner (2016), have resisted any conventional ranking done for purposes of esteem, advantage, or tournament participation. We disagree, and hold that ranking, when done well, can meet adequately the 'structural goals' of sport "to measure, compare, and rank competitors according to athletic performance" (Loland, 2013, p.53), as well as the 'intentional goals' - "the subjective reasons leading individuals to participate in sport," (Finn, 2009, p.70) such as commercial interests, or the sense of engagement with a community.

This chapter also draws on the foundational sports philosophy work, *The Grasshopper: Games, Life and Utopia* (Suits, 1978), where game playing is defined by four

elements. The first element is the goal, which itself is split into two parts, the 'lusory goal' of winning, and the 'pre-lusory goal', the state of affairs that a competitor seeks to bring about. For example, in golf we might describe the pre-lusory goal as getting the ball in the hole, or in a race as crossing the line. The second element is the means. These are necessarily restricted. The most efficient way of getting the ball in the hole may be to pick it up, hop on a golf buggy and drive to the hole. Permitted means are called 'lusory' means and prohibited ones 'illusory'. The third element is the rules. These define the means that may be employed to achieve the pre-lusory goal. Within the rules, Suits includes the informal rules of how to play a game well, which might be handed down by a coach, for example. The final element is what he calls the 'lusory attitude', the acceptance of the rules and the goals. Under this understanding, cheating represents a rejection of the 'lusory attitude' and so a competitor who does so is no longer playing the game. These elements are brought together in the following definition:

> "To play a game is to attempt to achieve a specific state of affairs, using only means permitted by the rules, where the rules prohibit use of more efficient in favour of less efficient means, and where the rules are accepted just because they make possible such activity." (Suits, 1978, p.43)[1]

While there is widespread controversy about much of this framework, especially in the nature of the rules and the interpretation of cheating, the construct of 'lusory' and 'pre-lusory' goals has wide acceptance and is the one used in this work.

3.1.1 NCAA DI basketball

The primary structural goal in NCAA DI basketball is in identifying an overall winner. The season is split into two parts — a regular season that consists of what may be thought of as a generalized league tournament; and a post-season, a standard six-round knock-out tournament. Approximately half of the teams qualifying for the post-season do so automatically by winning their conference championship during the regular season. The others are determined by a selection committee, based on teams' performance in the regular season. In making this determination, the committee are mandated to look at particular quantitative metrics. Historically, they used the Ratings Percentage Index (RPI), a heuristic measure that aimed to account for the strength of opposition that a team had faced (for more details, see Section 1.7.4). For the 2018/19 season, after sustained criticism of the RPI, it was replaced with the

[1]Suits (1978) also provided the somewhat pithier, and oft-quoted, characterisation that 'playing a game is the voluntary attempt to overcome unnecessary obstacles.'

machine-learning based NCAA Evaluation Tool (NET) and the Quadrant System, a categorisation of results that aimed to capture the role of venue (home/away/neutral) and strength of opposition in the performance assessment.

3.1.2 Aims

The argument presented here is that fairness in ranking of generalized league tournaments is important and can be usefully guided by principles, which may be derived from the norms of sports competition. Where these generalized league tournaments are a part of a wider tournament, such as in NCAA DI basketball, our contention is that if methods are available that meet these principles as well as the structural goals of the wider tournament, then such methods should be preferred.

The aims for this chapter are thus twofold. The first aim is to propose and defend a minimum set of principles to guide ranking methods. These will be applicable where the ranking itself constitutes the tournament outcome or where the purpose is the identification of participants for championship tournaments or post-season games. These principles are that a ranking should:

1. be anonymous — a team should not be (dis)advantaged due to their identity;
2. reflect a positive response with respect to the beating relation — a win is better than a loss against the same opposition;
3. depend on current season games only;
4. have no recency weighting to the evaluative weight of games;
5. be based solely on wins and losses;
6. adequately account for strength of opposition;
7. adequately account for venue.

The claim is not that the principles would exclude all potentially objectionable ranking methods in all scenarios. Instead, it is that they may be used to guide ranking, so that failing to meet them should raise questions about the ranking approach. The second aim is to assess the methods used by the NCAA in the administration of their annual basketball tournament. Given the popularity of the tournament, this question is of interest itself, but also demonstrates that the principles are not so general as to be uninformative.

The role of wins is of particular interest; round-robin tournaments (in binary outcome sports) are typically decided on wins, but as Torres and Hager (2005, p.211) notes: "The philosophy-of-sport literature is replete with either condemnations of the obsession with winning so prevalent in sport communities across the world or reminders that winning is neither the only value that matters in competitive sports nor the most important," and tournament administrators, such as the NCAA, resist employing wins uniquely in generalized tournament ranking. We contend that in sports where wins alone are customarily used for ranking in balanced tournaments they should remain the guiding principle even in unbalanced tournaments, and not other metrics that may reflect a broader understanding of athletic superiority. In particular, we find that prediction-based ranking that includes performance data other than wins alters the lusory and prelusory goal of the sport and violates other basic features of best ranking practices. The discussion here considers primarily sports where the outcome is a binary win/loss, rather than points. This aids consistency, interpretability and brevity, and is the case in basketball, which provides the main working example. However, many of the arguments are readily adaptable to sports where it is points rather than wins that are customarily the outcome of an individual contest, with points then becoming the definitive data.

The chapter proceeds in Section 3.2 by discussing round-robin tournaments — why they might be regarded as a useful base for inferring principles for generalized tournaments, and how the round-robin ranking method might be justified. In Section 3.3, this analysis is built on to derive seven principles which it is claimed a generalized ranking method should meet. In Section 3.4, the NCAA basketball ranking method is judged against those seven principles and evidence found that it violates all seven. Finally, some concluding remarks are made.

3.2 Round-robin tournaments

The round-robin tournament is a useful starting point for the consideration of generalized league tournaments. If it can be established that the standard ranking approach in round-robin tournaments is appropriate in that context then it can provide two insights for generalized ranking. First, if there are particular principles that such an approach meets then these may be applicable to generalized ranking. Second, if a preferred round-robin ranking method can be identified then it is desirable for the generalized ranking method to give the same ranking when applied to a round-robin results set.

There is a clear norm as to how a round-robin ranking method should work, by ranking participants in order of their number of wins. However, the ubiquity

of the practice alone does not confirm its efficacy, so we must consider on what grounds this ranking method might be justified. This question was addressed by Rubinstein (1980). He takes a round-robin tournament composed of a single round with binary win/loss game outcomes and considers three axioms. Following the notation of Rubinstein (1980), $A \to B$ denotes 'A beats B', and $A \succeq(T)\ B$ denotes 'A is ranked higher or equal to B in tournament T'. The axioms are:

a Anonymity – Let T be a tournament, σ a permutation on the set of teams N, and i and j teams. Denote by σT the tournament which relabels the teams so that $\sigma i \to \sigma j$ in σT if and only if $i \to j$ in T. Then $i \succeq(T)\ j$ if and only if $\sigma i \succeq(\sigma T)\ \sigma j$.

b Positive response with respect to the beating relation – Suppose i and j are distinct players in T and $i \succeq(T)\ j$. Let T' be identical to T except for the existence of a third player k such that $k \to i$ in T and $i \to k$ in T'. Then $i \succeq(T')\ j$.

c Independence of irrelevant alternatives – Let i, j, k and l be four distinct players. Suppose T and T' are identical, except $k \to l$ in T but $l \to k$ in T'. Then $i \succeq(T)\ j$ if and only if $i \succeq(T')\ j$.

Rubinstein (1980) demonstrates that under these axioms the unique ranking method is that defined by ordering teams by their number of wins. Henriet (1985) extends the axioms and the result to include ties, and Nitzan and Rubinstein (1981) extends to round-robin tournaments with multiple rounds. Van Den Brink and Gilles (2000) provide an alternative axiomatisation also leading to a ranking by number of wins.

A second argument for ordering by wins is based on its being the accepted norm, but notes that such norms might have particular force given the normative nature of the outcome of a tournament. There appears to be high acceptance of the combination of round-robin tournaments and ranking teams by their number of wins, as evidenced by four factors: ubiquity — it is used worldwide across sports; uniformity — almost everywhere it is used in a similar format with a sufficiently fair distribution of opposition, with structural controls for variables that may meaningfully influence the outcome (e.g. venue and neutral officiating); undisputedness — the outcomes of these tournaments are very rarely disputed as they relate to the structure or ranking method; certainty — no-one, not even the most ardent statistician, proposes that points accumulation should be presented with some sort of error bar to indicate to what degree there is uncertainty in the points aggregation as an expression of quality.

As was noted above, established norms need not be decisive in the question of how we ought to perform ranking, but there would seem to be philosophical and

pragmatic grounds for emphasizing them in the present context given the desire to achieve the 'intentional goals' of the sport. People would be less willing to watch post-season games if they did not believe that the teams they were seeing were deserving, and a coach would be less likely to get the sack for finishing down the rankings if the team would be held in no lower regard for doing so. Other outcomes — financial incentives from further games or professional incentives from job security — are derived from this prestige. In this way, prestige related to the determined ranking is required for a tournament to realise these 'intentional goals' of the sport. Since prestige exists purely as a function of the opinions of stakeholders, the degree to which a tournament achieves these 'intentional goals' is based on the acceptance of the tournament results by stakeholders. Therefore, ranking methods that have widespread acceptance are due special regard in the question of appropriate ranking methods.

As an established and hugely popular tournament, some might claim that NCAA DI basketball is an example itself of a tournament where the ranking method has achieved popular acceptance. But while there is generally no popular dispute about the winner of March Madness, there is regular controversy around the selection for the post-season tournament. Indeed the move to use NET and then amend it, and the expansive list of alternative ranking methods (see, for example, Massey (2019)) are evidence of this controversy and the lack of a generally accepted method.

3.3 Generalized ranking

A generalized tournament here refers to any fixed period league-based tournament or sub-tournament where the games are between two competitors. The round-robin format is a subset of generalized tournaments. As argued above, there are both axiomatic and normative grounds for taking total wins as the ranking method in round-robin tournaments. But it seems less clear how one ought to rank in unbalanced tournaments.

A number of generalized ranking methods that accord with round-robin ranking have been proposed. For example, one might consider a statistical generalization where game outcomes are taken to be independent conditional on the team strengths, with the consistency to the round-robin tournament achieved by requiring that those strengths depend on the results only through the number of wins that each team achieves. As was shown in Section 1.2.4, this leads uniquely to the well-known Bradley-Terry model (Bühlmann and Huber, 1963). Alternatively Slutzki and Volij (2005, 2006) describe sets of plausible axioms, similar to those of Rubinstein (1980), that lead to what they call the 'Fair Bets' model, originally described by Daniels

(1969) (see Section 1.7 for further details); or Chebotarev (1994) describes a family of models meeting particular axioms that he calls the 'generalized row sum model'. González-Díaz et al. (2014) review a number of ranking methods against axiomatic criteria. But these works generally describe how particular methods accord with particular principles, rather than arguing for those principles. In some instances the principles are also more specific than we would be prepared to grant necessary. Instead, a broader minimum set of principles is argued for here.

3.3.1 Principle 1: Anonymity

Given the desired consistency with round-robin ranking, we start by considering the axioms employed by Rubinstein (1980). First, the condition of anonymity requires that no team has an advantage in the ranking because of who they are. Perhaps some might dispute this in the context of a play-off system, believing that a team with a history of raising their game in the post-season should have a greater claim to a spot, even given an equivalent or lesser playing record in the regular season. But this sort of structural advantage seems unfair. It entails that another team is disadvantaged due to historic results, potentially that no member of the present team had a part in.

The mathematical definition of anonymity taken by Rubinstein (1980) is rather restrictive for the present context and not well-suited to the conference setting. With the conference system it would not be possible for a team to exchange their results in the way that the relabeling definition envisages. Conference membership, which is a key determinant of much of a team's schedule, is defined by history and geography and is largely inflexible. So here a broader definition is taken, that no team is (dis)advantaged by their identity, including a team's membership of a particular conference. This is consistent with arguments on the use of statistical methods in legal proceedings, where, for example, Jorgensen (2022) argues that algorithms used to inform legal outcomes should not depend on any 'unchosen properties'.

3.3.2 Principle 2: Positive response with respect to the beating relation

Positive responsiveness to the beating relation, broadly understood, seems to be an expression of the goal of competitive sport to win. The specification of Rubinstein (1980) states that a team that wins against a particular opposition is ranked at least as high as a team that loses against the same opposition if the rest of their record is the same. However, there may be alternatives. For example, it could be posed as:

2^*. Suppose i and j are distinct players in T and $i \succeq(T)\ j$ ($j \succeq(T)\ i$). Let T' be identical to T except for the existence of a third player k such that there is an additional match in T' between i and k, where $i \to k$ ($k \to i$). Then $i \succeq(T')\ j$ ($j \succeq(T')\ i$).

This specification states that the addition of a win (loss) to a record cannot result in a lower (higher) ranking.

2^* is stricter than the specification of Rubinstein (1980), since 2^* is a sufficient condition for the specification of Rubinstein (1980). To see this, suppose we have three tournaments T, T', T''. These tournaments are identical except that T' has one fewer game. In T there is an additional game where $k \to i$. In T'' there is an additional game where $i \to k$. If $i \succeq(T)\ j$, then $i \succeq(T')\ j$ by 2^*, and if $i \succeq(T')\ j$, then $i \succeq(T'')\ j$ by 2^*. Therefore, if $i \succeq(T)\ j$, then $i \succeq(T'')\ j$ and the specification of Rubinstein (1980) is met.

Here it would seem desirable to take the weaker condition. Notwithstanding that we later argue for a dependence solely on wins and losses, some might argue that a narrow win against a weak team ought, in some circumstances, to see a ranking reduced. Taking the condition of Rubinstein (1980) obviates this objection as a counter to the principle would require that there may be circumstances where a win would have a more negative ranking impact than a loss against the same opposition. Such a position seems indefensible.

3.3.3 Principle 3: Dependence on current season only

In considering a third principle, it is tempting to consider the third axiom of Rubinstein (1980). Independence of irrelevant alternatives may be appealing if one understands the quality of a team's performance to be a function purely of their own records. Building on analysis by Berker (2014), a closely related concept is argued for as 'autonomous relative ranking' (ARR) in a round-robin soccer context in Pakaslahti (2019), where it is taken as a decisive argument against using head-to-heads as tie-breakers. However, in the unbalanced tournament this principle need not hold. Consider four teams A, B and X, Y, where A and B have one win each against each other and X and Y have one win each against each other. The convention in round-robin ranking tells us that A and B should be ranked equal and X and Y should be ranked equal. But now suppose that $B \to Y$ and we are required to rank the teams. It would seem that A, B ought to be ranked higher than X, Y, which entails ranking A higher than X, as a result of games in which neither were involved, thus violating the axiom. Such dependence on 'irrelevant alternatives' in the context of an unbalanced tournament thus appears reasonable, even necessary. So, while we

might expect this axiom to obtain in the special case of a round-robin, it is not applicable to the generalized case. Thus, only the first two principles from Rubinstein (1980) — anonymity and the positive response with respect to the beating relation — are taken.

Instead, we may draw on three principles present as assumptions in Rubinstein (1980) in considering a round-robin tournament, but not presented explicitly as axioms. First, the ranking method should be dependent on performances in the relevant season alone. In a sense that is the definition of a season, that each team starts with a clean slate. Note that a 'season' here may have a wider interpretation than the conventional one, including contexts in which a 'season' would be a small part of a tournament, for example in the group stages of a World Cup.

There is a sense in many sports tournaments in which the results of previous seasons are considered. Promotion and relegation is common in sports tournaments in Europe and elsewhere. Suppose the eventual tournament winner is determined by a play-off between the top four teams in the top division of a regular season. In order to be able to qualify for the play-off and win the tournament overall, a team needs to be in that top division; something that will depend on their performance in previous seasons. However, conditional on the set of competing teams being defined at the beginning of the season these previous seasons' results play no role in the ranking. This would seem to be the relevant point here, where the set of competing teams is set, any ranking ought to take account solely of information from the current season.

3.3.4 Principle 4: No recency weighting to the evaluative weight of games

The second assumption implicit in Rubinstein (1980) is that game results are not treated differently due to when in the season they occurred. This principle may be controversial to many. Some may argue that in a play-off system more recent games should be weighted more heavily as these provide a better indicator of potential post-season success. The norm in most sports, including those in basketball and other professional North American leagues, is for there to be no weighting. Regular season games in balanced leagues could be weighted, but that is not done in any of the major professional North American sports or in the major European club tournaments in soccer and rugby union where play-off formats are employed. Pragmatically, weighting games disincentivises teams from caring about all games in the season equally, which may be to the detriment of sports competition, but it may also be unfair when it interacts with scheduling if, for example, a team is able to play weaker teams later in the season when they have greater weight.

In NCAA DI basketball, it is perhaps not so clear what the norm is. There appears to be no official acknowledgement that more recent matches carry more weight, much less an official argument that they should — Mitch Barnhart notably did not include "when'd you play" in his list of criteria. The quantitative metrics mandated to be used by selection committees are consistent with this observation; neither RPI nor its successor NET included any such recency weight. When determining the 32 qualifiers who qualify directly for the post-season based on winning their conferences there is also no recency weighting used in those conference determinations. Thus, adoption of this principle seems consistent with the sports norms we have argued ought to form the basis of the ranking method.

3.3.5 Principle 5: Based solely on wins and losses

A third implication of the Rubinstein (1980) construct that we seek to make explicit is that the team-based data on which the ranking will depend will be solely wins and losses. This is a central point of our argument and may be controversial to many. The quote from Mitch Barnhart is somewhat ambiguous on this point, using the term "result", which may be interpreted in several ways. The term 'team-based data' is used here to distinguish it from game features like venue — home/away/neutral — or strength of opposition, accounting for which is consistent with round-robin ranking and is discussed later. Instead the argument here is to exclude the use of additional information such as team demographic data, or team performance data such as possession or territory statistics, or even score margin.

In the context of a tournament like March Madness, some may see the goal as being to select the teams that are most likely to perform well in the post-season. Others may accept the primacy of wins and losses and the requirement for consistency with round-robin ranking, but interpret this as a requirement to determine what wins and losses would have been achieved were all teams to have faced each other. These are both predictive tasks, and, broadly speaking, prediction works better the greater the number of relevant factors one allows to be used in the method. Thus, the argument goes, relevant information is wrongly ignored by not considering factors outside of wins and losses. Alternatively, others might be in favour of a retrodictive approach, an evaluation of who has done best rather than who is expected to do best, but argue that such an evaluation should include performance data other than just wins and losses.

A potentially useful categorisation is to consider four levels of team-based data:

1. win/loss;

2. score;

3. performance data from matches played e.g. possession, territory, shots, fouls, net efficiency etc.;

4. non-performance data e.g. shirt colour, age, height, race etc..

Starting with the last of these, color of uniform (Hill and Barton, 2005; Attrill et al., 2008) or the proportion of players of particular heights, ages or races may be predictive features of a team's success, but few would advocate that they would be appropriate to include in the ranking of a team. Indeed, their exclusion is specifically mandated by the anonymity principle. The second and third categories are differentiated by score being defined in the rules of the sport as the means by which wins and losses are determined. In contrast, the non-score performance data are not a specified goal of the sport. Using them in ranking teams then becomes problematic. They have no intrinsic defined value from the rules of the sport, and their value can be viewed only actuarially — the degree to which they have been predictive of success in the past. Using these data for ranking purposes then alters the lusory and prelusory goals of the sport (Suits, 1978). The goal of a team becomes to most closely match the performance attributes of teams that have been successful in the past, rather than to win. As well as being objectionable in its own right, this might have the effect of actively discouraging tactical innovation, which is often lauded as a feature of sport, since such innovation might result in optimising to a different profile of performance metrics to those that have been successful in the past.

Using just scores or score margin would seem to avoid these objections but their use might still constitute misunderstanding the prelusory goal of the sport. For example, Pakaslahti (2019, p.358) states: "in a soccer match the betterness of teams is determined by how much ability to score and prevent goals (by using means that are permitted by the official rules and the ethos of soccer) each team demonstrated." In a soccer context, the *ability to score and prevent goals* is reflected as much in the difference between a 3-0 win and a 2-1 win as between a 1-0 win and a 0-1 loss, with the score margin being different by 2 in both cases. This could be expressed in a ranking system. It would be possible to rank in a round-robin tournament by aggregate goal difference, with number of wins perhaps acting as a tie-breaker. But tournament administrators, reflecting the view of their relevant sports communities, choose not to do that. We might reasonably conclude that they believe, as Herm Edwards famously expressed in what became a meme,"You play to win the game." That is, we might better understand betterness expressed thus: "in a soccer match the betterness of teams is determined by how much ability to score and prevent *more*

goals than their opposition (by using means that are permitted by the official rules and the ethos of soccer) each team demonstrated," with 'more' here being primarily an ordinal rather than cardinal evaluation. It is also notable that NCAA resisted the inclusion of an uncapped "score margin" in the algorithmic rankings that they have used to support their decisions, RPI and NET, on the grounds that "running up the score" may embarrass opponents and would conflict with the goal of sportsmanship (NCAA, 2018; Paul and Wilson, 2015).

Considering a predictive approach more directly, basing on prediction presupposes that ranking ought to be defined in terms of probability of winning in some sense, but this does not seem consistent with what we observe. An informative example is the 2015/16 English Premier League soccer season. Leicester City were the champions. But Leicester were not favourites in the betting markets for the next season, nor were they favourites in any of their games against the four other main contenders for the title that season — Arsenal, Tottenham Hotspur, Manchester City, Manchester United (football-data.co.uk, 2016). It therefore seems unlikely that people took Leicester City to be better than all the other teams in the tournament, in some sense consistent with 'most likely to win'. Yet they were fêted worldwide; no-one disputed who the champions were or expressed concerns that the tournament structure and ranking method had failed in meeting the requirements of ranking. Thus, the aggregation of past results was accepted as the appropriate model of ranking rather than a concept of betterness that would be modelled by predictive metrics. That unlikely events happen should not, on its own, preclude us from using predictive measures as a guide to ranking. But the example is informative in showing that when predictive measures and achieved performance lead to different ranking conclusions then it is the achieved performance that stakeholders in sport take to be decisive.

Another objection to limiting team-based data to a consideration of wins and losses is in the consideration of 'good' or 'bad' wins or losses. There would seem to be wide normative acceptance of the idea that a 'good' loss can be a better indication of quality than an 'ugly' win. Taking only wins and losses does not seem to allow for the identification of these. However, there is a strong norm in sports tournaments worldwide and in the USA, including in basketball in the NBA, for using rankings based solely on wins and losses and there is widespread acceptance of these rankings, despite there still being a widespread acknowledgement of the idea of 'good' and 'bad' losses in those contexts. Therefore, the acknowledgement of 'good' and 'bad' wins does not seem to be a constraint on using wins and losses to determine rankings.

However, if a stakeholder community feels it sufficiently important to acknowledge 'good' and 'bad' wins and losses then a principles-based approach does not preclude this. We have made the argument in this chapter in terms of wins and losses to aid

interpretability and maintain consistency with the work of Rubinstein (1980), but in the more general case it is points rather than wins and losses that are the relevant data. With that in mind, it can be noted that some sports do include factors other than just win/loss in their points systems. For example, in the NHL a team losing in overtime gains a bonus point, or in rugby union, points are awarded for scoring a certain number of tries or for losing within a particular score margin. Where this norm exists within a common sports community in the round-robin format it seems reasonable to maintain this points system in the generalized ranking through considering points rather than wins. In all these examples, the additional point-scoring factor is considered within the same dimension — two overtime bonus points in the NHL are equal to one win — with the other factors always earning less than the win, so that Principle 2, positive response with respect to the beating relation, will not be violated. When a points method of this nature is combined with a principle that we argue for later, that opposition strength should be adequately accounted for, then it could even be that a narrow or overtime defeat to a very strong team increases a ranking more than a wide win against a very weak team, under ranking methods concordant with the principles argued for here. Importantly under such approaches the lusory goal of winning and the pre-lusory goal of scoring more than the opposition are not compromised.

Relatedly, it should also be noted that principle-based methods relying solely on wins and losses will not automatically rate unbeaten teams over teams with losses on their record. If the strength of opposition is sufficiently different, then the unbeaten team may be ranked lower (see, for example, the discussion of using priors in Hamilton and Firth (2021), or realtedly the discussion of the use of penalties in Chapter 4).

Finally, methods that use prediction or data other than wins and losses may cause violations of some of the other principles. The example was given as to how non-performance data violates anonymity, but later with the example of NCAA DI basketball in Section 3.4, we show how the use of performance data may also cause violations to principles 2 and 3 — positive response with respect to the beating relation and dependence on current season games only.

3.3.6 Principles 6&7: Adequately account for strength of opposition and venue

Two final principles relate to the other explicit criteria that Mitch Barnhart cited — "who'd you play, where'd you play". In a round-robin tournament these are addressed by the balance of the tournament structure. A win(loss) against the strongest

team counts the same as a win(loss) against the weakest team. This works because everyone plays the same opposition, so conditional on having the same number of wins, a win against a stronger team is offset by a loss against a weaker team. In an incomplete league, the offset is not guaranteed by the schedule and so not accounting for opposition strength would allow teams to artificially inflate their rating by playing weaker teams. Here we describe the principle as 'adequately accounting for strength of opposition', by which we mean that a win(loss) against a stronger opponent will increase(decrease) a team's rating more(less) than a win(loss) against a weaker opponent. The methods discussed at the start of this section — Bradley-Terry, Fair Bets, generalized row sum model — are examples of retrodictive ranking models that achieve this.

Similarly, most balanced league tournaments account for home advantage through tournament design. Typically, teams play a similar number of home and away games overall, often with each opposition played an equal number of times home and away. This structure might be interpreted as an expression that a team's ability ought to be an average of their home and away strengths, or alternatively that we understand the home advantage to be a symmetric one where the advantage to a team A playing at home against a team B is equal to the advantage of team B playing at home against team A. These are alternative perspectives consistent with the round-robin norm. We are not prescriptive on this point but require that venue is adequately accounted for in line with some reasonable interpretation of the advantage that it confers.

3.3.7 Summary

An argument has been presented as to why the predominant ranking method used in round-robin tournaments ought to be respected. Based on that, seven principles have been derived with which generalized ranking ought to accord, namely that a ranking should:

1. be anonymous — a team should not be (dis)advantaged due to their identity;
2. reflect a positive response with respect to the beating relation — a win is better than a loss against the same opposition;
3. depend on current season games only;
4. have no recency weighting to the evaluative weight of games;
5. be based solely on wins and losses;

6. adequately account for strength of opposition;

7. adequately account for venue.

These are now used to assess the ranking method used in NCAA basketball.

3.4 NCAA Basketball

In this section, we compare the NCAA March Madness seeding and selection to the principles developed in Section 3.3. NCAA DI basketball consists of a regular season followed by a knock-out tournament commonly known as March Madness. The regular season consists of non-conference games, sometimes as part of tournaments, followed by a conference season, and finally conference championship tournaments. Due to disproportionate non-conference matchups and the varying size, strength, and format of conferences, team schedules can vary greatly. The March Madness tournament consists of 64 teams in the women's competition and 68 in the men's. In both cases, 32 teams are taken as champions of their respective conferences. The additional teams are selected by a committee, as are the seedings.

No details are given of the committee deliberations. However, the committee are mandated to use particular metrics in their deliberations. Until the 2018/19 season, the Ratings Percentage Index (RPI) provided a key instrument. This was often criticized for not respecting the positive response with respect to the beating relation, and for unfairly advantaging teams from the strongest conferences, due to overweighting the strength of a team's opposition, thus also for being readily gameable (Coleman et al., 2010; Baker, 2014). In response to these criticisms, for the 2018/19 season, the NCAA replaced RPI with the NCAA Evaluation Tool (NET), which used performance indicators. The model, which used machine learning techniques, was optimized against late season and NCAA tournament games as test sets (NCAA, 2018).

For the 2020/21 season, a new version was introduced with a reduced set of indicators:

> "...the NCAA Evaluation Tool will be changed to increase accuracy and simplify it by reducing a five-component metric to just two. The remaining factors include the Team Value Index, which is a result-based feature that rewards teams for beating quality opponents, particularly away from home, as well as an adjusted net efficiency rating. The adjusted efficiency is a team's net efficiency, adjusted for strength of opponent and location (home/away/neutral) across all games played."

These descriptions still leave significant lacunae in our knowledge. It could be that the method takes wins and losses in a manner consistent with the principled approaches already mentioned, and net efficiency is used only as a tie-breaker for example. However, the description of NET's development, as a model optimized against late season and NCAA tournaments, seems inconsistent with this understanding.

We have argued that rankings in generalized tournaments should be guided by the best practices of round-robin tournaments. The most basic problem for the NCAA's regular season ranking is that if a round-robin tournament were to be played, NET would likely give a different ranking to win percentage. But the NCAA provides a valuable case study for our principles more broadly. Since we know that net efficiency is applied as a significant factor, the NCAA's ranking most clearly violates Principle 5, that a ranking should be based solely on wins and losses, which we defended at length. This violation, in turn, creates forms of breaches of three other principles: anonymity, positive response with respect to the beating relation, and dependency on the current season results only. We now proceed to examine these three violations in more detail.

Net efficiency is a technical term in the context of NET with a specific definition (Jones, 2018). However, for the purposes of the argument here and without compromising that argument, it can be approximated by the ratio of net score margin to freneticism, by which we mean an approximation to the inverse of the average possession time.

$$\text{Net efficiency} \propto \frac{\text{Net score margin}}{\text{Number of possessions during game}} \propto \frac{\text{Net score margin}}{\text{Freneticism}}$$

Generally, net efficiency will be associated with higher propensity to win, since it is proportional to score margin. In order to see how a consideration of net efficiency might violate anonymity, let us suppose that each conference has developed its own independent culture of playing style. Conference X is unique in having a particularly frenetic style of play where pride is taken in a more aggressive defense, in turning over possession, and in displaying more stamina than the opponent. Scores are similar or higher than in other conferences, but net efficiency is lower. Stakeholders of other conferences prize patient build-up and accuracy more. When these teams meet in the post-season, teams from the more patient conferences are less successful in dealing with the more frenetic style, so teams from Conference X have an advantage. However, in general, teams with higher net efficiency will win — most games will be between two teams with a patient style, and even in games between a frenetic style team and a patient style team, differences in team quality will often be large enough to dominate any frenetic style advantage. In this way, higher net efficiency comes to

be associated with wins in the training of the algorithm and so yields a higher NET value, disadvantaging teams from Conference X in their NET rating. There may be no compelling reasons for them to change their playing style in response, as they will increase the probability of losing their conference games, and fans may object to the change in style on the grounds of it not being part of their team identity. As such, when teams from Conference X come to be compared to teams from other conferences under NET for the purpose of March Madness selection and seeding they may be disadvantaged by who they are with respect to the conference they play in, thus violating anonymity.

Now let us consider Principle 2, the positive response with respect to the beating relation. Suppose there are two tournaments T and T' where the wins and losses are identical except that $k \to i$ in T and $i \to k$ in T'. Now suppose that in T, all i's wins come by wide margins, and their losses by small margins so they have a high positive net efficiency, but in T' all their wins come by narrow margins and their losses come by wide margins so they have a high negative net efficiency. In T, team j has an identical record to i, the same wins and losses against the same opponents, but a very slightly lower net efficiency so that $i \succeq(T)\ j$ under NET. Team j achieves exactly the same results and net efficiency in T' as it does in T, so that in T' its net efficiency is much greater than team i, to an extent that despite team i's additional win, $j \succeq(T')\ i$ under NET, violating the positive response with respect to the beating relation. However, one might object that this assesses the adherence to the principle by taking T and T' to be identical in wins and losses (excluding the games between i and k), but that the ranking is dependent on net efficiency, so one must instead consider the net efficiencies as well when defining identical tournaments.

Such an objection violates Principle 5, that the ranking be based solely on wins and losses. But suppose we relax that condition and concede that T and T' should be identical in both wins and losses and net efficiency except for the games against k where $k \to i$ in T and $i \to k$ in T'. NET may still violate the positive response with respect to the beating relation. First, suppose, for example, that all teams play in either frenetic or patient conferences and that all games in the regular season are close so that signals as to the value of higher net efficiency are weak within conferences, but in the post-season frenetic teams always beat patient teams. An algorithm trained on these previous seasons might come to recognise lower net efficiency as being an indicator of higher ability. If this effect were sufficiently large and the score margin by which i beat k in T' were sufficiently large then the negative rating effect of the net efficiency increase could outweigh the positive rating effect of the win such that $i \succeq(T)\ j$ and $j \succeq(T')\ i$, with team i's game against team k being the only difference in the tournaments. Second, while we have characterised net efficiency as

being approximately proportional to score margin for the purpose of explication, the formal definition of net efficiency does not produce this exact proportionality, and indeed it would be possible for a team to win but achieve a lower net efficiency than their opponent. In this way we may have that $k \to i$ in T, but with k having lower net efficiency and $i \to k$ in T' but with i having lower net efficiency, so that if the weighting of net efficiency within NET were very large and positive then we could get that $i \succeq(T)\ j$ and $j \succeq(T')\ i$.

The above examples concerning anonymity and the positive response with respect to the beating relation also show how Principle 4, dependency on the current season only, is violated, since the results in one season become dependent on those in previous seasons. Even where net efficiency and wins are in a consistent direction, their relative influence in NET will depend on the results from previous seasons, since these comprised the test sets used to train NET. Indeed, it could be the case that a team in the current season would be ranked higher had they lost some game in the previous season, based on how that previous result impacted the training of the algorithm.

It is important to note that the NET is not definitive in NCAA seedings. It influences a ranking by being considered directly by committee members, but also in determining the Quadrant system that informs the committee's decision (Reinig and Horowitz, 2019). Under the Quadrant system each team's win-loss record is summarised by means of splitting opposition into four Quadrants based on their NET rating:

Quadrant 1:	Home games vs NET top 30, neutral vs top 50, road vs top 75.
Quadrant 2:	Home vs teams ranked 31-75, neutral vs 51-100, road vs 76-135.
Quadrant 3:	Home vs teams ranked 76-160, neutral vs 101-200, road vs 136-240.
Quadrant 4:	Home vs teams ranked 161+, neutral vs 201+, road vs 241+.

The equating of results within a Quadrant is itself objectionable. Teams rated equally in a round-robin tournament will have different ratings under the Quadrant system. For example, suppose we have a round-robin tournament where every team plays every other team twice, once at home and once away. If two teams A and B had each won one and lost one against teams ranked 31-75, but team A had won their matches at home and team B had won their matches on the road then this would yield different Quadrant system representations despite being identically rated under the round-robin ranking method. The Quadrant system also has an element of arbitrariness. A road win against the top ranked team is seen as equivalent to one against the 75th ranked team, or road wins against the first and 76th ranked

teams constitutes a worse record than road wins against the 74th and 75th ranked teams. There are not clear empirical or principled grounds for why a win on the road against the 75th ranked team is better than a win at home against the 31st ranked team. Even to the degree that this were true in any given season it would seem highly likely that there is variation across seasons. Thus the arbitrary treatment of home and away wins and of results against different opposition under the Quadrant system violate Principles 6 and 7.

Moreover, winning by a large margin might increase a team's NET rating, but will also decrease their opponent's NET rating, through its consideration of net efficiency. If that opponent were close to a Quadrant lower boundary then this might result in the winning team's Quadrant rating appearing worse, as it would then have a win against a team from a lower Quadrant. In such cases, a team may have a perverse incentive to minimize their margin of victory and therefore net efficiency. This reverses the prelusory goal of the sport, in requiring that fewer points are scored, and is also self-contradictory. If net efficiency is seen as suitably valuable to be included in the NET rating then we should not want situations where teams are incentivized to try to decrease it.

The analysis of the selection and seeding process against the principles has so far been based solely on an analysis of the quantitative approaches that contribute to the final ranking. As we have shown, of the seven identified principles, six are potentially violated under the NCAA's current approach — anonymity, positive response with respect to the beating relation, dependence on the current season only, dependence solely on wins and losses, and adequately accounting for opposition strength and venue. It could be that the deliberations of the committee members ameliorate these issues, but there is no evidence for that being the case and it may very well be that they are in fact exacerbated, with Reinig and Horowitz (2019) finding evidence that the Quadrant system is influential on their deliberations. The remaining principle, Principle 4, was that there should be no recency weighting to games. NET and the Quadrant system explicitly adhere to this (NCAA, 2020). However, one of the biggest controversies in the 2020/21 season involved the decision not to select Louisville despite them having a higher NET rating than other selected teams. Louisville's performances had been strong at the start of the season but weak at the end, which might suggest that the committee decision violated the remaining principle.

3.5 Concluding remarks

We have argued in this chapter for a principle-based approach to generalized ranking in meeting the structural goal of measuring performance. Seven principles were de-

rived from the common practices and axioms of ranking in the round-robin setting. Included among the principles is the central claim that wins and losses should be the defining factor; prediction-based ranking for the purpose of performance measurement compromises the goals of sport and violates other basic features of best ranking practice. We applied these principles to assess the ranking method applied in the NCAA DI basketball tournament. This highlighted the utility of such an approach, and it demonstrated the NCAA's ranking method as deficient, violating some, and potentially all, of the principles.

It is our contention that these principles should guide ranking in all generalized league tournaments. However, there can be competing goals acting on these tournaments, especially in the case where they act as a qualifying tournament. The primary goal of the overall tournament in these cases is in identifying the best team. The post-season knock-out format is widely accepted as a means of identifying a winner conditional on all realistic contenders being in that knock-out tournament. So the primary goal of the qualifying regular season may be understood by some as being to ensure that teams with the best chance of triumphing qualify. It may also be that there are intentional goals in the selection of teams for a post-season, in maximising public or commercial interest, that would compete with the principles we have set out. These goals may be more sharply in tension in a sport like college football where teams play fewer games in the regular season, making any quantitative assessment of their performance more arbitrary, and only one game in the post-season, meaning the validity of the final winner is more sensitive to the selection for that final game.

However, in the case of NCAA DI basketball, the presence of teams who could realistically be expected to win the overall tournament is likely to be ensured by the high ratio of the number of teams participating in the post-season tournament compared to the number of teams who may realistically win. Since March Madness was expanded to 64 teams in 1985, the lowest seed to win was seeded 8 (Villanova in 1985) and the lowest seeds to reach the final four were seeded 11 (George Mason 2006; LSU 1986; VCU 2011; Loyola Chicago 2018; UCLA 2021). With this being the case, then it is highly likely that a credible principle-based ranking would be capable of selecting contenders. On the other hand, there is evidence that the committee is influenced by the Quadrant system (Reinig and Horowitz, 2019), a patently misleading measure. Therefore, it is likely that every year teams miss out on playing in March Madness due to unfair ranking systems. Hence even for those who see the role of the regular season rankings to be identifying potential winners, a principle-based approach should be preferred.

In conclusion, we make four recommendations — two high level recommendations applicable to all tournament administrators, and two further recommendations

specific to NCAA DI basketball. The first of the high level recommendations is that tournament administrators should make explicit both the goals that they are seeking to meet and any principles supporting those. The second is that they should, as far as possible, be transparent about the method they intend to employ to meet those goals and principles. Enacting these would allow and encourage the relevant sports communities to openly debate the sort of questions addressed here, and for the ranking methods to therefore be better grounded. Transparency is not a topic discussed explicitly here but there would seem to be sound principles echoing those from a legal context that may apply. For example, Vredenburgh (2022) argues, in the legal context, for a right to an explanation, based in its necessity for protecting informed self-advocacy. Indeed arguments for transparency may have particular force given the normative outcomes of sport discussed earlier.

For NCAA DI basketball, we recognise that Selection Sunday itself meets intentional goals of the tournament in providing a focus for public engagement. We therefore moderate our recommendations from advocating for the replacement of committee decision by principle-based ranking to something that may be agreeable to a wider audience. First, we recommend the elimination of the Quadrant System. Since it is a categorical simplification of what NET or other ranking methods can capture in a more nuanced way, it serves only to arbitrarily misrepresent those data in a way that influences the outcome (Reinig and Horowitz, 2019). Second, we recommend the replacement of NET with a transparent ranking method, or methods, based solely on wins and losses and in line with the other principles.

As the COVID-19 pandemic highlighted, we anticipate more tournament administrators will have to address similar challenges in future. Unbalanced schedules may be necessary due to canceled games or shortened seasons. The continual expansion of leagues, especially in North America, preclude round-robin formats. These situations will call for the careful consideration of generalized principles, and not predictive metrics, for the purposes of official rankings in those sports. The type of principles we outline should be used to guide and constrain these deliberations, to be supplemented and informed by other factors agreed upon by stakeholders of a competition.

Methods have been discussed here in terms of broad principles, but it will be clear to readers that the Bradley-Terry model, the subject of much of this thesis, is consistent with the principles advocated, at least once an adaptation is made to account for venue (see, for example, Davidson and Beaver (1977) or Firth (2022)). Although it was not included in the seven identified principles, which were of a broader nature, the desirability of a selected ranking method returning the same as standard 'accumulated wins' ranking when applied to a round-robin tournament was

discussed. As the example of the round-robin tournament and PageRank in Section 1.7.1 demonstrated, this is not automatically the case even where ranking methods meet the other principles. Therefore, taken in conjunction with the arguments of Chapter 1, this chapter may be seen as highly supportive of the use of the Bradley-Terry model in a generalized tournament setting.

Chapter 4

Measures of reliability in Comparative Judgement

Abstract

Comparative Judgement is an assessment method by which item ratings are estimated based on rankings of subsets of the items. As an alternative assessment technique it has been important to establish the credibility of the ratings produced. In order to do this, studies have employed statistical measures to assess the reliability of those ratings. Those measures have been shown to be misleading under some conditions, and this has driven choices in how Comparative Judgement has been implemented. In this chapter those measures are discussed in more detail than has heretofore appeared in the literature. Relevant considerations are highlighted and alternatives proposed that address existing shortcomings.

4.1 Background

'Comparative judgement' (CJ) is a term used to describe a method of assessment by which a set of items are assessed based on rankings of subsets of the items via direct comparison. Most commonly the subsets consist of two items, with judges asked to provide a binary response indicating which item they consider to have more of some relevant quality, or simply to be better. The idea of using comparative judgement in this way builds on the idea that people are better at making comparative than

absolute judgements (see, for example, Goffin and Olson (2011))[1]. In this chapter, we focus on educational assessment, but the method can be applied more broadly to get ratings on any subjective quality.

A comparative approach to rating subjective qualities dates back to the work of Louis Thurstone (Thurstone, 1927a,b,c). He extended previous psychophysical work which had used comparison methods to investigate people's perceptions of physical properties such as loudness, weight or brightness to applying the methods to subjective questions such as 'seriousness of crime', 'attitudes towards gambling' and 'excellence of handwriting'. Bramley (2007, Sec. 2) provides an excellent summary of Thurstone's work and its development from a CJ perspective.

As well as the expansion in scope addressed by the method, Thurstone also proposed novel modelling techniques. In particular, Thurstone (1927a) proposed a model whereby the probability p_{ij} that an item i is preferred to an alternative j is given by

$$\text{Probit}(p_{ij}) = \lambda_i - \lambda_j,$$

where $\lambda_i \in \mathbb{R}$ is a rating of item i. This has become known as the Thurstone-Mosteller model after the elaboration in Mosteller (1951). More commonly in the CJ literature, the Bradley-Terry model (Bradley and Terry, 1952; Zermelo, 1929) is used to analyse these data, where

$$\text{Logit}(p_{ij}) = \lambda_i - \lambda_j, \tag{4.1}$$

or alternatively expressed

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j},$$

where $\lambda_i = \log(\pi_i)$, and will be referred to as the log-strength of i.

The widespread adoption of the Bradley-Terry model in the CJ literature is perhaps due to its form as a dichotomous Rasch model (Andrich, 1978), with the family of Rasch models being familiar to educational researchers, who have been active in employing Rasch models for analysis of education data. Alternatively, the Bradley-Terry model has also been presented as an analytically attractive approximation to Thurstone's model (see, for example Bramley (2007)). Stern (1992) showed that they are empirically highly similar. More broadly, as described in Chapter 1, the

[1]The website of the company most prominent in applying CJ assessments, `www.nomoremarking.com`, includes a very good demonstration of the advantages of comparative judgement under its Demo section. The assessor is shown a range of shades of purple, from very light to very dark, and then challenged to order them by using absolute and comparative judgement. Readers of this thesis are encouraged to attempt this task to get a sense for why one might be prepared to believe that comparative judgement is an advantageous method.

Bradley-Terry model has statistical appeal in being the unique model for which the number of 'wins' per item is a sufficient statistic (Bühlmann and Huber, 1963), and as the entropy and likelihood maximising model subject to the appealing retrodictive criterion that, for all items, the expected number of preferences given the comparisons made is equal to the actual number of preferences (Henery, 1986; Joe, 1988); these properties resulting from its identification as a full exponential family model (see Section 1.3.2).

4.1.1 CJ in practice

CJ has seen considerable growth in its use over the last two decades. The first systematic use in education was as a means of comparing the standards of English exam boards (D'Arcy, 1997; Pollitt and Elliott, 2003; Bramley, 2007). It has since been used to assess a variety of academic work including, for example, visual arts (Newhouse, 2014), engineering design (Strimel et al., 2017), mathematical proof comprehension (Davies et al., 2020), and translation (Han, 2022). Of particular interest to readers of this thesis, may be its use in tertiary Mathematics education as both a formative and summative assessment mechanism, especially in the context of peer-assessment (Jones and Alcock, 2012; Jones et al., 2013; Jones and Alcock, 2014; Jones and Wheadon, 2015; Jones and Sirl, 2017; Davies et al., 2020). However, its most widespread use currently is in English primary schools to assess writing.

Primary school writing offers a good example of an assessment task well-suited to the method (Wheadon et al., 2020). Children are given a visual prompt and are asked to respond to it through a piece of writing. This exercise is undertaken at approximately the same time across many schools. The pieces of writing are then scanned and uploaded for assessment. The assessment consists of teachers viewing pairs of responses, either both from their own school or both from other schools to avoid inter-school bias, and stating which they consider to be better. These judgements are aggregated and ratings applied to each piece of work by applying the Bradley-Terry model. There are a number of advantages compared to a traditional rubric-based marking approach in this setting. First, the marking of so many responses requires many judges. It would be difficult to elicit consistent judgements from so many judges under a rubric-based marking scheme, but the comparative method eliminates the influence of the variable severity of judges seen, for example, in two judges giving different marks to work of equivalent quality. Second, what is being assessed is holistic. It is hard to quantitatively define what is a good piece of writing for application under a rubric. CJ allows judges to consider a response's holistic merits (Pollitt, 2012b; van Daal et al., 2019) and negates attempts to game

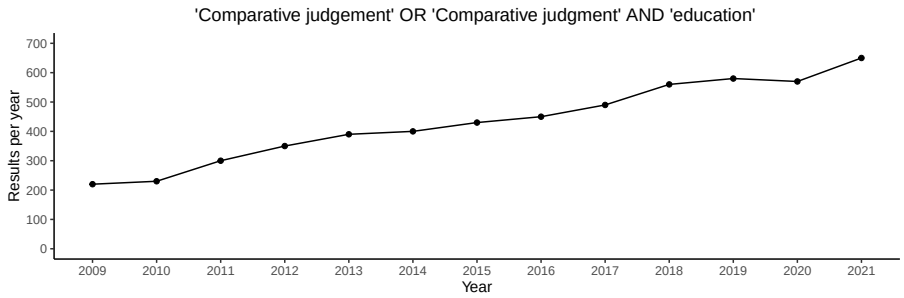

Figure 4.1: Google Scholar results by year for search "comparative judgement" OR "comparative judgment" AND "education". Data collected on 6th September 2022.

a marking scheme.[2] Third, the items are quick and cognitively easy to compare (Laming, 2003). Fourth, by including sample pieces of work from other assessment exercises, progress, as well as year-on-year and cohort comparisons, can be readily and interpretably assessed (Christodoulou, 2022).

4.1.2 CJ research

The increase in the use of CJ has seen a concomitant increase in its appearance in the research literature as can be seen in Figure 4.1. Much of this work has investigated its implementation in particular contexts, assessing its ability to reliably and efficiently rate the items under consideration (see Bartholomew and Yoshikawa (2018) for a systematic review of CJ use in the context of K-16 education in the USA, and Bartholomew and Jones (2021) for a systematic review of CJ use in higher education).

The method of scheduling the pairwise comparisons has also been a topic of investigation. In particular, Adaptive Comparative Judgement (ACJ) is a scheme for scheduling that has become popular and is embedded in one of the prominent commercial CJ platforms, Digital Assess. It was originally proposed by one of the leading advocates of CJ, Alastair Pollitt, formerly of Cambridge Assessment, with a claimed increase in the efficiency of the assessments, achieving ratings of equivalent

[2]An example given by Daisy Christodoulou, Director of Education at No More Marking, is that of GCSE English where a mark was awarded for correct use of a hyphen (Christodoulou, 2017). The response of one teacher was to train her pupils to use a character with the hyphenated name "Anne-Marie", thus gaining the mark, without the writer requiring any understanding of the correct use of hyphens. The author of this thesis remembers a similar piece of rote learning being employed to gain marks for the correct use of the subjunctive in the letter-writing part of their GCSE French exam!

reliability from fewer judgements (Pollitt, 2012a). The algorithm is described here as it is widely used in academic studies and in practice, and it provides useful context for the investigations that follow in this chapter. The method begins with a single round of random allocation – where each item is paired with one other – followed by three rounds of a Swiss tournament. Under a Swiss tournament scheme, in each round, items with the same (or as similar as possible) number of wins are paired. The items are then rated by fitting the Bradley-Terry model to the results set after these four rounds. These strength estimates are used to determine the next round of comparisons by pairing items i and j such that the absolute difference in estimated log-strengths, $|\lambda_i - \lambda_j|$, is close to a pre-specified 'gap'. This procedure of fitting the Bradley-Terry model after each round with the following round of comparisons scheduled according to their strength estimates is repeated until the desired number of rounds have been completed, or until an estimated reliability threshold is reached.

In the initial formulation, the intuition behind the selection of 'gap' began by considering the usual statistical concept of information, noting that the $(i, j)^{\text{th}}$ value of the expected information matrix, often known as the Fisher information, of log-strengths is

$$I_{ij} = -\mathbb{E}\left[\frac{\partial^2}{\partial\lambda_i \partial\lambda_j} \log p_{ij} \mid \boldsymbol{\lambda}\right] = p_{ij}(1 - p_{ij}). \tag{4.2}$$

Pollitt (2012a, p.163) goes on to note: "Information is maximised when $p_{ij} = 0.5$, but this makes the decision process difficult for judges: their decision will be easiest when the two portfolios are very different in quality and p is close to 0 or 1. We have chosen to optimise the assessment process by choosing a comparator for each portfolio so that p is approximately 0.67 or 0.33." This equates to a gap of approximately 0.7 logits. The validity of the claims for increased efficiency of ACJ when compared to random scheduling were challenged in a pair of papers that took a simulation-based and empirical approach respectively (Bramley, 2015; Bramley and Vitello, 2019). In response, Digital Assess conducted a further simulation-based investigation and a modified 'gap' of 1.5-2.5 logits was recommended (Rangel-Smith and Lynch, 2018). It also seems to have caused No More Marking to cease their investigations into using their own adaptive method (Wheadon, 2015b).

4.1.3 Reliability

Underlying all of these investigations are the measures used to assess the reliability of the CJ assessment exercises, and indeed the concept of reliability that underpins any measures. Ofqual, the English exam regulator, applied the insights from a comprehensive review of reliability (Tisi et al., 2013) to give this definition:

> ...consistency of examination results is referred to as reliability - the repeatability of results from one assessment to the next, be they assessments taken on different days, or from one year to the next.
>
> In everyday use, "reliable" means "that which can be relied on", but the technical definition in educational assessment is narrower: "the extent to which a candidate would get the same test result if the testing procedure was repeated." The technical definition of reliability is a sliding scale – not black or white, and encourages us to consider the degree of differences in candidates' results from one instance to the next.(Ofqual, 2013)

Hallgren (2012) emphasises the important distinction between *reliability* and *validity*. Roughly, we might say that validity is the degree to which an assessment assesses the understanding or skill it is intended to assess, and reliability is the degree to which it does so consistently, or as Hallgren (2012) states it, validity "assesses how closely an instrument measures an actual construct rather than how well coders provide similar ratings". For example, an electronically-marked multiple choice test on English grammar has high reliability as answers are either right or wrong, so there is no room for marker bias or error, but it may have lower validity if one is seeking to assess the ability of the examinee to correctly use grammar in written work. Proponents of CJ claim that the method can be well-suited for achieving high validity in some scenarios where alternatives, such as rubric-based marking, might struggle, though this is not a topic for investigation here.

The most frequently used statistic for assessing reliability in the CJ literature is the so-called Scale Separation Reliability (SSR).[3] It is based on classical test theory (Lord, 1959; Novick, 1966) where

$$\text{Observed Score} = \text{True Score} + \text{Measurement Error},$$

the idea being, in this context, that there is some 'true' underlying strength for each item that we seek to measure in performing an assessment.

Conceptually, it is perhaps questionable if such a 'true' strength exists. Similar to the argument made in Section 2.2 of Chapter 2, it is difficult to conceive what those fixed elements of merit would be. It seems highly plausible that what is generally valued in a piece of writing, say the relativity of correct spelling, good grammar, expressive vocabulary etc., changes over time and is dependent on the marker. One

[3]As Bramley (2015) notes, it is unfortunate that the equations given for SSR in the literature are frequently either wrong, including in three of the most highly-cited works (Heldsinger and Humphry, 2010; Pollitt, 2012b; Verhavert et al., 2019) , or not very clearly defined, as in Pollitt (2012a) or in the correction to Pollitt (2012b).

might understand those changes and marker preferences to be part of the measurement error, but the nature of the quality being assessed in CJ would seem to be inherently subjective. Indeed, the ability to deal with subjective assessment is a distinctive strength of the method. In practice, in educational research, 'true' score is often interpreted to be the score that would have been given by a Lead Moderator had they marked the item. In this setting, it could be interpreted in this way but that provides no way for the measurement error to generally be evaluated. It also seems at odds with the philosophy of the judgements being, in a sense, crowd-sourced. Instead, we might understand it as being the asymptotic score that would be achieved were time and the number of judgements not to be limitations. Even then, it is somewhat unclear which judges we would have doing these judgements and in what ratios, but this does seem to provide, at a high level at least, an interpretible and tractable conception of how the classical test theory might be applied in this setting.

Under the classical test theory conception, the SSR may be thought of as an estimate of a coefficient of determination, or R^2, the proportion of the observed variation of log-strengths that can be accounted for by the true variation of log-strengths, defined by

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(\lambda_i - \lambda_i^*)^2}{\frac{1}{n}\sum_{i=1}^{n}(\lambda_i - \bar{\lambda})^2}, \tag{4.3}$$

where λ_i^* is the 'true' strength for item i, λ_i is the estimate or observation of this strength, and $\bar{\lambda} = \sum_i \lambda_i / n$ is the mean of the λ_i.

Clearly, in a CJ assessment the 'true' strengths are unknown. SSR seeks to estimate the numerator of the fraction in equation (4.3) by taking the mean of the squares of the standard errors of the log-strength estimates. However, the standard errors of the log-strength estimates and the average of their squares are not well-defined, since the log-strengths are identifiable only up to a constant, as is apparent from equation (4.1). Nevertheless, there seems to be a consensus within the CJ literature. Pollitt (2012b, p.283), the most highly-cited paper in the CJ literature, asserts:

> "The *information* in each decision is calculated and summed over all decisions involving each script.
>
> $$\text{Information on script A: } I_a = \sum_{i \neq a}^{n} (p_{ia}(1 - p_{ia}))$$
>
> from which the standard error for the estimate of v_a is given by
>
> $$\text{Standard error for script A: } se_a = \frac{1}{\sqrt{I_a}}$$"

Thus, the standard error for the estimated log-strength of the item is defined as the inverse of the square root of the diagonal of the information matrix. This is incorrect. The standard error is found as the square-root of the diagonal of the inverse of the information matrix. In the notation of the expected information matrix defined in equation (4.2), under the Pollitt (2012b) method, $\left(\sqrt{I_{ii}}\right)^{-1}$ is taken as the standard error for item i, rather than $\sqrt{(I^{-1})_{ii}}$. Pollitt's definition appears to be the one applied in the vast majority, perhaps all, of the CJ literature and practice through its use in the **btm** function in the **sirt** package in **R** (R Core Team, 2021), which is used by No More Marking (Wheadon, 2015a), as well as in the Facets (Linacre, 2022a) package used in Bramley and Vitello (2019), in the study produced by a team from the Digital Assess platform (Rangel-Smith and Lynch, 2018) and in the code used in Crompvoets et al. (2020).

While SSR is by far the most common reliability statistic in the CJ literature, others also appear. Jones et al. (2013) and Bisson et al. (2016) introduced a 'split-half' procedure whereby judges were partitioned at the analysis stage into two approximately equal-sized sets and the Pearson correlation of the ratings produced by those two sets calculated. This procedure was repeated over a number of such possible partitions and summary statistics were presented. This measure has begun to be adopted more widely (for example, Jones and Wheadon (2015); Davies et al. (2020); Han (2022)). It is generally understood to be a measure of inter-rater reliability, the degree to which different judges agree in their assessment decisions.

However, we might reasonably understand any measure of reliability in this context to be a measure of inter-rater reliability. If we are to understand reliability as "the extent to which a candidate would get the same test result if the testing procedure was repeated" (Ofqual, 2013), then the variation in the test result may be due to one of three elements: variation in the performance of the candidate; intra-judge variation — variation in the assessments of any individual judge; and inter-judge variation — variation between the assessments of different judges. Given a single item of work from a candidate, we cannot consider the first of these. The second will be difficult to detect and even harder to quantify given the typical CJ data set. Cyclic chains of comparisons where, for example, A is preferred to B, B to C and C to A, may be detected. But on many data sets, the network of judgements by any individual judge will be extremely sparse and likely unconnected, so that non-detection of intransitivities is uninformative, and even the detection of an intransitivity may tell us little about how much variation there is in an individual judge's preferences. It also seems highly reasonable to suggest that the degree of variation between judges' assessments is likely to be significantly higher than the variation within any one judge's assessments, given natural trait-preference differ-

ences between judges. Split-halves might therefore be interpreted more broadly as a reliability measure in the CJ context.

Alternatively, Holmes et al. (2017) uses a Spearman rank correlation when comparing rankings produced by traditional rubric-based marking vs pairwise comparison vs an anchored rank order approach; Pinot de Moira et al. (2022) uses Krippendorf's alpha (Hayes and Krippendorff, 2007) to compare the attainment category applied through a traditional rubric-based marking vs lead examiner marking vs CJ by assuming that the number of members of each category should be the same as that under the rubric-based method. Additionally there have been a number of simulation-based studies, typically investigating adaptive scheduling procedures such as ACJ (Bramley, 2015; Rangel-Smith and Lynch, 2018; Bramley and Vitello, 2019; Crompvoets et al., 2021). In this situation, the true log-strengths are known, in being those used to simulate the data, and can be used to calculate a similarity statistic of interest.

Finally, we note that the amenability of any CJ assessment to an estimation of reliability as a byproduct of the assessment process should itself be understood as a strength of the method. Alternatives such as rubric-based marking generally do not have the density of judge-item networks to allow such a measure to be robust.

4.1.4 Parameter estimation

An important consideration in assessments of this kind that has gone unrecognised in the CJ literature is the estimation approach. Maximum likelihood estimation is standard practice in fitting model parameters in CJ, providing an asymptotically unbiased estimator under standard regularity conditions. However, when data is sparse, the maximum likelihood log-strength estimates may be materially biased, indeed, they will not even be finite if an item has been preferred (dis-preferred) in all its comparisons. The large number of items and relatively small number of judgements in many CJ assessment exercises means these issues occur with high frequency. Additionally, adaptive scheduling schemes, such as ACJ or Swiss tournaments, can further increase bias by accentuating differences at the extreme ends of the strength scale. In general, the CJ literature is frustratingly unclear on exactly what estimation methods are used and regrettably there does not appear to be a norm around making code and data available along with publication. Consequently, here the discussion is confined to investigating three estimation approaches associated with software that has appeared in the CJ literature, plus three others with particular appeal or relevance. These are all within the family of penalised estimation approaches. Later bootstrap bias-correction methods are also investigated.

4.1.5 Aims

The aim of this chapter is to review two of the central practices in CJ — parameter estimation and the measurement of reliability — and where appropriate to suggest better alternatives. These are important questions for the field as the current focus on SSR, and the requirement that it be an accurate indicator of reliability under current estimation methods, has caused the rejection of adaptive scheduling approaches by the largest platform, No More Marking, and a substantial dilution of the practice on one of its competitor platforms, Digital Assess (Rangel-Smith and Lynch, 2018). The consequent reduction in efficiency, the time taken for an assessment of equal reliability, may be holding back the approach from wider adoption (Pinot de Moira et al., 2022).

Section 4.2 introduces the estimation procedures that will be considered in this analysis, highlighting how they might be related in a generalised form, and selecting four for the later analyses. In Section 4.3, the SSR measure is reviewed. The effect on SSR of the interaction of estimation method with the scheduling scheme and underlying log-strength distribution is investigated. The section concludes with a discussion of the conceptual desirability of using SSR. In Section 4.4 the split-halves measure is discussed, noting its flexibility but also how it may underestimate reliability due to the loss of information inherent in the procedure. The conceptual desirability of the measure as most commonly implemented is also discussed. In Section 4.5, a bootstrap method for reliability estimation is introduced and compared to SSR. It is found to perform substantially better, though not to be a solution on its own. In Section 4.6, the use of a bias-corrected estimator is investigated. It is demonstrated how the approach improves the performance of the parameter estimation and thus of both estimated and achieved reliability. In Section 4.7, an alternative measure for reliability is introduced that addresses some of the critiques of other measures and may be calculated within the framework suggested by Section 4.5.

The investigations up to this point will be based on simulation studies. In Section 4.8, an empirical study is undertaken using a data set from Bramley and Vitello (2019). It is shown how the improved parameter and reliability estimation methods lead to radically different conclusions. In Section 4.9, the notable success of one of the estimation methods is discussed further, and cast in terms of a more detailed examination of the inference being performed and the importance of the schedule as an ancillary statistic. Section 4.10 presents some concluding remarks, including recommendations for current practice and possibilities for future research.

Notation is consistent with that used in Chapter 1 but will be reintroduced to aid readability and will be extended in Section 4.9. Given the specific context, the descriptions will use the terms: items, comparators, preferences, dispreferences and

comparisons in place of: teams, opponents, wins, losses and matches respectively. 'Tournament' will be used to denote a schedule of comparisons, not including the outcomes of those comparisons. The term 'strength' is used in a general sense such that as the 'strength' of an item increases the probability that it will be preferred in any comparison increases, but does not refer to any specific parametrisation.

4.2 Estimation

Various methods of estimation have been proposed for contexts relevant to CJ. For example, one stream of literature considers bias reduction in relevant model families (see, for example, Firth (1993); Kosmidis and Firth (2009, 2011); Kosmidis (2014); Kenne Pagui et al. (2017); Kosmidis et al. (2020); Kosmidis and Firth (2021)). Another, independent, stream of investigation considers estimation in the context of Rasch models, typically considering both bias and predictive accuracy (see, for example, Molenaar (1995); Linacre (2004); Haberman (2004); Robitzsch (2021)). A third stream considers Bayesian estimation and the selection of an appropriate prior within a Bradley-Terry model (see, for example, Davidson and Solomon (1973); Leonard (1977); Chen and Smith (1984); Whelan (2017); Phelan and Whelan (2017)). In this section, those literatures are leant on in reviewing six approaches. The first three approaches appear explicitly in the CJ literature. Next, two further methods are included that have particular interpretive appeal and relate closely in form to the three from the CJ literature. These have appeared in the literature on prior distribution selection with Bayesian estimation. Finally, the method from the most-cited work on bias reduction, Firth (1993), is also considered.

The score function under the Bradley-Terry model for an item r is

$$\frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda_r} = w_r - \sum_j m_{rj} p_{rj},$$

where $w_r = \sum_j c_{rj}$ is the number of observed preferences for item r over all other items and m_{rj} is the number of comparisons between r and j. Under maximum likelihood estimation, this is set to zero and the parameters estimated from this system of equations.

Under a penalised likelihood approach, a penalty term, a_r, is introduced into the score equation,

$$w_r + a_r = \sum_{j \neq r} m_{rj} p_{rj}.$$

This has the effect of ensuring the finiteness of estimates even where items have been (dis)preferred in all comparisons, and of providing shrinkage to the estimates, such that bias may be reduced. Here, in order to highlight the relative nature of the penalties, they will be compared in the generalised form,

$$a_r = \text{constant} \times (1 - 2\Omega_r),$$

where Ω_r is a function that can depend on the data or on the log-strengths of the items. Ω_r is chosen so that it increases with the strength of the item, such that a_r is negative for strong items and positive for weaker ones. Roughly speaking, different Ω_r will express different relationships between strength and the penalty, for example, linear or sigmoid, or related to the observed performance or the estimated strength, while the constant reflects the strength of the penalty.

The approaches will be presented individually and then further investigated by using simulation to calculate their bias and expected error under various conditions. Based on these considerations, four will be selected for further analysis in later sections.

4.2.1 ϵ-adjustment (Bertoli-Barsotti et al., 2014)

Many CJ studies have been performed using the No More Marking platform, which is made freely available to researchers. For example, a Google Scholar search for "comparative judgement" OR "comparative judgment" AND "nomoremarking" yields 99 results as at 6$^{\text{th}}$ September 2022. Wheadon (2015a) indicates that No More Marking use the **btm** function from the **sirt** package (Robitzsch and Robitzsch, 2022) in **R** (R Core Team, 2021) for their analysis. This function uses a bias reduction method proposed by Bertoli-Barsotti et al. (2014), which they call the ϵ-adjustment approach, where the number of preferences for item r, $w_r = \sum_j c_{rj}$ is adjusted by the offset term

$$a_r = \epsilon \left(1 - 2\frac{w_r}{m_r}\right), \tag{4.4}$$

where $m_r = \sum_j m_{rj} = \sum_j (c_{rj} + c_{jr})$ is the number of comparisons involving r, and ϵ is a constant, set to be 0.3 by default in **sirt**. The number of observed preferences, a sufficient statistic for the log-strengths, is therefore transformed from being in the interval $[0, m_r]$ to $[\epsilon, m_r - \epsilon]$ for each r, ensuring that log-strength estimates are finite. In our generalised form the constant is taken to be ϵ, 0.3 by default, and Ω_r is the proportion of comparisons involving item r in which it was preferred. Robitzsch (2021) found the ϵ-adjustment to be one of the best-performing of a wide variety of proposed estimation methods when considering the wider context of Rasch models,

looking at bias and root mean squared error. None of the other estimation methods discussed here was considered in that study, though the method of Warm (1989), which was considered and rejected, is closely related to that of Firth (1993).

4.2.2 Facets (Linacre, 2022a)

Another approach used by software that appears in the CJ literature is to be found in the commercial package Facets (Linacre, 2022a). Facets was used for the data analysis in Bramley and Vitello (2019), which was an empirical follow-up study to Bramley (2015). Based on analysis presented in Section 4.8, it appears that the model in Bramley and Vitello (2019) was fitted without penalty, with an ad hoc approach used to address issues of finiteness, but as a software tool used in education research, its penalisation approach is still of interest, and is a useful example of a somewhat intuitive method that fails to work well. The approach taken is described in an article on the website for the product (Linacre, 2022b). It introduces two dummy items, a hypothetical 'best' and 'worst'. Each item is then assumed to have been preferred once in comparison with the 'worst' item and dispreferred once in comparison with the 'best' item. The 'best' and 'worst' items are assigned log-strengths of 10 and -10 respectively. There seems to be no empirical or theoretical basis for the choice of these values, with Linacre (2022b) merely stating: "[l]et's hypothesize that a reasonable logit distance between those two hypothetical performances is, say, 20 logits". The likelihood is therefore augmented by a multiplicative term

$$\prod_i p_{bi} p_{iw},$$

where b and w denote the dummy 'best' and 'worst' items respectively, and the log-strengths of the dummy items are set to

$$\lambda_b = 10, \lambda_w = -10.$$

This translates to an additive term in the log-likelihood of

$$\sum_i \lambda_b - \log(e^{10} + e^{\lambda_i}) + \lambda_i - \log(e^{\lambda_i} + e^{-10}),$$

so that the Facets adjustment is equivalent to an adjustment to the number of preferences of

$$a_r = 1 - \frac{e^{\lambda_r}}{e^{10} + e^{\lambda_r}} - \frac{e^{\lambda_r}}{e^{\lambda_r} + e^{-10}} = 1 - p_{rb} - p_{rw}. \tag{4.5}$$

In the generalised form this therefore takes the constant as 1, and Ω_r as the average of the probabilities of item r being preferred to the dummy 'best' and 'worst' items. Rangel-Smith and Lynch (2018) report observing standard deviations of the log-strengths of between 0.9 and 3.6 logits in CJ exercises. At this level, the Facets approach offers a very minimal penalty, addressing the issue of finiteness but not of bias. Figure 4.2 gives an example of this. It shows the bias under the three penalisation methods mentioned in the literature for 15-round randomly scheduled tournaments with normally distributed log-strengths. As we will see later, this represents a 'well-behaved' example, a combination of log-strength distribution and scheduling scheme where estimation methods generally perform well, and yet the bias is close to a third of the log-strengths for the majority of the items. In the context of this work, the weakness of the penalisation in the Facets method is useful in giving an indication of the sort of bias we might expect with no penalisation, while dealing with the problem of finiteness of the log-strength estimates.

4.2.3 All v All (Crompvoets et al., 2020)

Another approach from the CJ literature appears in the Supplementary Material to Crompvoets et al. (2020). There it is assumed that each item is preferred against every other item an additional 0.01 times, augmenting the likelihood with the multiplicative term

$$\prod_{i \neq j} p_{ij}^{0.01},$$

giving an additive term in the log-likelihood of

$$0.01 \sum_{j \neq i} \log p_{ij} = 0.01 \sum_{j \neq i} \lambda_i - \log(e^{\lambda_i} + e^{\lambda_j}),$$

and therefore a score penalty of

$$a_r = 0.01 \sum_{j \neq r} \left(1 - 2\frac{e^{\lambda_r}}{e^{\lambda_r} + e^{\lambda_j}}\right) = 0.01(n-1)\left(1 - 2\frac{\sum_{j \neq r} p_{rj}}{n-1}\right).$$

In the generalised form, Ω_r is then the average probability of being preferred when compared to all other items. The constant is dependent on the size of the set of items. This seems undesirable. It provides a bias reduction method that depends on the total number of items, so that shrinkage will increase with the number of items. An example of this can be seen in Figure 4.2, where with just 250 items the penalty becomes large enough that there is very substantial over-shrinkage. However,

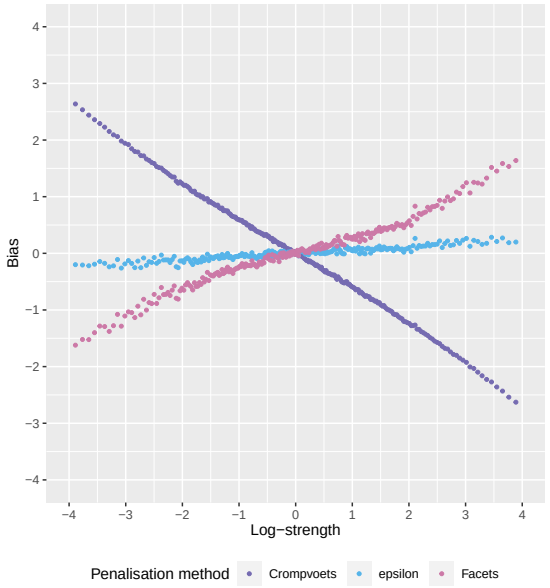

Figure 4.2: Bias under penalisation methods appearing in the CJ literature. Simulation based on 250 items, with normally distributed log-strength in a 15-round randomly-scheduled tournament. Method of Crompvoets et al. (2020) shows substantial over-shrinkage. Facets penalisation shows substantial under-shrinkage.

using the average expected number of wins against all comparators seems a sensible suggestion for Σ_r in the generalised form and leads to the following proposal.

4.2.4 α-adjustment

Consideration of the method of Crompvoets et al. (2020) suggests an alternative, where the results are augmented by assuming that each item has been preferred to each other item $\alpha/(n-1)$ times. This implies a score penalty of

$$a_r = \alpha \left(1 - 2\frac{\sum_{j \neq r} p_{rj}}{n-1}\right). \tag{4.6}$$

This method will be referred to as the α-adjustment approach here. In the generalised form, the constant is then α, to be determined, and Ω_r is the average probability of item r being preferred in comparisons with all other items. With $\alpha = 1$, this represents a form of Laplace's rule of succession, with an observed ratio of preferences to comparisons of $(w_i{+}1)/(m_i{+}2)$ for all items i. More generally, it can be understood

as a Bayesian estimation with a conjugate Beta$(\alpha/(n-1)+1, \alpha/(n-1)+1)$ prior on each of the pairwise preference probabilities.

It might also be interpreted as an information-based penalty in the sense that the likelihood is augmented by a multiplicative term that is a function of the pairwise information,

$$\prod_{i,j} p_{ij}^{\alpha/(n-1)} = \prod_{i<j} (p_{ij}(1-p_{ij}))^{\alpha/(n-1)} = \prod_{i<j} (I_{ij})^{\alpha/(n-1)}.$$

4.2.5 Dummy item (Phelan and Whelan, 2017)

Phelan and Whelan (2017) propose a method that takes a dummy item of average quality, against which each item is preferred and dispreferred an equal number of times, c_0, which need not be an integer. This method seems to have been used in early implementations of the Bradley-Terry model applied to rank teams in college hockey in the USA with c_0 taken to be a half (Wobus, 2007). As Phelan and Whelan (2017) note, this is equivalent to taking a Bayesian approach where the prior is a Beta(c_0+1, c_0+1) distribution on the probability of an item of zero log-strength being preferred to an item i. In the 'dummy item' approach the likelihood is therefore augmented with a multiplicative term

$$\prod_i p_{i0}^{c_0} (1-p_{i0})^{c_0},$$

where 0 denotes the dummy item, and p_{i0} the probability of item i being preferred to the dummy item. The log-strength of the dummy item is set to zero, which also provides an identifiability constraint. The penalty translates to an additive term in the log-likelihood of

$$\sum_i c_0 \lambda_i - 2c_0 \log(1+e^{\lambda_i}).$$

The dummy item adjustment is therefore equivalent to a score penalty of

$$a_r = c_0 \left(1 - \frac{2e^{\lambda_r}}{1+e^{\lambda_r}}\right) = c_0 \left(1 - 2p_{r0}\right). \tag{4.7}$$

In the generalised form, the constant is therefore taken to be c_0, and Ω_r is the probability of item r being preferred to the dummy item, an item of zero log-strength. As with the α-adjustment, with $c_0 = 1$ this is a form of Laplace's rule of succession, but here the additional preference and comparisons are added by means of the dummy item rather than additional comparisons with all other items.

4.2.6 Firth (1993)

The most cited work on the topic of bias reduction of maximum likelihood estimates is Firth (1993). It notes that previous proposals to use a jackknife method (Quenouille, 1949, 1956) or 'corrective' bias-correction rely on estimates being finite, but this cannot be guaranteed in general. Instead a 'preventive' correction is proposed, based on eliminating the "1/n" bias term (see, for example, McCullagh and Nelder (1989, p.455-456)). In the context of exponential family distributions, such as the Bradley-Terry model, this may be described by an adjustment to the log-likelihood by the log of the square-root of the determinant of the information matrix, $i(\boldsymbol{\lambda})$,

$$l^*(\boldsymbol{\lambda}) = l(\boldsymbol{\lambda}) + \frac{1}{2}\log\left|i(\boldsymbol{\lambda})\right|,$$

which is alternatively viewed as the Jeffreys (1946) invariant prior for the problem. Note that here, we have not specified if this is observed or expected information, since for the Bradley-Terry model, conditional on the comparisons observed being an ancillary statistic, they are equal. The implications of the schedule not being an ancillary statistic, as is the case under adaptive schedules such as ACJ or Swiss tournaments, is a topic that will be discussed at more length in Section 4.9.

In terms of the score function, the penalisation proposed is an additive term to the the number of preferences w_r of

$$a_r = \frac{1}{2}\mathrm{tr}\left\{i(\boldsymbol{\lambda})^{-1}\left(\frac{\partial i(\boldsymbol{\lambda})}{\partial \lambda_r}\right)\right\}. \tag{4.8}$$

Firth (1993) observes that the method is equivalent to solutions of the maximum likelihood equations on adjusted data of $c^*_{ij} = c_{ij} + h_{ij}/2$ and $m^*_{ij} = m_{ij} + h_{ij}$, where h_{ij} is the leverage of comparisons between items i and j, a measure of how far away the observation of pair i, j is from other observations, derived from the hat matrix. The 'adjusted data' score equation is thus

$$w^*_r = \sum_{j\neq r} m^*_{rj} p_{rj}$$

where

$$w^*_r = \sum_{j\neq r} c^*_{rj} = \sum_{j\neq r}\left(c_{rj} + \frac{h_{rj}}{2}\right) = w_r + \frac{1}{2}\sum_{j\neq r} h_{rj}.$$

The score equation being solved in the maximum likelihood estimation would then be

$$w_r + \frac{1}{2}\sum_{j\neq r} h_{rj} = \sum_{j\neq r} p_{rj}\left(m_{rj} + h_{rj}\right).$$

Rearranging into the generalised form presented in this section,

$$w_r + \frac{1}{2}\sum_{j\neq r} h_{rj}\left(1 - 2\frac{\sum_{j\neq r} p_{rj}h_{rj}}{\sum_{j\neq r} h_{rj}}\right) = \sum_{j\neq r} p_{rj}m_{rj}.$$

So that for an item r, the penalty term is

$$a_r = \frac{1}{2}\sum_{j\neq r} h_{rj}\left(1 - 2\frac{\sum_{j\neq r} p_{rj}h_{rj}}{\sum_{j\neq r} h_{rj}}\right), \tag{4.9}$$

with Ω_r a leverage-weighted average preference probability of the observed comparisons since $h_{ij} = 0$ for unobserved comparisons, and the constant equal to half of the sum of the leverages for the item.

4.2.7 Comparison

Here and in later analysis it is desirable to consider some of the main distributional features that we might expect to observe in such data. In order to do this we consider three distinct underlying distributions — normal, bimodal and skew normal — that will be used to simulate results. In all simulation distributions the standard deviation of the log-strengths is taken to be 2. Rangel-Smith and Lynch (2018) reports an observed log-strength standard deviation range of between 0.9 to 3.6 logits for CJ assessments. This will have been based on ACJ scheduling as these are data from the Digital Assess platform, so, as we will see later, these may have been somewhat inflated. On the other hand, it is helpful to have a slightly larger range in order to highlight some of the features of the distributions and hence a standard deviation of 2 is chosen here. All have a mean of zero. For each distribution, 100 items are considered. This is in line with the order of magnitude of many assessments based on a university or school cohort. For the three distributions, the log-strength of the kth item is taken to be:

1. Normal: $2\Phi^{-1}\left((k-0.5)/100\right)$
2. Bimodal: $\frac{2}{3.174}\left(\Phi^{-1}\left((k-0.5)/50\right) - 3\right), k = 1, \ldots, 50;$
   $\frac{2}{3.174}\left(\Phi^{-1}\left((k-50.5)/50\right) + 3\right), k = 51, \ldots, 100$
3. Skew normal: $2\Psi^{-1}\left((k-0.5)/100; \alpha = 8, \omega = 3.274, \xi = 2.592\right),$

where Φ is the cumulative distribution function for a standard normal distribution, and Ψ is the cumulative distribution function for a skew normal distribution with $\Psi_X(x;\alpha,\omega,\xi) = \Phi((x-\xi)/\omega) - 2T((x-\xi)/\omega,\alpha)$ where $T(h,a)$ is Owen's T function. These give the densities shown in Figure 4.3

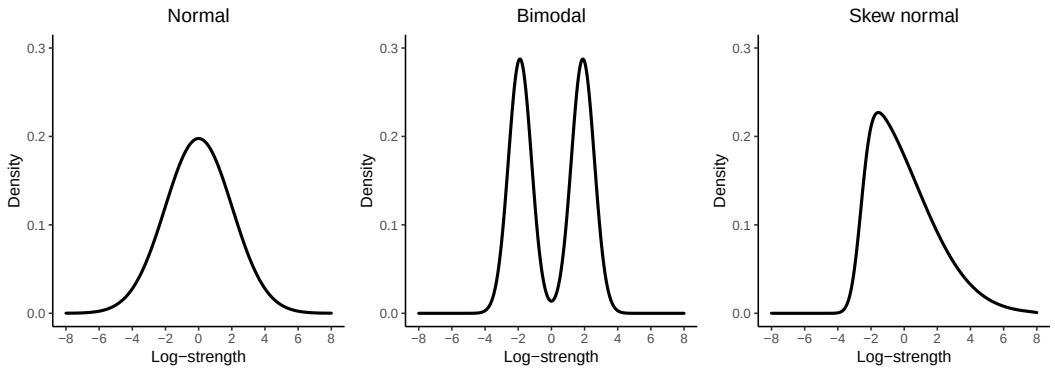

Figure 4.3: Simulated densities

Simulation study

To investigate the performance of these methods, the expected bias and expected absolute error of the log-strength estimates are considered through a simulation study. The approach of Crompvoets et al. (2020) will not be considered due to its undesirable scaling property by which the strength of the penalty changes depending on the number of items. Facets is also not considered, since it can be expected to address the finiteness of estimates but not the bias. For the dummy item method, c_0 will be taken to be 0.25. In testing based on a round-robin tournament, where the expected penalty was calculated analytically, this value was found to provide a good approximation to the method of Firth (1993). For the ϵ-adjustment approach, ϵ will be taken to be 0.3. Robitzsch (2021) suggests that in the wider context of a Rasch model a choice of $\epsilon = 0.24$ is optimal, but here we are interested in what practitioners might do and it seems not unreasonable to consider that many would take the default setting when applying the function. In order to be able to make a direct comparison, α is also taken to be 0.3. For the purpose of this investigation, the assessments will be taken to consist of 20 rounds of comparison with each item appearing in one comparison in each round. The choice of 20 is based on personal correspondence with No More Marking where they confirmed that this was the standard request they made of submitting schools. That is, if a school submits 100 items to be assessed then the school would be expected to perform a thousand pairwise comparisons, so that each item is compared 20 times.

Two scheduling schemes will be investigated. Both will consist of rounds of comparison where each item is compared once in each round. In the first scheduling

scheme, the pairs in each round are selected unifromly at random. In the second scheduling scheme the pairs are selected according to a Swiss system. Under the Swiss scheduling scheme, the first round of pairings is random. In each subsequent assessment round, items are paired with other items with as similar as possible (typically the same) number of preferences up to that point.

While the Swiss system is not formally used in CJ, it forms the basis for the ACJ method, being used in the first four rounds of that scheme and having the same underlying philosophy of matching items of similar strength. For the purposes of this investigation it allows for the approaches to be assessed against an intuitive scheme that is computationally cheaper to simulate, not requiring the re-estimation after each round that ACJ does, and avoids the complication of determining an appropriate method for the intermediate strength estimation. For each scheduling scheme, the three distributions previously specified for the log-strengths — normal, bimodal, skew normal with 100 items, mean zero, and standard deviation of 2 — will be considered. 1000 assessments were simulated under each of the six distribution and scheduling scheme combinations using a Bradley-Terry data generating process. All simulations are performed in **R** (R Core Team, 2021). The Firth (1993) adjustment is fitted using the **brglm2** package (Kosmidis, 2020), the other estimation methods are fitted using a Gauss-Siedel algorithm based on the code used in the **btm** function in **sirt** (Robitzsch and Robitzsch, 2022).

Based on these simulations, estimates of the bias and expected absolute error are calculated as

$$\text{Bias}_i = \frac{1}{N}\sum_{k=1}^{N}(\lambda_{ik} - \lambda_i^*), \tag{4.10}$$

$$\text{Expected Absolute Error}_i = \frac{1}{N}\sum_{k=1}^{N} \mid \lambda_{ik} - \lambda_i^* \mid, \tag{4.11}$$

where λ_{ik} is the estimate for the log-strength of the i$^{\text{th}}$ item ($i = 1, \ldots, 100$) from the k$^{\text{th}}$ simulation ($k = 1, \ldots, N$), with $N = 1000$ in this case, and λ_i^* is the 'true' log-strength of item i used to generate the simulation.

For the randomly generated schedule, Figures 4.4 and 4.5 suggest that all four estimation methods do an effective job in constraining the bias and reducing error. The α-adjustment notably over-constrains at the extremes, especially for the skewed population of log-strengths.

Under the Swiss scheduling scheme there is more divergence between the estimation methods. The α-adjustment gives notably lower absolute error and bias than the other methods, with the expected absolute error comparable to the results from the

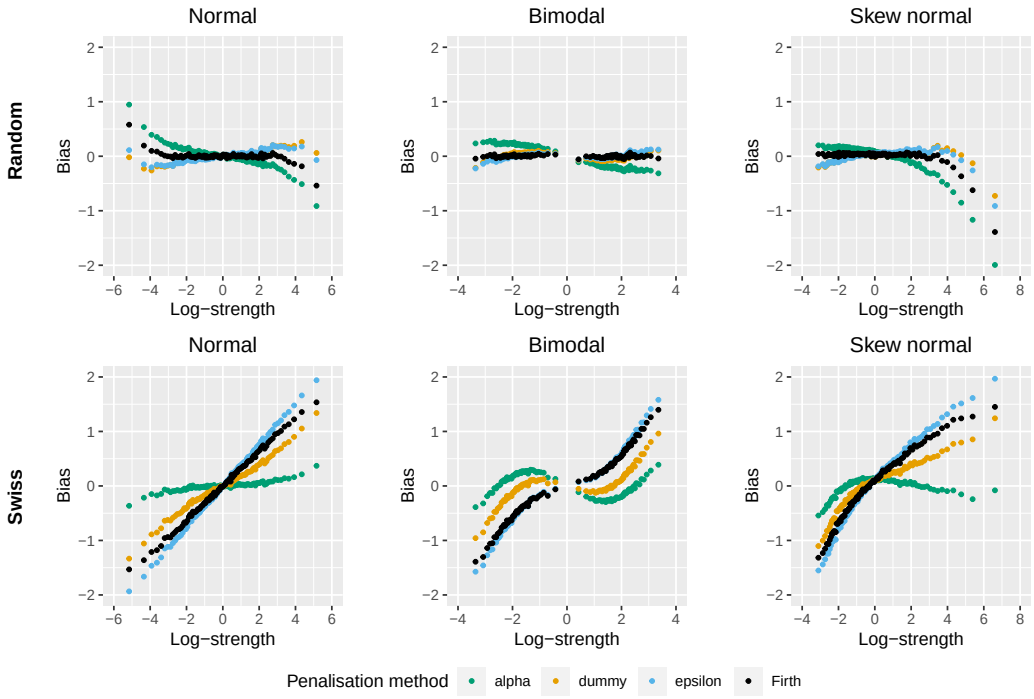

Figure 4.4: Bias of log-strength estimates under different scheduling scheme, log-strength distribution and penalisation method combinations. All four methods achieve substantially reduced bias for random schedules, but only α-adjustment is effective in reducing bias under a Swiss scheme.

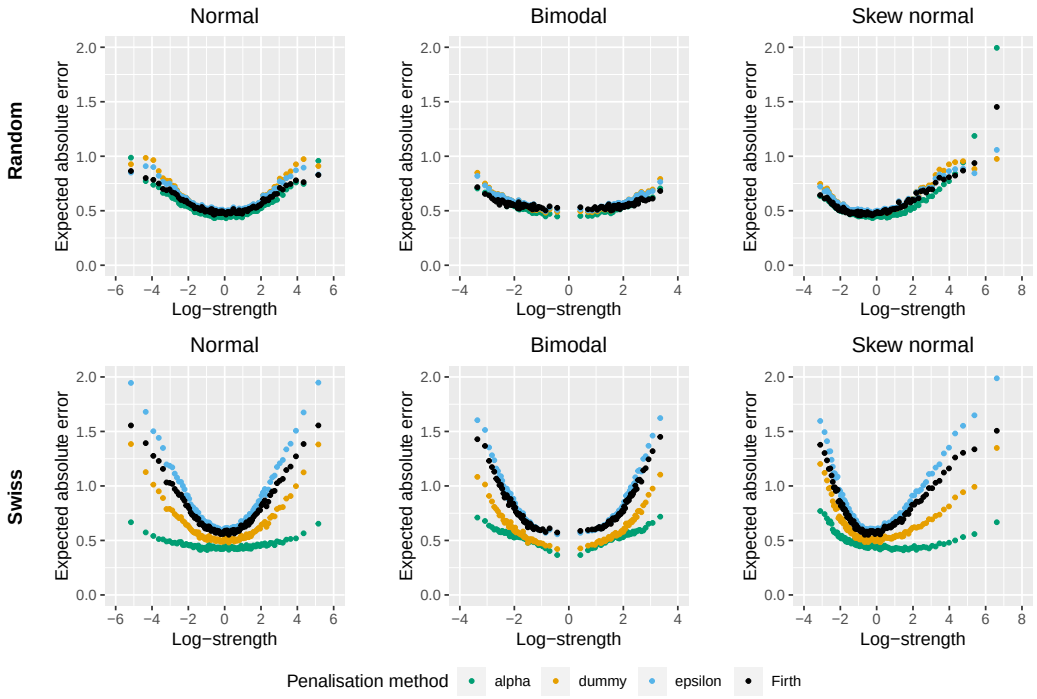

Figure 4.5: Expected absolute error of log-strength estimates under different scheduling scheme, log-strength distribution and penalisation method combinations. All four methods achieve substantially reduced error for random schedules, but only α-adjustment is effective in reducing error under a Swiss scheme.

randomly scheduled scheme. Figure 4.4 shows that there is an extremising tendency in the estimation of the log-strengths for all the methods, where the lower strength item estimates have a negative bias and the higher strength item estimates have a positive bias. The intuition behind this effect is that the selective scheme leads to a higher proportion of strong vs strong or weak vs weak comparisons, extremising the estimates. In the next sections, the implications for SSR and split-halves are considered, as well as a discussion of the conceptual applicability of the underlying metrics to the context.

4.3 Scale Separation Reliability (SSR)

Recall that the SSR can be considered as an estimate of a coefficient of determination, R^2, the proportion of variation of the observed log-strengths that can be accounted for by the variation of the 'true' log-strengths, as represented by equation (4.3). It is then natural to consider the success of this estimate through simulation, comparing the proportion estimated by SSR to the true proportion. Figure 4.6 plots this comparison for the first 50 simulated tournaments under each of the estimation methods in order to get a sense for the success of this estimate.

For the randomly scheduled tournaments, the results are encouraging for the use of SSR as an estimate of R^2. Most simulations show an R^2 of between 0.85 and 0.9 and estimates within 0.05 of the true value. A close look seems to suggest that for the bimodally distributed log-strengths the α-adjustment gives a slightly lower SSR, and a slightly lower R^2 for the skewed distribution.

For the Swiss scheduled tournament, the results are far less encouraging for the use of SSR. For the Firth (1993), ϵ-adjustment and dummy methods the SSR seems to increase as the R^2 decreases and is inflated by between 0.05 and 0.15 with respect to the true value. However, it is also notable that with normally distributed item log-strengths, the SSR is a good estimate of R^2 when using the α-adjustment, and that it performs better than the other methods for the bi-modal and skew normal distributions. It is also worth noting that for normal and skew normal distributions, R^2 is around 0.05 higher than under the randomly scheduled scheme. This suggests that for any given level of reliability it should be possible to achieve those results with meaningfully fewer judgements if an adaptive scheme is utilised. So that the correct response does not seem to be to reject adaptive scheduling but rather to consider the reliability measure.

The SSR seems to carry a large amount of credibility in the CJ literature and research community, so before advocating its rejection, it is worth understanding the cause of its error, and perhaps even seeking to address that, as Bramley and Vitello

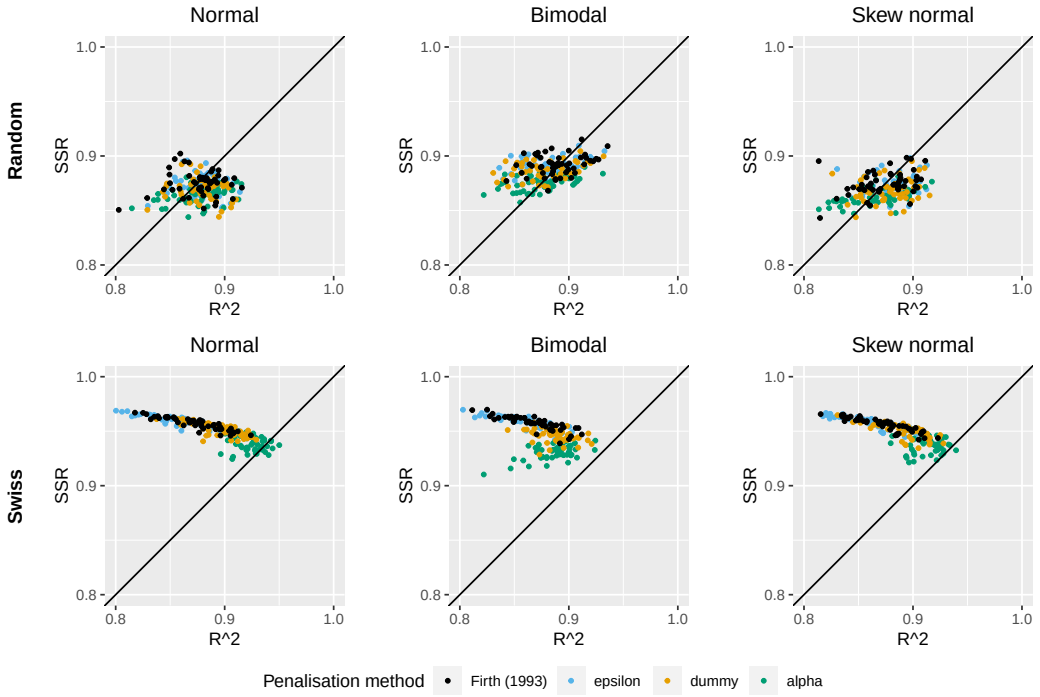

Figure 4.6: SSR accuracy under different scheduling scheme, log-strength distribution and penalisation method combinations. SSR is a reasonable estimate for R^2 with all four methods under a random schedule. α-adjustment gives higher R^2 and a closer match between SSR and R^2 under a Swiss scheme than the other estimation methods, for which SSR is a substantially inflated estimate of R^2.

(2019) suggests. The calculation includes two parts — the observed variance of log-strength estimates and the mean of the squared standard errors as an estimation of the mean of the squares of the actual errors. These are not independent and it is their ratio that is important for the calculation of SSR but it is helpful to consider them separably nevertheless.

Figure 4.7 shows boxplots of the standard deviation of the log-strength estimates. The 'true' standard deviation of the underlying log-strength distribution of 2 is highlighted by the dashed line. This representation reflects on a population level what was observed in Figure 4.4, where all estimation schemes under Swiss scheduling other than the α-adjustment showed an extremising bias. It is notable, however, that this bias is consistent enough that the other estimation methods almost never achieve the underlying standard deviation, while the α-adjustment achieves a narrow band of standard deviation with the mean close to 2. Again, the results for the randomly scheduled scheme are similar across estimation methods and close to the underlying standard deviation of the log-strengths, with the α-adjustment giving a consistently lower standard deviation but by a small absolute amount.

Another source of error in SSR is in its estimation of standard error. Under the SSR, the estimation of the error for each item is taken to be the square root of the inverse of the diagonal of the information matrix. This is incorrect. The standard error is the square-root of the diagonal of the inverse of the information matrix. However, care needs to be taken in the definition of the information matrix here. The information matrix for may be defined as

$$i(\boldsymbol{\lambda})_{ij} = \begin{cases} \sum_k m_{ik} p_{ik}(1 - p_{ik}) & i = j \\ -m_{ij} p_{ij}(1 - p_{ij}) & i \neq j. \end{cases}$$

But this is not of full rank, because the log-strengths are identifiable only up to a constant, and so it is not invertible. The problem of identifiability is generally resolved by applying an identifiability constraint. For example, we might apply the corner point constraint $\lambda_1 = 0$. But the standard errors will then be different depending on the identifiability constraint applied. One alternative is to take a generalised inverse to the information matrix and make the computation on the diagonal of that. In particular, the Moore-Penrose pseudoinverse meets a number of desirable criteria (Penrose, 1955) and so we consider it here.

An alternative to calculating the standard errors, would be to note that the data are pairwise and so we might wish to choose item-level standard errors that can best approximate the standard errors of the pairwise contrasts. This is the intuition behind quasi-variances (Firth and De Menezes, 2004). In the present context, given the variance-covariance matrix for the log-strength contrasts, the method is to identify

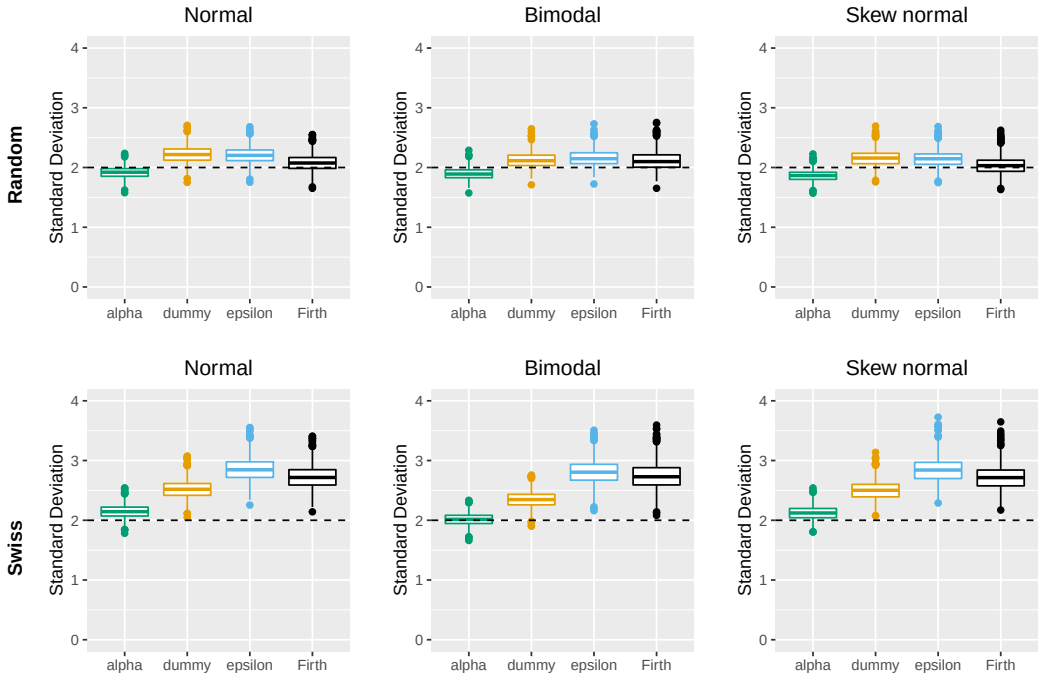

Figure 4.7: Distribution of the standard deviation of log-strength estimates under different scheduling scheme, log-strength distribution and penalisation method combinations. All four estimation methods produce accurate estimates of standard deviation under random schemes. Only α-adjustment achieves a good estimate in the case of Swiss tournaments.

quasi variances q_i such that

$$\text{var}(\lambda_i - \lambda_j) \approx q_i + q_j,$$

by selecting q_i for each item to minimise the penalty function

$$\sum_{i<j} \left(\log \left(\frac{\text{var}\left(\lambda_i - \lambda_j\right)}{q_i + q_j} \right) \right)^2.$$

Figure 4.8 shows the results from taking the first 10 simulations and calculating the three alternative mean squared error methods — the one used in the CJ literature and advocated by Pollitt (2012b), that taken from the diagonals of the Moore-Penrose pseudoinverse to the information matrix, and the quasi-variances — for each estimation method. The small number of simulations charted here is to allow an intuitive graphical representation without those charts being too cluttered, but the findings remain the same when looking at the complete population of simulations.

We observe that across all simulations the estimates from taking the quasi-variances and the diagonal of the Moore-Penrose pseudoinverse of the information matrix are very close. They are also consistently higher than the Pollitt method. For the random scheduling scheme, the difference between the Pollitt method and the other two methods is approximately 0.05 . Given that the variance of the log-strength estimates is close to 4, then this will have minimal impact on the SSR (see equation (4.3)). Both the estimated and true MSE under the α-adjustment method are slightly lower than from the other methods but this seems to be mostly offset by the lower observed standard deviation of the log-strength estimates under α-adjustment observed in Figure 4.7, so that it does not appear to have a material effect on SSR.

For the Swiss schedule, note that there is a change in the range of the axes. The results are much more stark. The estimated MSE is materially lower than the actual MSE, especially under ϵ-adjustment and Firth (1993) estimation methods. The α-adjustment method, taking either the Moore-Penrose diagonal or quasi-variance specifications accurately estimates the MSE. It should also be noted that the Swiss scheduling scheme gives lower actual MSE when using the α-adjustment method than the random scheduling scheme, indicating again the potential advantages from an adaptive scheme.

This analysis supports a number of findings with respect to the SSR. For a randomly generated tournament, if there are a sufficient number of comparison rounds, the SSR is likely to be a good estimate for R^2 under a number of estimation methods, though not all (for example, non-penalised, Facets and Crompvoets et al. (2020)). Adaptive scheduling schemes can achieve higher reliability than random scheduling schemes, but not as high as SSR would suggest. In applying SSR to adaptive

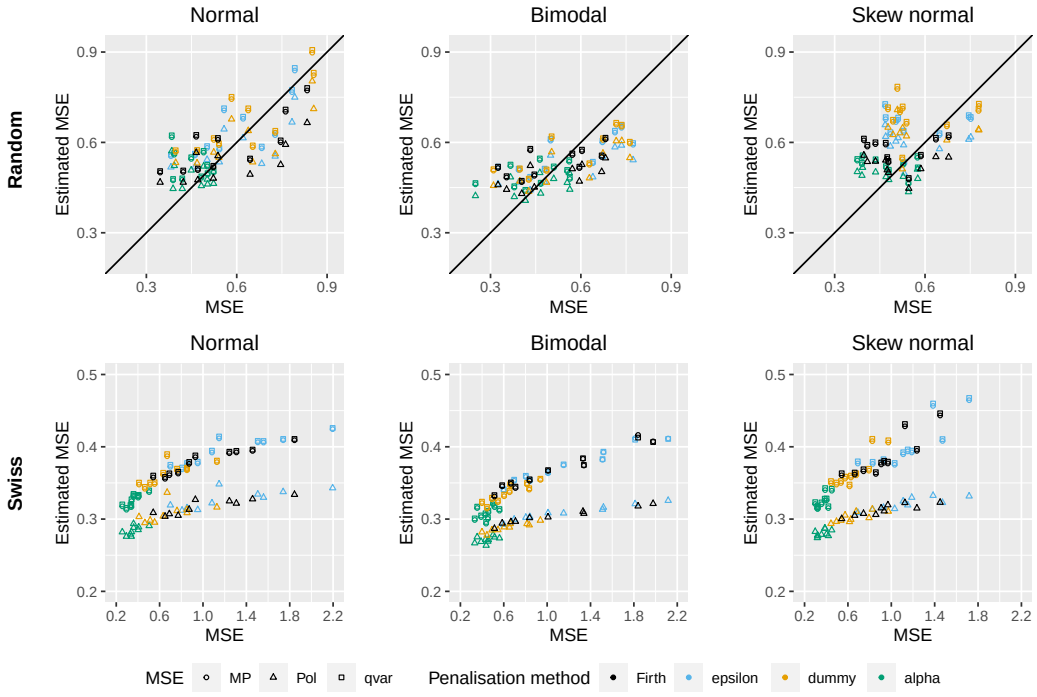

Figure 4.8: Mean squared error (MSE) under different scheduling scheme, log-strength distribution and penalisation method combinations. MSE estimates under Swiss scheduling scheme are highly inaccurate for all estimation methods except α-adjustment.

scheduling schemes, the estimation method matters. The findings here suggest that the α-adjustment method may allow SSR to be an accurate estimation of R^2 for some adaptive scheduling schemes for some underlying log-strength distributions. However, without significantly more empirical or theoretical evidence, this conclusion cannot be asserted more generally. Furthermore, outside of empirical concerns about the use of SSR, there may be conceptual reasons to question its efficacy as a measure of reliability in this context.

Underlying the use of SSR is the assumption that R^2 applied to the log-strengths from the Bradley-Terry model is an appropriate indicator of reliability that we should seek to estimate. There are at least two reasons why we might question this. First, since the measure relies on the average of squares, its conclusions can be dominated by extreme points. In particular in the CJ context, assuming, say, a normal distribution, then the measure is likely to have heavy dependence on the items at the two ends of the distribution where accurate estimation is harder and the errors likely to be greater. In some contexts, a reliability measure more sensitive to the extreme items may be desirable. For example, if we were looking at estimating a league position in a sports tournament where the major consequence of the rating was promotion or relegation, then we would care more about the accuracy of the estimates at the extreme ends, but in the academic context the opposite is generally true. Typically, marks are used to segregate students into grade bands and so whether a student is at an extreme end of the top grade band or just marginally above others in the top grade band is of little consequence. Second, the measure relates directly to the log-strength scale but this is an artefact of the model rather than an intrinsic property. Another way to express this would be that we can apply any monotonically increasing function to the log-strengths to get an alternative set of ratings. For example, the specification of the item strengths within the Bradley-Terry model as $\pi_i = \exp(\lambda_i)$ is a natural one as shown by the motivations of Sections Sections 1.3.3 and 1.3.4. These alternative ratings can be used to derive the same pairwise probabilities, and so would fit the data in exactly the same way, but would give a different R^2. The R^2, even if its estimation is accurate, should therefore be interpreted cautiously as a useful absolute or relative indicator of reliability.

4.4 Split-halves

In this section, we consider the use of the split-halves measure as a measure of reliability. Recall that the split-halves reliability measure partitions the judges in a CJ assessment at the analysis stage into two equally-sized groups. The log-strengths are estimated for each of these groups independently and the Pearson correlation

between them calculated. This procedure is repeated with a number of different partitions (100 times in Bisson et al. (2016), 20 times in Davies et al. (2020)) and a summary of these correlations reported. Given the concerns on using SSR highlighted in the previous section, it might be attractive to interpret split-halves as a broad reliability measure based on the rationale of measurable reliability being essentially the same as inter-rater reliability in this context, as we discussed in Section 4.1.3.

A problem with the interpretation of split-halves as a broad reliability measure, at least in the framework of classical test theory, is that the correlation with the 'true score' has error in only one of the data sets, the 'true score' being fixed. Whereas, the correlation between two estimates includes error in both, and so split halves would likely underestimate the relevant reliability. Even understood more narrowly as inter-rater reliability, it would seem to produce an underestimate because it is a measure produced by comparing two sets of estimates from only half the data. Recall that the definition of reliability was "the extent to which a candidate would get the same test result if the testing procedure was repeated." This suggests repeating the assessment with the same number of judgements. If we were to do so, we might reasonably expect the correlation to be higher, and conceptually it would seem that this is the level of inter-rater reliability that is most relevant to the reported estimates. As a countervailing impact, Pearson correlation may suffer from the same issue as SSR in producing a higher figure when bias is present and an over-dependence on the assessment of extreme items. This may have an interaction with the estimation method.

Here we seek to investigate these elements — the impact of taking the correlation between two populations that are estimated with error, the sensitivity to estimation method, the impact of halving the data, and the degree to which the split-halves method can be applied to adaptive scheduling schemes. In order to do this we use a simulation study.

For the purposes of this investigation, we will consider 100 items, 10 different tournaments, and 10 splits within each tournament, giving 100 split-half assessments in total for each of the six combinations of strength distribution and scheduling scheme. It will be assumed that there are 50 judges such that they each perform one judgement in each round and 20 judgements in total. The judgements are simulated based on a Bradley-Terry data generating process using the underlying log-strengths, with these judgements randomly encoded to judges. In each simulation, three correlations will be calculated: between the log-strength ratings of the two halves ('Half-Half' in Figure 4.9), between the log-strength ratings of the halves and the 'true' log-strengths (Half-True), and between the log-strength estimates from the full data and the 'true' log-strengths (Full-True). Where the partition of comparisons means that an item

does not appear in one of the subsets then the correlation will be calculated only on the items that are common to both sets. The procedure followed for each of the six log-strength distribution by scheduling scheme combinations and for tournaments of 10, 15, 20, 25 and 30 rounds is summarised in Algorithm 1. We take $m = 10$ and $n = 10$.

Algorithm 1 Split-half simulation algorithm

Require: $\boldsymbol{\lambda}^*$,m,n.

1: Simulate Bradley-Terry assessment from log-strengths $\boldsymbol{\lambda}^*$.
2: Split judges randomly into two equal-sized sets, A and B.
3: Estimate log-strengths
   a: based on judgements from set A alone, $\boldsymbol{\lambda}^A$
   b: based on judgements from set B alone, $\boldsymbol{\lambda}^B$
   c: based on all judgements, $\boldsymbol{\lambda}$.
4: Calculate correlation between
   a: $\boldsymbol{\lambda}^A$ and $\boldsymbol{\lambda}^B$ (Half-Half),
   b: $\boldsymbol{\lambda}^A$ and $\boldsymbol{\lambda}^*$ and between $\boldsymbol{\lambda}^B$ and $\boldsymbol{\lambda}^*$ (Half-True),
   c: $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^*$ (Full-True).
5: Repeat steps 2-4 m times.
6: Repeat steps 1-5 n times.

Given the analysis in the previous section, and with the goal of clear presentation, results will be reported only for the α-adjustment and ϵ-adjustment methods. These represented the extremes of fit for the Swiss scheduling scheme and so their comparison may demonstrate the range of sensitivity of the split-halves correlation measure to estimation method.

In practice, we might expect more structure to the judge behaviour than we are imposing in this simulation. For example, one subset of judges in a writing assessment might put more weight on correct use of grammar while another subset might put more weight on use of descriptive words. These tendencies would see judges with similar preferences be positively (negatively) inclined to the same items. However, the size and nature of this structure is very unclear, and so, for the purposes of the questions being investigated here, all judges will be assumed to follow the same Bradley-Terry data-generating model. If such structure did exist, then we would expect it to increase the size of the differences between the Half-Half, Half-True, and Full-True measures, as partitions that align with judge preference differences would show more variation from the overall and in an opposite direction to the other half. It may also be expected to increase the range of the Half-Half and Half-True estimates

for the same reason.

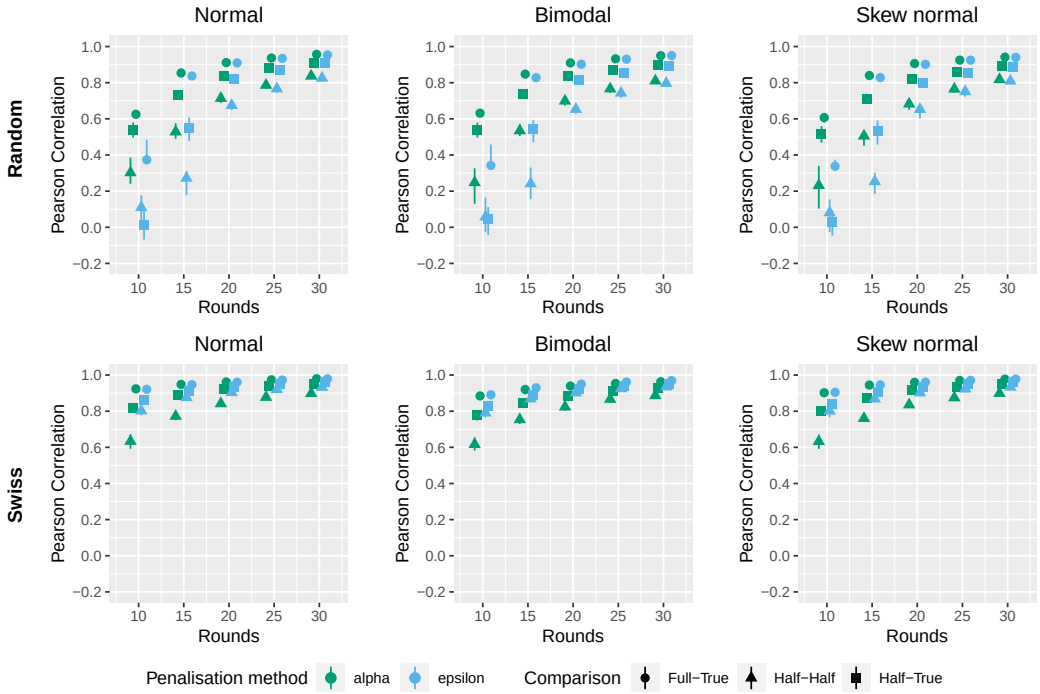

Figure 4.9: Median and inter-quartile range of split-half Pearson correlations for combinations of number of rounds of judgement, scheduling scheme, log-strength distribution, estimation method and the item sets being correlated. Pearson correlation increases substantially with number of rounds. Pearson correlation is higher under Swiss scheduling than random. Pearson correlation between two population halves is substantially lower than between entire population and the 'truth'.

Figure 4.9 presents the results showing the median and inter-quartile range for each correlation. First, the impact of using only half the data can be seen in the difference between the 'Half-True' and 'Full-True' correlations. This difference is around 0.09 for the 20-round tournaments and larger for tournaments of fewer rounds. This represents a material difference. Second, the impact of comparing two populations estimated with error can be seen by comparing the 'Half-Half'; and 'Half-True' correlations. This difference is around 0.11 for the 20-round tournaments and larger for tournaments of fewer rounds. This again represents a material difference. Third, considering the influence of estimation method, it can be seen that for the random scheduling scheme the correlations are similar for 20 or more rounds, but for 10 or 15 rounds the difference between the three different correlations is greater for the

ϵ-adjustment. For the Swiss scheduling this result is reversed with only a small difference between the three correlations under the ϵ-adjustment. The difference between estimation methods under Swiss scheduling is consistent with the evidence presented in Figure 4.4, where the ϵ-adjustment was shown to have an extremising bias that would produce higher correlations. Perhaps the most striking aspect of the results are the high correlations observed under the Swiss scheduling scheme, particularly under the ϵ-adjustment method. The high value of the 'Full-True' measure under the Swiss tournament, even with a lower number of rounds of judgement, suggests that the adaptive scheme is producing some benefit compared to random scheduling, consistent with the results presented in Figure 4.6.

There should be caution in the interpretation of the results under the Swiss scheme however. It is tempting to understand the split-half method as running the assessment exercise twice and comparing the results, but under an adaptive scheme the scheduling benefits from all the comparisons, those in both sides of the partition, so this interpretation is incorrect. As expected, the split-half measure improves considerably with the number of rounds of judgement. For example, using a random scheduling scheme, with 15 rounds of judgement a split-half correlation of below 0.6 is observed, but the split-half from 30 rounds where two 15-round judgements are compared gives a correlation of above 0.8. This suggests that the agreement we might expect were we to repeat the assessment with 15 rounds of judgement is substantially higher than the split-half method suggests. In all cases, the split-half correlation is substantially below the 'Half-True' correlation, so understood in the context of classical test theory and a desire to understand the measurement error, the split-half method again seems conservative.

Finally, we consider the appropriateness of Pearson correlation as a measure in this context. As with SSR, it suffers from being a measure on the somewhat arbitrary scale of log-strengths, so that a monotonic transformation of the estimates in a way that would not change the model could lead to a different correlation. An alternative would be to employ a correlation on the rank orders as Verhavert et al. (2018) does. However, this could lead to lower reliability despite more accurate estimates of relevant strengths if a population of items happened to have a smaller variation in quality. An alternative that addresses these concerns will be proposed in Section 4.7.

4.5 Bootstrap measures

In this section, we discuss the use of bootstrap methods for estimation of reliability measures. Bootstrap methods (Efron, 1992) are a commonly used tool that use

random sampling to estimate the properties of an estimand, with the sampling based on a distribution that approximates the one of interest. In the present context, the idea is to use a parametric bootstrap where the estimated log-strengths are used to simulate CJ assessments and then measures are calculated which relate the log-strengths from those simulated CJ assessments to the original estimated log-strengths. This may then be understood to approximate the measure we would derive if we were able to compare our original estimated log-strengths to the 'true' underlying log-strengths. The approach aligns well with our definition of reliability — "the extent to which a candidate would get the same result if the testing procedure was repeated" — in directly simulating repeated assessments.

A parametric bootstrap of this type might be expected to address some of the concerns that were highlighted in the previous sections, especially in the estimation of reliability under adaptive schemes. The key issue that arose for both the SSR and split-halves measures was that there was an extremising bias to the log-strength estimates. In the case of the SSR, the MSE was also underestimated for the Swiss scheduling scheme. A small part of this was due to the particular definition of standard error that has commonly been applied in CJ. But the much larger effect was due to taking an asymptotic measure of variance with the inappropriate assumption that the schedule is an ancillary statistic. This will be discussed further in Section 4.9.

These issues manifested for all estimation methods other than the α-adjustment. That the α-adjustment seemed to address this in the simulation studies presented in Sections 4.3 and 4.4 is encouraging and an intuition behind this will also be discussed in Section 4.9, but without substantial further testing or some theoretical grounding it is not clear that it would continue to do so for all adaptive schemes, underlying strength distributions, judge consistency levels and variations in numbers of items, rounds and judges. It is therefore desirable to have a method for estimating reliability that we might reasonably expect to address these problems, or at least reliably mitigate them. Bootstrap methods would seem to be a natural candidate for this task, due to their non-reliance on asymptotic theory and close relation to the concept of reliability. On the other hand, one reason that we might be cautious of this approach is that there is still potential for inflation of the R^2 estimate from the extremising bias, since simulating from an extremised distribution could be expected to bring the R^2 estimate closer to unity. We also note that in producing multiple simulations, the bootstrap approach very naturally gives the distribution for any measure which may be helpful in the interpretation of any reliability measure.

In this section, one particular bootstrap method will be investigated, with an alternative discussed. We will apply it to the R^2 measure, for which SSR is an al-

ternative estimate, but it could be applied to produce estimates of any measure of similarity between the 'true' log-strengths and the estimated log-strengths. The approach begins by estimating the log-strength of the items based on the comparisons in the observed assessment in the conventional way. These estimated log-strengths are then used as the underlying strengths to simulate further assessments. Crucially, these simulation assessments apply whatever scheduling scheme was used in the original assessment, and will be fitted using the same estimation method. The estimates from these simulated assessments may then be compared to the original estimates to derive any statistic that is of interest. In the context of estimating R^2, the errors can be estimated by taking the errors between the simulated log-strength estimates and the original log-strength estimates.

The simulation will consist of five assessments, with 100 items, and 20 rounds in each assessment. For each assessment and each estimation method, the log-strengths will be estimated, and these used to simulate a further 50 assessments according to the scheduling scheme that was used to generate the original assessment. Log-strength estimates for each of the 50 simulations are then calculated, applying the same estimation method. The R^2 is calculated between the original log-strength estimates and each of the 50 simulated log-strength estimates. Summary statistics for these 50 R^2 estimates are calculated. The R^2 is then calculated between the original estimate and the underlying 'true' log-strengths. This is understood to be the value that is being estimated through the simulation and against which the success of the procedure can be assessed. This procedure is performed for each of the six distribution and scheduling scheme combinations. A summary of this procedure is provided in Algorithm 2. We take $m = 50$ and $n = 5$.

Algorithm 2 Bootstrap R^2 simulation algorithm

Require: $\boldsymbol{\lambda}^*$,m,n.

1: Simulate assessment A with Bradley-Terry outcomes based on log-strengths $\boldsymbol{\lambda}^*$.
2: Estimate log-strengths, $\boldsymbol{\lambda}$, based on judgements from assessment A.
3: Simulate assessment A_s with Bradley-Terry outcomes based on log-strengths $\boldsymbol{\lambda}$.
4: Estimate log-strengths, $\boldsymbol{\lambda}^s$, based on judgements from simulated assessment A_s.
5: Calculate R^2 between $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}^s$.
6: Repeat steps 3-5 m times.
7: Calculate R^2 between $\boldsymbol{\lambda}^*$ and $\boldsymbol{\lambda}$ to compare to the R^2 calculated in step 5.
8: Repeat steps 1-7 n times.

The results of this exercise are summarised in Figure 4.10. It is informative to compare with those of Figure 4.6, as an alternative approach for the estimation of

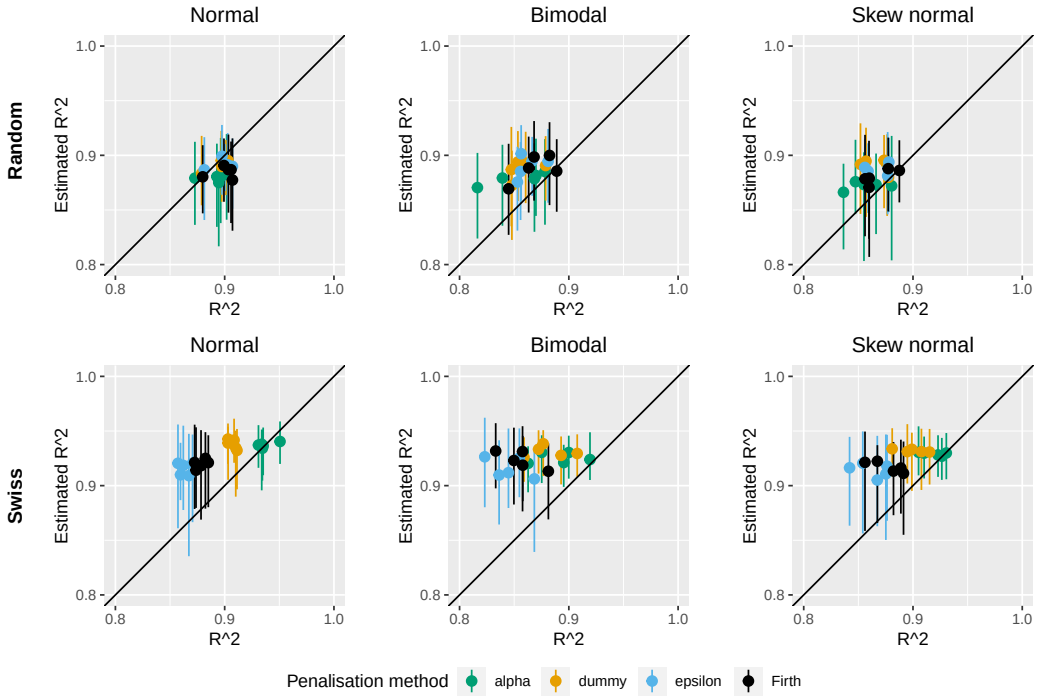

Figure 4.10: Median and 95% range for bootstrap R^2 estimation under different scheduling scheme, log-strength distribution and penalisation method combinations. All four estimation methods perform similarly under a random schedule. α-adjustment performs better than alternatives under Swiss schedule.

R^2. The performance on the randomly scheduled assessments is reasonable across all estimation methods and log-strength distributions. For the Swiss scheduled assessments, the bootstrap R^2 performs better than SSR for the dummy, ϵ-adjustment and Firth (1993) estimation methods with the difference between the true R^2 and the R^2 estimate reduced compared to the difference observed with SSR. However, the difference is still notable. It is uniformly flattering, indicating greater reliability than is the case. The true value is mostly not even within the 95% range of estimates. This seems likely to be due to the extremising bias of these methods when applied to Swiss scheduling schemes. The α-adjustment continues to perform well across all assessment scheme and log-strength distribution combinations, though marginally less so for the bimodal distribution.

This is perhaps the most intuitive bootstrap method to apply, but there are alternatives that could be explored. For example, there may be a desire to try to reflect the notion of an inter-rater reliability more directly without the loss of information inherent to the split-halves approach. The idea would be to determine individual judge-level probabilities for each of the pairwise comparisons. The population of judges used for each simulation would then be sampled with replacement from the original population of judges, and each judgement would be determined by that judge's probabilistic preference rather than the one due to the overall log-strengths. The method thus closely resembles the first approach, but perhaps comes closer to a more literal interpretation of the concept of reliability by envisioning repeating the assessment with a potentially different population of judges, each expressing their own preferences.

In order to simulate assessments in this way, one would require the ability to estimate the probability, p_{ijk}, of an item i being preferred to an item j by a judge k, for all i, j, k. In some environments it may be possible to identify important item traits and the degree to which each item possesses each trait and then use a judge's comparison decisions to estimate the degree to which each trait is valued by each judge and therefore how highly each item is rated by a judge. This approach is similar to the one used to determine overall feature preference in Floridi and Lauderdale (2022). However, in a typical CJ setting there is not sufficient trait information for such an approach, and there is also likely not to be sufficient data for each judge.

A more generally applicable approach would be to rely on the insight that judges are allocated to comparisons randomly so that

$$p_{ij} = \frac{1}{J} \sum_{k=1}^{J} p_{ijk},$$

where J is the number of judges. For any judge a, we can consider the probability

that i is preferred to j if a were not part of the judging population.

$$ {}_{-a}p_{ij} = \frac{1}{J-1} \sum_{k=1, k \neq a}^{J} p_{ijk}, \quad \text{for all } i, j. $$

Therefore

$$ p_{ija} = J p_{ij} - (J-1)\, {}_{-a}p_{ij}. $$

In computing p_{ija}, p_{ij} can be taken to be the estimate for the probability that i is preferred to j using all the results from the original assessment, and ${}_{-a}p_{ij}$ can be taken to be the estimate for the probability that i is preferred to j when all results except those due to judge a are used. Thus, judge-specific pairwise probabilities may be calculated and then used for simulation purposes.

This procedure might be understood to be a form of jackknife estimation (Quenouille, 1949), though here we are leaving out a set of observations — those from a single judge a — rather than a single observation as in the classic jackknife, and the purpose is to infer something about the omitted judge's preferences rather than the population overall. In doing so, there is an assumption that any single judge will have generally similar preferences to the wider population. If better data on the preferences of each individual judge were available, either because each judges performs more judgements or because each judgement has more information, then it may be possible to consider more sophisticated factor analysis approaches that would allow both the identification of judge preferences and latent qualities of the items being assessed. In many CJ assessments, however, it is common for a large number of judges to perform a relatively small number of judgements, so such approaches may be infeasible.

4.6 Bias-corrected estimation

In this section, we consider the use of a bootstrap method for correcting the bias of the log-strength estimates. This represents another common use of bootstrap methods and seems well-suited to this problem. As was seen in Figure 4.4 and discussed in Section 4.2, under a Swiss scheduling scheme, all except the α-adjustment method, led to biased estimates. The bias-correction method here works first by estimating the log-strengths of a Bradley-Terry model based on the results of the CJ assessment of interest. These log-strength estimates are then used to simulate other CJ assessments using the same scheduling scheme and a Bradley-Terry data-generating process. The log-strengths are then estimated for these simulated CJ assessments. For the bias-correction calculation, in resimulating the assessments, we maintain the part of the

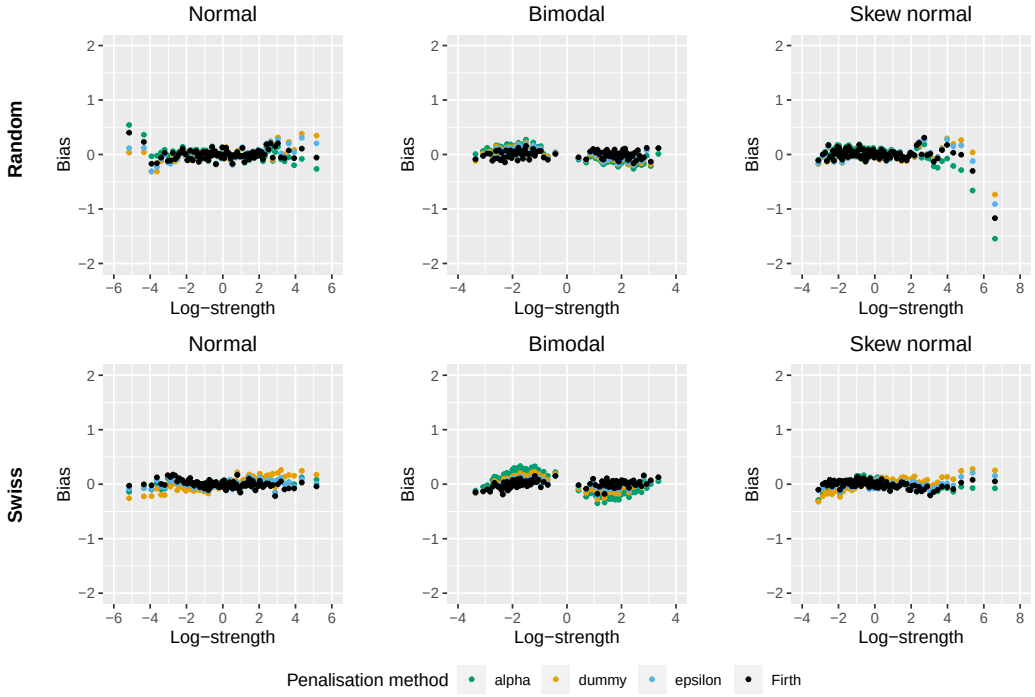

Figure 4.11: Bias using bias-corrected estimator under different scheduling scheme, log-strength distribution and penalisation method combinations. Bias is reduced to close to zero in all cases.

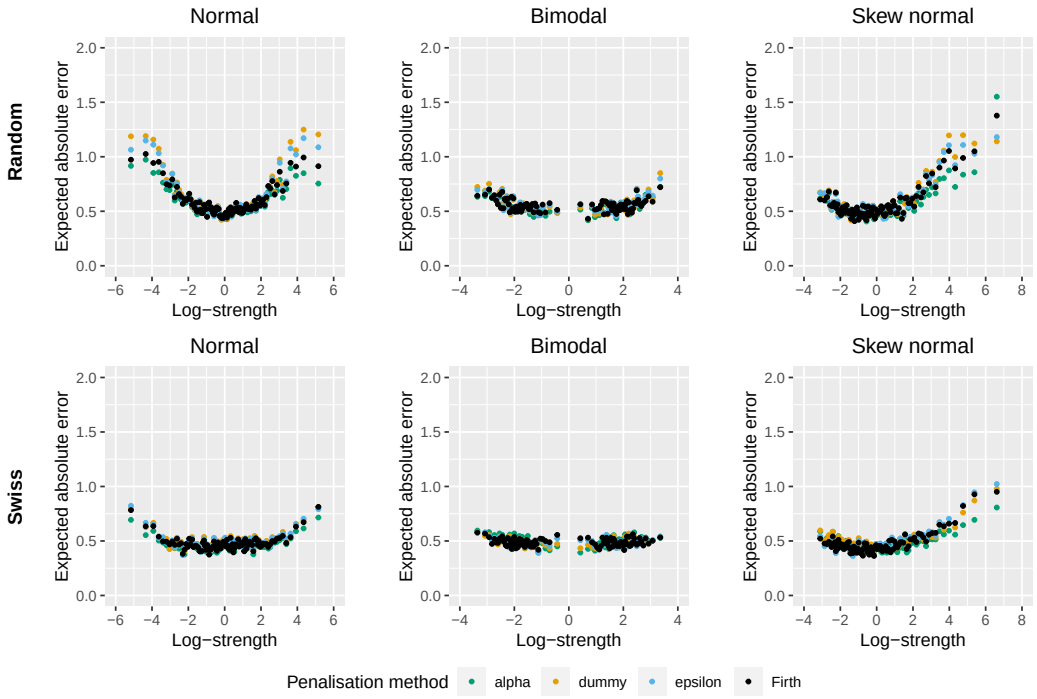

Figure 4.12: Expected absolute error using bias-corrected estimator under different scheduling scheme, log-strength distribution and penalisation method combinations.

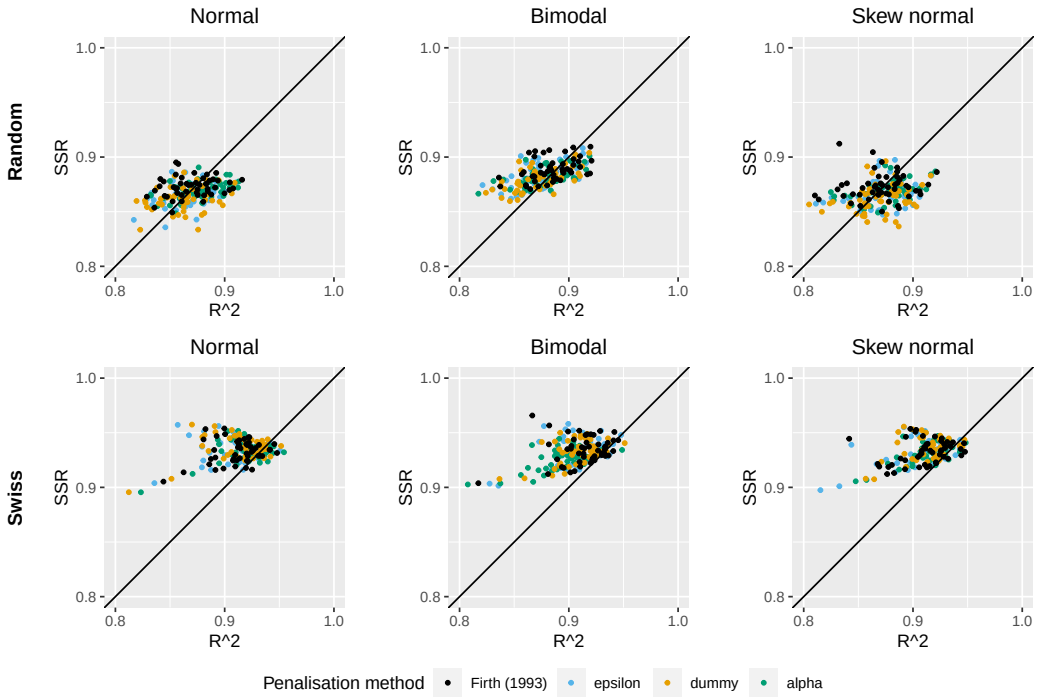

Figure 4.13: SSR using bias-corrected estimator under different scheduling scheme, log-strength distribution and penalisation method combinations. Error is substantially reduced by bias-correction in the case of Swiss tournaments. Using bias-corrected estimator improves the performance of SSR as an estimate of R^2 for Swiss tournaments.

scheduling that is an ancillary statistic. For the randomly scheduled tournament that is the entire tournament. For the Swiss scheme that is just the first round. this is in line with the insight of Cox (1958) that inference should be conditioned on any ancillary statistic.

The bias of an original estimate is then calculated as the mean of the difference in the item log-strength estimates of the simulations and the original estimate. This bias is then subtracted from the original log-strength estimate to get a bias-corrected estimate. That is, the bias-corrected item log-strength estimates are

$$ {}_{BC}\lambda_i = {}_0\lambda_i - \left(\frac{1}{S} \sum_{s=1}^{S} {}_s\lambda_i - {}_0\lambda_i \right), \tag{4.12} $$

where ${}_0\lambda_i$ is the original estimate for the log-strength of item i, and ${}_s\lambda_i, (s = 1, \ldots, S)$ is the log-strength estimates for item i from simulation s. Here we take $S = 40$, and produce results by averaging over 100 tournaments for each of the six log-strength distribution and scheduling scheme combinations. Each tournament has 100 items and 20 rounds. This procedure is summarised in Algorithm 3, where we take $m = 40$ and $n = 100$. In the algorithm we utilise the distinction between a tournament, which we defined as a schedule of comparisons not including the outcomes of those comparisons, and an assessment, which is a schedule of comparisons including their associated outcomes.

Having estimated these bias-corrected log-strengths, $\boldsymbol{\lambda}^{BC}$, we calculate the estimated bias and expected absolute error as defined in equations (4.10) and (4.11). We can also calculate the SSR as described in Section 4.1.3. The results of these are presented in Figures 4.11, 4.12 and 4.13 respectively.

Figure 4.11 shows that the bias-correction was largely successful in minimising bias, though the most extreme item under the skew normal distribution and the random scheduling scheme maintained some bias. It is not altogether clear why that would be the case, but it is unlikely to be impactful to overall results. Figure 4.12, when compared to Figure 4.5, shows that the bias correction has also materially reduced the expected absolute error in all cases and especially for the Swiss tournaments. Consequently, in Figure 4.13, the SSR based on the bias-corrected estimates is significantly improved, and once again shows that the Swiss scheduling scheme achieves higher R^2 values than the randomly scheduled tournament.

Algorithm 3 Bias-correction simulation algorithm

Require: $\boldsymbol{\lambda}^*$,m,n.

1: Simulate tournament T and assessment A using a Bradley-Terry data-generating process and log-strengths $\boldsymbol{\lambda}^*$.
2: Estimate log-strengths, $\boldsymbol{\lambda}$, based on judgements from assessment A.
3: **if** Scheduling = random **then**
4: Simulate assessment A_s from tournament T using a Bradley-Terry data-generating process and log-strengths $\boldsymbol{\lambda}$.
5: **else**
6: **if** Scheduling = Swiss **then**
7: Simulate assessment A_s using only first round of tournament T and later rounds based on Swiss scheduling using a Bradley-Terry data-generating process and log-strengths $\boldsymbol{\lambda}$.
8: **end if**
9: **end if**
10: Estimate log-strengths, $\boldsymbol{\lambda}^s$, based on judgements from assessment A_s.
11: Repeat steps 3-10 m times.
12: Calculate bias-corrected log-strength estimates $\boldsymbol{\lambda}^{BC}$ using equation (4.12).
13: Repeat steps 1-12 n times.

4.7 Alternative measures

So far, we have continued to consider the most common reliability measure, SSR as an estimate of R^2, and have shown how selection of an appropriate underlying penalised likelihood estimation method, use of a bias-corrected estimator, and application of a bootstrap for estimating the measure are all likely to contribute to provide an improved estimate of R^2 and hence allow for the use of adaptive scheduling schemes that are more efficient. However, in Sections 4.3 and 4.4 the appropriateness of R^2 and Pearson correlation as measures of reliability, even if accurately estimated, were questioned. In this section we therefore propose an alternative that has not appeared before in the CJ literature.

A criticism that applied to both SSR and the split-halves measures was that they were dependent on the log-strength scale, which was essentially arbitrary. A monotonically increasing function could be applied to these estimates along with an appropriate adjustment to the model specification, such that neither the probabilities associated with the pairwise comparisons nor the likelihood that the model yields would change. Yet measures based on R^2 or Pearson correlation could change substantially when using these transformed parameters. Conceptually the log-strengths are a non-unique means to parametrise the probabilities of the pairwise outcomes. They are not in themselves especially meaningful.

This insight might encourage us to consider alternative item ratings and related measures based on those pairwise probabilities. An obvious choice, consistent with the arguments made in Chapter 2, is to consider Expected Preferences per Comparison (EPC) with

$$\mathrm{EPC}_i = \frac{1}{n-1} \sum_{j=1;j\neq i}^{n} p_{ij}, \tag{4.13}$$

where the EPC_i is estimated based on the p_{ij} derived from log-strength estimates. This rating has the conceptual advantage of being based on the type of data that we actually observe — the pairwise comparisons, being naturally constrained to the interval $[0, 1]$ as many academic assessments are, and having an immediate intuitive meaning as the mean number of preferences that an item would be expected to receive if it were compared to all others. This idea has been proposed previously in the context of sport in Hamilton and Firth (2021). This can be estimated by applying a Bradley-Terry model to the observed judgements to estimate the strength parameters and thus each of the p_{ij}.

Having adopted this as a rating, we can define an overall reliability score by considering the errors in the item EPCs. One way of doing this is to consider the average absolute error in EPC, defining a measure that will for convenience just be

referred to as κ.

$$\kappa = 1 - \frac{1}{n}\sum_{i=1}^{n} \mid \mathrm{EPC}_i - \mathrm{EPC}_i^* \mid, \tag{4.14}$$

where EPC_i^* is the 'true' EPC of item i and EPC_i its estimate. As discussed with respect to log-strengths in Section 4.1.3, we might reasonably understand EPC_i^* to be the proportion of preferences that item i would receive if it were compared against all other items a sufficiently large number of times.

Total EPC is fixed since

$$\sum_{i=1}^{n} \mathrm{EPC}_i = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1;j\neq i}^{n} p_{ij} = \frac{n}{2},$$

so the EPCs can be considered an allocation of the $n/2$ total. Thus, the measure may also be interpreted as the proportion of correctly allocated EPC,

$$\kappa = 1 - \frac{1}{n}\sum_{i=1}^{n} \mid \mathrm{EPC}_i - \mathrm{EPC}_i^* \mid = 1 - \frac{\sum_{i=1}^{n} \mid \mathrm{EPC}_i - \mathrm{EPC}_i^* \mid}{2\sum_{i=1}^{n} \mathrm{EPC}_i},$$

where the factor of two on the denominator appears because an absolute error in EPC_i will cause an equal total absolute error in other items and hence the absolute error in proportion to the total EPC would be double-counted. The distribution of EPC for each item and of κ may be estimated using the bootstrap method presented in Section 4.5. Algorithm 4 describes how the results of a CJ assessment can be used to estimate EPC and κ. The use of κ will be explored based on empirical data in the next Section.

Algorithm 4 Bootstrap EPC estimation from CJ assessment

1: Conduct CJ assessment A.
2: Estimate log-strengths, $\boldsymbol{\lambda}$, based on judgements from assessment A.
3: Estimate EPC_i for all i based on $\boldsymbol{\lambda}$ and equation (4.13).
4: Simulate assessment A_s according to the same scheduling scheme as used in A with Bradley-Terry outcomes based on log-strengths $\boldsymbol{\lambda}$.
5: Estimate log-strengths, $\boldsymbol{\lambda}^s$, based on judgements from simulated assessment A_s.
6: Estimate EPC_i^s for all i based on $\boldsymbol{\lambda}^s$ and equation (4.13).
7: Estimate κ^s based on the EPC_i^s and EPC_i and equation (4.14).
8: Repeat steps 4-7 S times ($s = 1, \ldots, S$).

4.8 Empirical study

In order to complement the simulation studies that have formed the basis for this chapter so far, this section seeks to investigate reliability measures using an empirical data set. For this purpose, the results from Bramley and Vitello (2019) are reanalysed. Bramley and Vitello (2019) was a follow-up study to Bramley (2015), which used a simulation-based study to demonstrate that adaptive scheduling schemes could have an inflationary effect on SSR. That work was criticised for being based on simulated rather than empirical data (Pollitt, 2015). In response, Bramley and Vitello (2019) sought to collect relevant empirical data.

The data consist of pairwise comparisons from three CJ assessment exercises. The assessment exercises are referred to here, consistently with Bramley and Vitello (2019), as studies 1a, 1b, and 2. They were designed as follows:

> 1a — a study of 150 GCSE English essays, with comparisons scheduled by an adaptive scheme, made by 18 judges.
>
> 1b — a study of a subset of 20 GCSE English essays from the wider set of 150, with a round-robin format where every one of the 20 items was compared with every other item once. The judges were the same as those in 1a.
>
> 2 — a study of the same 150 GCSE English essays, with comparisons scheduled randomly. Essays were judged a similar number of times. There were 16 judges, none of whom had participated in 1a and 1b.

The judges were all examiners of English GCSE for the Oxford, Cambridge and RSA (OCR) examining body. In study 1b, 19 judges were recruited to do 10 judgements each. However, one judge dropped out late in the process and the results of another were excluded based on their poor consistency with other judges and their response time (average of 1 second per judgement). This means that each item is compared to just 17 rather than 19 others in study 1b. A summary of the data is given in Table 4.1. For further details of data collection, see Bramley and Vitello (2019).

The adaptive scheme used in study 1a was based on the progressive method proposed in Revuelta and Ponsoda (1998) and further discussed in Barrada et al. (2008, 2010). This method has intuitive appeal as it provides for a gradual transition from scheduling pairs on a random basis to scheduling pairs with the most information. Unfortunately for our purposes, the implementation details of the method are not given, and have not been discernible from the available evidence of the relevant literature and data. Since being able to simulate consistently with the scheduling scheme

	1a adaptive	1b round-robin	2 random	Combined (1a, 1b, 2)
Essays	150	20	150	150
Judges	18	17	16	34
Judgements per essay	14.4	17	14.6	31.3

Table 4.1: Data summary for studies from Bramley and Vitello (2019)

is a requirement of both the bootstrap measure and the bias-corrected estimation, these methods are not empirically testable here for the data collected in study 1a.

However, the data set is still useful in at least two ways. First, the assessment of the same items through an adaptive and a random scheme enables us to examine the interaction between estimation method and scheduling scheme in determining reliability. Second, the all-play-all structure of study 1b gives a large amount of data on these 20 items, so that we can use the log-strength estimates for these as 'quasi-true' values. In turn these allow for the calculation of a 'quasi-true' R^2. This allows for a comparison between SSR, a bootstrap R^2 and the 'quasi-true' R^2, giving us an indication as to the usefulness of SSR and bootstrap R^2 as measures of reliability.

Bramley and Vitello (2019) states that maximum likelihood estimation was used to calculate the log-strength estimates under a Bradley-Terry model. It does not discuss the use of a penalty, and indeed the results presented in Tables 4.2 and 4.3, showing respectively the SSR and standard deviation of the log-strength estimates, strongly suggest that none was used. For study 2, there are eight essays which were either preferred or dispreferred in all comparisons. It is reported that these "receieved a measure based on an extrapolation rule", though this is not specified. This accounts for the calculation of SSR in study 2 despite there being no finite estimate for their log-strength when no penalty is used.

The results are suggestive of bias in the estimates reported by Bramley and Vitello (2019), due to not using an adequate penalty, especially in the case of the adaptive scheme. This led to the unintuitive finding that the preferred estimate of reliability, SSR, was higher based on the data from study 1a alone than when analysing the data combined over all studies. It is theoretically possible that reliability could be lower for estimates based on the larger combined data set, for example, if the judges in study 1b had starkly different judging criteria to those in study 1a. But, given their common background as approved examiners, this seems unlikely. This finding is therefore consistent with the earlier observation that insufficiently penalised estimation of log-strengths will inflate SSR, especially in the case of an adaptive scheme.

Looking at Table 4.2, we also see that the SSR from the adaptive and combined

data is equal when using an α-adjustment penalty with α set to 0.3. One possibility is that $\alpha = 0.3$ does not provide a sufficient penalty for the analysis of data from study 1a. One way to investigate this is by looking at the standard deviations of the estimates in Table 4.3. The combined data has a large number of judgements per essay and so we might expect the standard deviation of the log-strength estimates to be consistent across different estimation methods and indicative of a true range. We do indeed see much more consistency to the standard deviation of log-strength estimates using the combined data of around 1.4 logits. This in turn suggests that $\alpha = 0.3$ fails to provide sufficient shrinkage for the data from study 1a, whereas taking the stronger penalty $\alpha = 0.5$ appears to constrain results closer to the likely 'true' standard deviation. While Sections 4.2 and 4.3 were supportive of using the α-adjustment, and Section 4.9 will go on to provide some intuition behind that, they provide no reason to believe that 0.3 would be an optimal choice for α for all adaptive schemes covering all numbers of comparisons per item and so it is not surprising that a different value may be better for analysing the data from study 1a.

	1a adaptive	2 random	Combined (1a, 1b, 2)
Bramley and Vitello (2019)	0.98	0.72	0.91
No penalty	0.98	N/A	0.91
Firth (1993)	0.97	0.72	0.89
$\alpha = 0.3$	0.89	0.69	0.89
$\alpha = 0.5$	0.82	0.64	0.87

Table 4.2: SSR as reported in Bramley and Vitello (2019) and using no penalty, Firth (1993) and α-adjustment.

	1a adaptive	2 random	Combined (1a, 1b, 2)
Bramley and Vitello (2019)	4.7	1.8	1.5
No penalty	4.7	N/A	1.5
Firth (1993)	4.0	1.4	1.4
$\alpha = 0.3$	1.9	1.3	1.3
$\alpha = 0.5$	1.4	1.1	1.2

Table 4.3: Estimated log-strength standard deviation as reported in Bramley and Vitello (2019) and using no penalty, Firth (1993) and α-adjustment.

The extra comparisons for the subset of 20 items considered in study 2 provide

a means of evaluating SSR directly. When the data from all studies are combined, the 20 items have been judged a mean of 45 times, including a direct comparison between each other in almost all cases. It might therefore be reasonable to take the log-strength estimates for these items from the combined data as a 'quasi-true' log-strength, against which an R^2 may be calculated and compared to the SSR. Here, this is done by estimating the log-strengths using the α-adjustment with $\alpha = 0.3$, though the results are not sensitive to this choice as the high R^2 across methods using the combined data in Table 4.5 show.

	1a adaptive	2 random	Combined (1a, 1b, 2)
Bramley and Vitello (2019)	0.97	0.70	0.93
No penalty	0.97	N/A	0.93
Firth (1993)	0.96	0.68	0.92
$\alpha = 0.3$	0.87	0.66	0.92
$\alpha = 0.5$	0.80	0.74	0.91

Table 4.4: SSR for the subset of 20 items from study 1b as reported in Bramley and Vitello (2019) and using no penalty, Firth (1993) and α-adjustment estimation methods.

	1a adaptive	2 random	Combined (1a, 1b, 2)
No penalty	0.48	N/A	0.99
Firth (1993)	0.53	0.74	1.00
$\alpha = 0.3$	0.86	0.77	1
$\alpha = 0.5$	0.91	0.70	1.00

Table 4.5: 'Quasi-true' R^2 for the subset of 20 items from study 1b using no penalty, Firth (1993) and α-adjustment estimation methods.

Tables 4.4 and 4.5 show SSR and 'quasi-true' R^2 respectively for the subset of 20 items. The 'quasi-true' R^2 is calculated against the log-strengths derived from an $\alpha = 0.3$ estimate using data from studies 1a, 1b and 2 combined. When using the data only from 1a, the estimations using no penalty gives an SSR of 0.97. This compares to a 'quasi-true' R^2 value of just 0.48, demonstrating the inflationary effect of the combination of an adaptive scheme and a weak (or no) penalty. The R^2 is highest at 0.91 under an α-adjustment penalty with $\alpha = 0.5$, though the SSR is substantially lower at 0.80. The 'quasi-true' R^2 when using $\alpha = 0.3$ and Firth (1993)

penalties are substantially higher than the SSR, suggesting that SSR provides an underestimate for R^2 in this case.

The lack of details on the adaptive scheduling method prevents the application of bootstrap methods to the data from 1a. However, they may be applied to the data from 2. Tables 4.6 and 4.7 show the results from doing so. Because it is not possible to get finite log-strength estimates without penalisation from the results of study 2, a weak penalty, taking $\alpha = 0.005$, is used instead. This ensures estimates are finite while limiting the degree to which the estimates are debiased.

	SSR	Bootstrap R^2	'Quasi-true' R^2
$\alpha = 0.005$	-0.18	0.44 (0.01, 0.72)	0.70
Firth (1993)	0.68	0.76 (0.48, 0.90)	0.74
$\alpha = 0.3$	0.66	0.77 (0.52, 0.89)	0.77
Bias-corrected(Firth (1993))		0.79 (0.44, 0.91)	0.74
Bias-corrected($\alpha = 0.3$)		0.79 (0.57, 0.91)	0.78

Table 4.6: SSR and mean and 95% range bootstrap R^2 compared to 'true' R^2 for the 20 items in study 1b based on the data from study 2.

	Bootstrap κ	'Quasi-true' κ
$\alpha = 0.005$	0.961 (0.944, 0.973)	0.952
Firth (1993)	0.958 (0.942, 0.972)	0.951
$\alpha = 0.3$	0.960 (0.945, 0.972)	0.956
Bias-corrected(Firth (1993))	0.959 (0.945, 0.971)	0.949
Bias-corrected($\alpha = 0.3$)	0.958 (0.943, 0.972)	0.955

Table 4.7: Mean and 95% range bootstrap κ compared to 'quasi-true' κ for the 20 items in study 1b based on the data from study 2.

Looking at Table 4.6, we can see that with the limited penalty the 'quasi-true' R^2 is perhaps surprisingly comparable to that from the other estimation methods, but that the SSR does a poor job of estimating this, while the bootstrap gives a very wide range with the 95% estimation interval spanning 0.71. For the other methods, the bootstrap mean provides a good estimate for R^2, though the ranges produced are wide, suggesting that a randomly scheduled assessment of these items can produce very variable reliability, something not discernible from the SSR point estimate.

Table 4.7 immediately highlights one feature of the κ measure, that it produces very high absolute values with this data set, regardless of estimation method. This

may be an issue if it constrains the ability of the measure to readily differentiate between assessments of varying reliability. It also lacks the intuition that perhaps, for some, comes with the R^2 measure. On the other hand, Figure 4.14 shows that the main outlier in the $\alpha = 0.005$ estimates was a single item. This item was dispreferred in all its comparisons in study 2. That it should be ranked substantially lower than the other items seems reasonable, but precisely how much is immaterial. It would be very unlikely to be preferred in any comparison with the other items. It therefore seems desirable that difficulties in estimating the log-strength of this item within this population should not have a material bearing on the assessment of reliability. This is reflected in the EPC item-strength estimates, which are consistent across estimation method. Consequently the κ measure is also substantially similar across estimation methods in contrast to the results seen for SSR in Table 4.6.

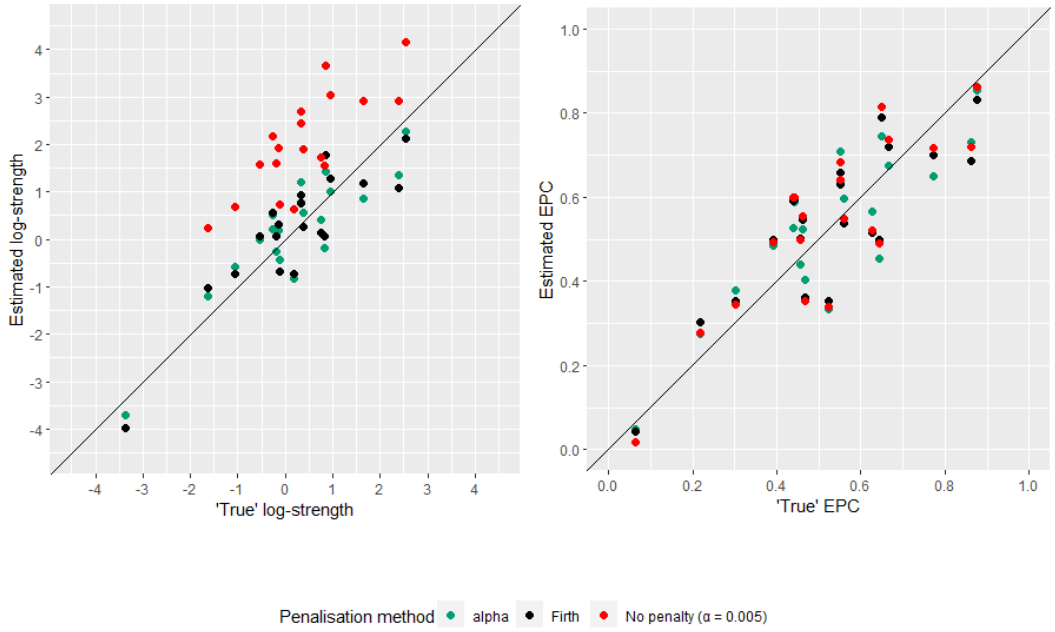

Figure 4.14: Log-strength and EPC estimates for the 20 items in study 1b based on comparisons from study 2 of Bramley and Vitello (2019) plotted against 'true' estimations using the combined data sets from studies 1a, 1b, 2. The lowest rated item had a log-strength estimate of -43.2 under the $\alpha = 0.005$ estimation method and is not plotted here.

4.9 Discussion

The aim of this section is to discuss in more detail the inference being undertaken when a CJ assessment is analysed and to provide an intuition for some of the results seen in the previous sections. We will show that the maximum likelihood estimate is a consistent estimator under any scheduling scheme — random or adaptive — where the conditional independence assumption of the Bradley-Terry model holds. We go on to discuss how the maximum likelihood estimate will be the same given the same data, whether that is collected under a random or adaptive scheme, but that the estimator properties, in particular bias, will be different. Next, by considering a pseudolikelihood that conditions on the comparisons observed, we make a proposal for why the α-adjustment proved to be a successful penalty in the previous analysis. Finally, we discuss the calculation of the information matrix under an adaptive scheme.

In order to do this, it is helpful to introduce some additional notation. Let

- ${}_rX_{ij}, (r \in \{1, \ldots, R\})$, be a random variable that takes value 1 if i is preferred to j in round r and 0 otherwise.
- ${}_rx_{ij}$ be the observed preference from the rth round of a CJ assessment for a pair of items (i, j), where ${}_rx_{ij}$ is 1 if i has been preferred to j in round r and 0 otherwise.
- $\boldsymbol{x} = ({}_1x_{12}, {}_1x_{13}, \ldots, {}_1x_{(n-1)n}, {}_2x_{12}, \ldots, {}_2x_{(n-1)n}, \ldots, {}_Rx_{12}, \ldots, {}_Rx_{(n-1)n})$ be the observed sample of preferences.
- ${}_rY_{ij}$ be a random variable that takes value 1 if i is compared to j in round r and 0 otherwise, so that ${}_rY_{ij} = {}_rY_{ji} = {}_rX_{ij} + {}_rX_{ji}$.
- ${}_ry_{ij}$ be the observed comparison from the rth round of a CJ assessment for a pair of items (i, j), where ${}_ry_{ij}$ is 1 if i has been compared to j in round r and 0 otherwise, so that ${}_ry_{ij} = {}_ry_{ji} = {}_rx_{ij} + {}_rx_{ji}$.
- $\boldsymbol{y} = ({}_1y_{12}, {}_1y_{13}, \ldots, {}_1y_{(n-1)n}, {}_2y_{12}, \ldots, {}_2y_{(n-1)n}, \ldots, {}_Ry_{12}, \ldots, {}_Ry_{(n-1)n})$ be the observed sample of comparisons.
- $C_r = \cap_{i,j}\{{}_rX_{ij} = {}_rx_{ij}\}$ be the event that the preferences in round r were as observed.
- $C = \cap_r C_r = \cap_r \cap_{i,j} \{{}_rX_{ij} = {}_rx_{ij}\}$ be the event that the preferences during each round of the CJ assessment were as observed.

- $c_{ij} = \sum_{r=1}^{R} {}_r x_{ij}$ be the total number of observed preferences for i over j during the whole CJ assessment.
- $M_r = \cap_{i,j}\{{}_r Y_{ij} = {}_r y_{ij}\}$ be the event that the comparisons in round r were as observed.
- $M = \cap_r M_r = \cap_r \cap_{i,j} \{{}_r Y_{ij} = {}_r y_{ij}\}$ be the event that the comparisons during each round of the CJ assessment were as observed.
- $m_{ij} = \sum_{r=1}^{R} {}_r y_{ij}$ be the total number of observed comparisons between i and j over the whole CJ assessment.
- p_{ij} be the probability that i is preferred to j in a comparison.
- $\boldsymbol{\lambda}$ be the vector of log-strengths of the items, where we assume a Bradley-Terry data generating process, such that $p_{ij} = \mathbb{P}({}_r X_{ij} = 1 |_r Y_{ij} = 1) = e^{\lambda_i}/(e^{\lambda_i} + e^{\lambda_j})$.

4.9.1 Estimator properties

Under a maximum likelihood approach, in the present setting, we observe a CJ assessment and wish to estimate the log-strengths, $\boldsymbol{\lambda}$. What we observed consists of both the comparisons made and the corresponding preferences, so we are considering the likelihood,

$$L(\boldsymbol{\lambda}) = L(\boldsymbol{\lambda}; \boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\lambda}) = \mathbb{P}(C, M; \boldsymbol{\lambda})$$

Considering a tournament of R rounds,

$$\begin{aligned}&\mathbb{P}(C, M; \boldsymbol{\lambda}) = \mathbb{P}(C_R, \ldots, C_1, M_R, \ldots, M_1; \boldsymbol{\lambda}) \\ &= \mathbb{P}(C_R, M_R \mid C_{R-1}, \ldots, C_1, M_{R-1}, \ldots, M_1; \boldsymbol{\lambda}) \quad \mathbb{P}(C_{R-1}, \ldots, C_1, M_{R-1}, \ldots, M_1; \boldsymbol{\lambda})\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}(C_R, M_R \mid C_{R-1}, \ldots, C_1, M_{R-1}, \ldots, M_1; \boldsymbol{\lambda}) \\ &= \mathbb{P}(C_R \mid C_{R-1}, \ldots, C_1, M_R, \ldots, M_1; \boldsymbol{\lambda}) \quad \mathbb{P}(M_R \mid C_{R-1}, \ldots, C_1, M_{R-1}, \ldots, M_1; \boldsymbol{\lambda})\end{aligned}$$

Conditional on the event M_R and given $\boldsymbol{\lambda}$, C_R is independent of $C_{R-1}, \ldots, C_1$ and $M_{R-1}, \ldots, M_1$ by the assumption of our data-generating process. Likewsie, conditional on the events $\{C_{R-1}, \ldots, C_1\}$, M_R is independent of $M_{R-1}, \ldots, M_1$ and has no

dependence on $\boldsymbol{\lambda}$ because, even under an adaptive scheme, pairings are based on the preferences observed, so that

$$\mathbb{P}(C_R, M_R \mid C_{R-1}, \ldots, C_1, M_{R-1}, \ldots, M_1; \boldsymbol{\lambda}) \\ = \mathbb{P}(C_R \mid M_R; \boldsymbol{\lambda}) \mathbb{P}(M_R \mid C_{R-1}, \ldots, C_1),$$

and

$$\mathbb{P}(C, M; \boldsymbol{\lambda}) = \mathbb{P}(C_R, \ldots, C_1, M_R, \ldots, M_1; \boldsymbol{\lambda}) \\ = \mathbb{P}(C_R \mid M_R; \boldsymbol{\lambda}) \quad \mathbb{P}(M_R \mid C_{R-1}, \ldots, C_1) \quad \mathbb{P}(C_{R-1}, \ldots, C_1, M_{R-1}, \ldots, M_1; \boldsymbol{\lambda}).$$

Applying the same reasoning iteratively we have that

$$\mathbb{P}(C, M; \boldsymbol{\lambda}) = \mathbb{P}(C_R, \ldots, C_1, M_R, \ldots, M_1; \boldsymbol{\lambda}) \\ = \prod_{r=1}^{R} \mathbb{P}(C_r \mid M_r; \boldsymbol{\lambda}) \prod_{r=2}^{n} \mathbb{P}(M_r \mid C_{r-1}, \ldots, C_1) \quad \mathbb{P}(M_1)$$

Since the first round of comparisons is scheduled randomly under any scheduling scheme, then $\mathbb{P}(M_1)$ is constant. Given $\boldsymbol{\lambda}$, the comparison preferences within any round r, ${}_rC_{ij}$, are independent of one another, so that,

$$\prod_{r=1}^{R} \mathbb{P}(C_r \mid M_r; \boldsymbol{\lambda}) = \prod_{r=1}^{R} \prod_{i,j} \mathbb{P}({}_rC_{ij} \mid {}_rM_{ij}; \boldsymbol{\lambda}) = \prod_{i,j} \prod_{r=1}^{R} p_{ij}^{{}_rx_{ij}} = \prod_{i,j} p_{ij}^{c_{ij}}$$

And so we have that

$$\mathbb{P}(C, M; \boldsymbol{\lambda}) = \prod_{i,j} p_{ij}^{c_{ij}} \prod_{r=2}^{n} \mathbb{P}(M_r \mid C_{r-1}, \ldots, C_1) \mathbb{P}(M_1). \tag{4.15}$$

Note that the term $\prod_{r=2}^{n} \mathbb{P}(M_r \mid C_{r-1}, \ldots, C_1) \mathbb{P}(M_1)$ is not dependent on $\boldsymbol{\lambda}$, and so the estimation of $\boldsymbol{\lambda}$ is entirely dependent on $\prod_{i,j} p_{ij}^{c_{ij}}$, whose form is dictated by the conditional independence assumption of the Bradley-Terry model, that $p_{ij} = \mathbb{P}(i \text{ is preferred to } j; \boldsymbol{\lambda}) = \pi_i/(\pi_i + \pi_j)$. This guarantees that the maximum likelihood estimator will be a consistent estimator under any scheduling scheme where the conditional independence assumption of the Bradley-Terry model holds.

However, it remains the case that adaptive schemes introduce additional bias into parameter estimation. Intuitively, this is because the comparisons under an adaptive scheme are more likely to be between two items closer in strength. For

example, if two strong items, close in strength, are compared, one must be preferred and its strength will then be estimated to be very high since it was preferred to another strong item. With enough comparisons, the relative strengths of these two items would be reflected in their proportion of wins, but under finite (and often sparse) sampling, randomness of preferences will induce bias, with this effect being inflated by an adaptive scheduling scheme. It may therefore be appealing to consider alternative ways of estimating the parameters.

4.9.2 Conditional likelihood

One might consider conditioning on the comparisons observed,

$$\mathbb{P}(C, M; \boldsymbol{\lambda}) = \mathbb{P}(C \mid M; \boldsymbol{\lambda})\mathbb{P}(M; \boldsymbol{\lambda}).$$

Under a wide variety of scheduling schemes (including random, Swiss, ACJ and the progressive scheme of Revuelta and Ponsoda (1998)), conditional on the comparisons observed, the preferences in different rounds and between different pairs are independent given $\boldsymbol{\lambda}$.[4] We may also condition the comparisons on earlier round comparisons so that

$$\begin{aligned}\mathbb{P}(C, M; \boldsymbol{\lambda}) &= \mathbb{P}(C \mid M; \boldsymbol{\lambda})\mathbb{P}(M; \boldsymbol{\lambda}) \\ &= \prod_{r=1}^{R} \mathbb{P}(C_r \mid M; \boldsymbol{\lambda}) \prod_{r=2}^{R} \mathbb{P}(M_r \mid M_{r-1}, \ldots M_1; \boldsymbol{\lambda})\mathbb{P}(M_1). \quad (4.16)\end{aligned}$$

Under a random scheduling scheme, given $\boldsymbol{\lambda}$, C_r is dependent only on M_r. The observed comparisons are independent of other rounds and of $\boldsymbol{\lambda}$, so that $\mathbb{P}(M_r \mid M_{r-1}, \ldots, M_1; \boldsymbol{\lambda}) = \mathbb{P}(M_r)$ is uniform — each possible tournament has equal probability of being observed. Alternatively expressed, $\{m_{ij} : i, j \in \{1, \ldots, n\}, i \neq j\}$ is an ancillary statistic, and we should condition on it (Cox, 1958). Under an adaptive scheme, the observed comparisons are dependent on the item-strengths. For example, under a scheme that sought to maximise information on a pairwise basis, items that are close in strength are more likely to be compared. Relatedly, it is no longer the case that C_r is independent of other round comparisons conditional on M_r.

To illustrate this, consider a two-round CJ assessment involving four items, A, B, C and D. Suppose that the same preferences were observed under a random

[4] An example of a scheduling scheme where this would not be the case would be one where items are paired in each round uniformly at random, but two items may only be compared a third time if they have each been preferred once in the two previous comparisons.

scheduling scheme and a Swiss scheduling scheme. In the first round A was preferred to B and C was preferred to D. In the second round, A was preferred to C and B was preferred to D. Under the Swiss scheduling scheme, the two first-round winners will face each other, as will the two first-round losers. Therefore, if A is compared to C in the second round, then the sample space for preferences from the first round excludes the preferences where A is preferred to B and D is preferred to C or where B is preferred to A and C is preferred to D. In this way, later round comparisons have information on the preferences of earlier rounds. The constituent parts to the two likelihood equations (4.15) and (4.16) are summarised in Table 4.8.

	Random	Swiss
$\mathbb{P}(C_1 \mid M_1; \boldsymbol{\lambda})$	$p_{AB}p_{CD}$	$p_{AB}p_{CD}$
$\mathbb{P}(C_2 \mid M_2; \boldsymbol{\lambda})$	$p_{AC}p_{BD}$	$p_{AC}p_{BD}$
$\mathbb{P}(M_1; \boldsymbol{\lambda})$	$1/3$	$1/3$
$\mathbb{P}(M_2 \mid C_1; \boldsymbol{\lambda})$	$1/3$	1
$\mathbb{P}(C_1 \mid M_1, M_2; \boldsymbol{\lambda})$	$p_{AB}p_{CD}$	$p_{AB}p_{CD}/(p_{AB}p_{CD} + p_{BA}p_{DC})$
$\mathbb{P}(C_2 \mid M_1, M_2; \boldsymbol{\lambda})$	$p_{AC}p_{BD}$	$p_{AC}p_{BD}$
$\mathbb{P}(M_2 \mid M_1; \boldsymbol{\lambda})$	$1/3$	$p_{AB}p_{CD} + p_{BA}p_{DC}$
$\mathbb{P}(M_1)$	$1/3$	$1/3$

Table 4.8: Likelihood function for two round CJ assessment

However, it is not clear how the terms in this likelihood formulation, $\mathbb{P}(C_r \mid M; \boldsymbol{\lambda})$ and $\mathbb{P}(M_r \mid M_{r-1}, \dots M_1; \boldsymbol{\lambda})$, may be modelled even knowing the scheduling scheme. For this reason, we might consider instead the pseudolikelihood,

$$\prod_{r=1}^{R} \mathbb{P}(C_r \mid M_r; \boldsymbol{\lambda}) \mathbb{P}(M_r; \boldsymbol{\lambda}) \tag{4.17}$$

Under the assumption of a Bradley-Terry data-generating process,

$$\mathbb{P}(C_r \mid M_r; \boldsymbol{\lambda}) = \prod_{i,j} {p_{ij}^{r}}^{x_{ij}}.$$

In any round, r, the number of comparisons is constrained such that each item is compared once. Thus $P(M_r; \boldsymbol{\lambda})$ is positively related to

$$\prod_{(i,j) \in \mathcal{M}_r} \mathbb{P}({}_rY_{ij} = 1; \boldsymbol{\lambda}),$$

where $\mathcal{M}_r = \{(i,j) : {}_r y_{ij} = 1\}$ is the set of pairs (i,j) that were compared in round r. Under an adaptive scheme, the probability $\mathbb{P}({}_r Y_{ij} = 1; \boldsymbol{\lambda})$ will be higher when the items are closer in strength. It is this insight that is suggestive of why the α-adjustment seems to work better than alternatives.

Here, this is illustrated graphically by computing the average probability of observing at least one comparison between pair (i,j) over the twenty comparison rounds of a Swiss scheduling scheme,

$$\mathbb{P}\left(\sum_{r=1}^{20} {}_r Y_{ij} \geq 1; \boldsymbol{\lambda}\right).$$

To calculate this, we use the simulations from the previous sections and for each pair (i,j) calculate the proportion of simulations where at least one comparison was observed. The calculation is made for each of the three log-strength distributions. Results are shown in Figure 4.15.

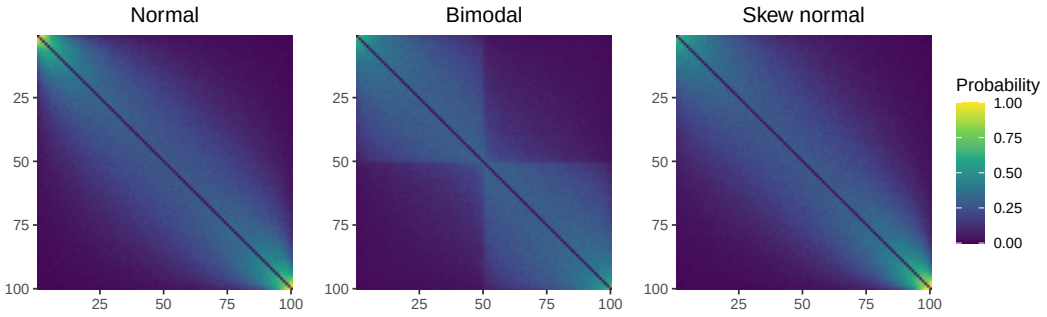

Figure 4.15: Probability of at least one comparison for each pair of items under a 20 round Swiss scheduling scheme. Items shown in strength order from 1, the weakest to 100, the strongest.

Recall that under the dummy penalty approach, we consider a likelihood penalty of

$$\prod_i p_{i0}^{c_0}(1-p_{i0})^{c_0} = \prod_{i<j}(p_{i0}p_{0i}p_{j0}p_{0j})^{c_0/(n-1)},$$

where the dummy item represented by 0 has log-strength of zero. Under the α-adjustment, we consider a likelihood penalty of

$$\prod_{i,j} p_{ij}^{\alpha/(n-1)} = \prod_{i<j}(p_{ij}p_{ji})^{\alpha/(n-1)}.$$

Thus, these likelihood penalties can be decomposed in terms of a value relating to pair (i, j) and compared to the average probability of observing a comparison between that pair. $p_{i0}p_{0i}p_{j0}p_{0j}$ and $p_{ij}p_{ji}$ respectively are plotted for the three log-strength distributions in Figures 4.16 and 4.17.

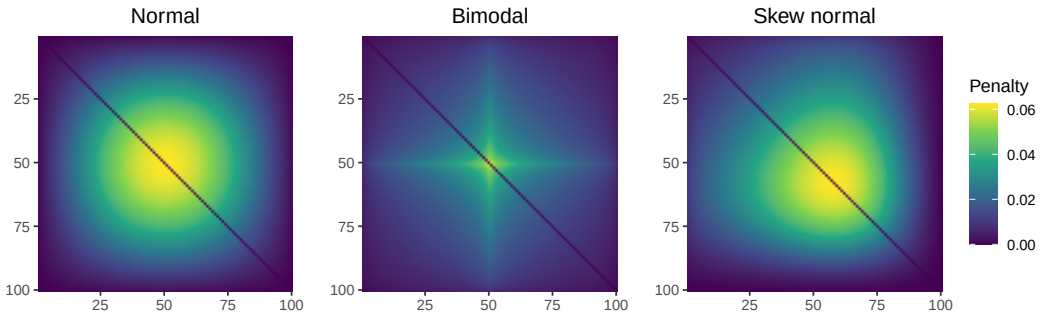

Figure 4.16: Pairwise value of dummy penalty. Items shown in strength order from 1, the weakest to 100, the strongest.

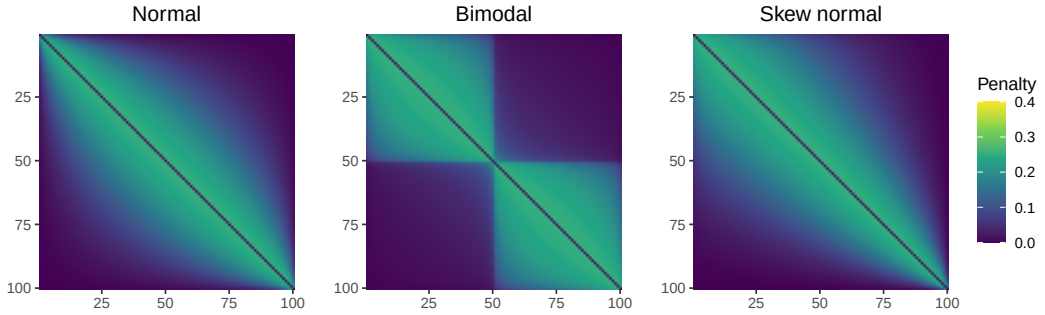

Figure 4.17: Pairwise value of α-adjustment penalty. Items shown in strength order from 1, the weakest to 100, the strongest.

The intensity of the profiles for these penalties is adjustable by varying the parameters c_0 and α respectively, but the profile will remain broadly consistent. There is a clear consistency of profile between the α-adjustment in Figure 4.17 and the probability of observing at least one comparison in Figure 4.15. On the other hand, the profiles diverge much more for the dummy item adjustment seen in Figure 4.16. This may be suggestive of why the α-adjustment, with an appropriately selected value for α seemed to provide better estimates in Sections 4.2 - 4.4 when using the adaptive scheme, since it shows graphically how the penalty that the α-adjustment provides is consistent with the term $\prod_r \mathbb{P}(M_r; \boldsymbol{\lambda})$ from the pseudolikelihood in (4.17).

The other two penalties, ϵ-adjustment and Firth (1993), do not give pairwise likelihood penalties that allow us to compare them in the same way. However, there are independent reasons to expect that they may not be performant for adaptive scheduling schemes. The ϵ-adjustment relies on the ratio w_r/m_r increasing with item strength. But this will be the case to a far lesser degree under an adaptive scheme, where comparisons are more likely between items of similar strength. Under a random scheduling scheme, the w_r/m_r term will approximate the $\sum_j p_{rj}/(n-1)$ term of the α-adjustment. Under an adaptive scheduling scheme, this results in too little bias correction when using the ϵ-adjustment as seen in Figure 4.4. It is possible that the penalty would perform better with a greater value of ϵ, but it seems likely to be highly sensitive to the adaptivity of the scheduling scheme, the number of rounds of comparison and the distribution of the item strengths, which may not all be well anticipated prior to analysis. It therefore seems likely to be an inferior penalty to the α-adjustment.

The generalized form of the Firth (1993) penalty expressed in equation (4.9) shows that it shares this same problem. Recall that Ω_r in that form was the leverage-weighted average of the pairwise probabilities of preference of the observed comparisons. So that an adaptive scheme where the p_{rj} will be closer to 0.5 for all items r will fail to provide a penalty of sufficient strength to items at the extremes of the strength distribution. Alternatively we might interpret this as the Firth (1993) penalty relying on an adjustment based on the asymptotic bias of the unpenalised estimator. The asymptotic bias will be different under an adaptive scheme and so the penalty cannot be relied upon. Consistent with this reasoning, the evidence of Figures 4.4 and 4.5 suggests that it produces a penalty very similar to the ϵ-adjustment.

4.9.3 Information matrix estimation

In this chapter, we have proposed the use of bootstrap measures of reliability, for which only good point estimation is required. However, bootstrap methods are computationally intensive, so it may be desirable to find ways to accurately compute the information matrix under an adaptive scheme, in order to calculate analytic measures of reliability. As was seen in Section 4.3, under the four penalisation approaches tested and a random scheduling scheme, all were able to return credible values for SSR, indicating good error estimation. However, this was notably not the case under the Swiss scheduling scheme, with the Mean Squared Errors shown in Figure 4.8 materially wrong for all except the α-adjustment approach.

Analytic error estimation is dependent on the calculation of the expected information matrix. Recall that the form of the information matrix under the Bradley-Terry

model is

$$i(\boldsymbol{\lambda})_{ij} = \begin{cases} \sum_k m_{ik} p_{ik}(1 - p_{ik}) & i = j \\ -m_{ij} p_{ij}(1 - p_{ij}) & i \neq j. \end{cases}$$

Under the random scheduling scheme, where the schedule is taken as an ancillary statistic and conditioned on in the inference, then the m_{ij} are known and constant. Under an adaptive scheme, $\mathbb{E}[m_{ij}]$ is dependent on $\boldsymbol{\lambda}$.

Figure 4.8 showed that the α-adjustment allowed for reasonably accurate error estimation under the Swiss scheduling scheme and in the tested scenario of 100 items and 20 rounds. Figures 4.15 and 4.17 might suggest that this was because the α-adjustment penalty produced a m_{ij} used in the inference that approximated $\mathbb{E}[m_{ij}]$. This is somewhat speculative however, and especially so where there are fewer comparisons, which is, after all, the aim of using adaptive scheduling schemes.

But there may be circumstances where it is possible to approximate the information matrix more reliably. Given a known adaptive scheme, a number of rounds of comparisons and a number of items being compared, which are all known prior to analysis, $\mathbb{E}[m_{ij} p_{ij}(1 - p_{ij})]$ is a function of $\boldsymbol{\lambda}$, and therefore the information matrix, will depend on the distribution of the item strengths. In general, the log-strength distribution of the items is unknown, even approximately, prior to analysis. It is not uncommon in academic settings for distributions of marks to show multimodality or skew, for example, but for this to be unanticipated. However, where log-strength distributions may be confidently anticipated prior to analysis, it may be possible to discern through simulation, under a particular adaptive scheme and for a defined number of rounds of judgement and number of items, how the term $\mathbb{E}[m_{ij} p_{ij}(1 - p_{ij})]$ relates to the item strengths and thus to calculate an approximate information matrix that may be used for inference. For example, in the large-scale assessment exercises conducted by No More Marking, tens of thousands of items are assessed and distributions from year to year may be observed. If the strength distributions observed are sufficiently stable then these large-scale marking exercises may provide an example where a different approach would be possible.

4.10 Concluding remarks

The work presented here provides recommendations for current practice in CJ and suggestions for future research. Recommendations include:

1. CJ analysts should be encouraged to give greater consideration to their parameter estimation methods. Authors should be encouraged to be more explicit

about these choices and to publish code and appropriately anonymised comparison data along with their work. Here we have demonstrated the strong dependence of conclusions on estimation methods. Therefore, without knowing what estimation procedure was used, it is not possible to adequately verify or challenge conclusions. Publishing data would allow later researchers to apply proposed methods to empirical data sets, accelerating the development of better practice.

2. Adaptive sampling schemes should be encouraged. With good inference practices, they increase the efficiency of CJ assessments, allowing more reliable conclusions to be drawn from the same amount of effort or equally reliable conclusions from less effort.

3. Even under random sampling, it is recommended that parameter estimation use a penalisation method. This is necessary for better point estimation and accurate estimation of errors, with these errors used in measures of reliability such as SSR. Not all penalisation methods are performant (e.g., Facets and the method used in Crompvoets et al. (2020)), but several were found to perform well in the scenario tested (ϵ-adjustment, α-adjustment, dummy adjustment, and that due to Firth (1993)). ϵ-adjustment, α-adjustment, and dummy-adjustment all rely on a constant, the value of which is not strongly suggested by theory. On the other hand, the penalisation method of Firth (1993) is free from such arbitrariness and has appealing asymptotic qualities (Firth, 1993; Kosmidis and Firth, 2009; Kosmidis, 2014). It is therefore recommended that penalisation using the method of Firth (1993) is applied when random scheduling schemes are used.

4. For adaptive scheduling schemes, it is recommended that point estimation be performed using a bias-corrected method based on an initial estimation using an α-adjustment penalty. The value for α can be estimated prior to analysis based on a simulation study using the scheduling scheme and the numbers of items and comparisons that were used to collect the data. The α-adjustment was found to provide an effective penalty providing point estimates with small bias and absolute error, allowing the bias-correction method to work effectively. Applying bias correction allows bias and errors to be minimised further.

5. For error and reliability estimation a bootstrap method is recommended. With adequate point estimation, this provides a credible, flexible and interpretable method. It can also be used to provide things such as item errors, and avoids

concerns around accurate estimation of the expected information matrix inherent to analytic alternatives. In general, given the sparse samples typical in CJ, analytic measures derived from asymptotic assumptions may not always be relied upon.

6. It is recommended that consideration be given to alternative measures of reliability than SSR and split-halves given their dependence on an essentially arbitrary parametrisation and their greater sensitivity to the extremes of the strength distribution. The suggestion made here was to consider a measure, κ, based on Expected Preferences per Comparison (EPC). However, appropriate measures are likely to be context specific. For example, if considering the degree to which CJ is able to accurately assign items within grade boundaries (which is a typical objective of assessments) then something like Krippendorf's α may be more appropriate. Also, if the EPC-based κ measure were to be used, it would require a better understanding of the scale. In contrast, R^2 measures may be more familiar for some researchers and practitioners and, at least in principle, allows for comparison directly to rubric-based marking schemes. An R^2 measure based on EPC using a bootstrap for error estimation would be a possibility if that were the case.

The work here also signals towards a number of interesting possible future research directions. In recommending adaptive schemes in general, it is clear that the selection of such a scheme is a topic of interest. There is relevant literature that may be drawn from for this task. Within the education literature, there is a large body of work on Computer Aided Testing (CAT) See Chapter 5 of Verhavert (2018) for a summary relevant to CJ. But these schemes rarely address pairwise comparison directly. There is some work specific to CJ (for example, Pollitt (2012b); Humphry and Heldsinger (2019); Crompvoets et al. (2020); Verhavert et al. (2022)), but these approaches tend to be heuristic in nature, and lack a theoretical grounding. Further insights from Statistics may also be usefully applied. For example, the scheduling scheme proposed by Glickman and Jensen (2005) and the advocation for Bayesian optimal design of Chaloner and Verdinelli (1995) or Lindley (1972, 1956), or a method for assessing judge reliability based on Dawid and Skene (1979) provide examples of useful proposals. The machine learning literature also deals with related topics. Mikhailiuk et al. (2020) provides a helpful recent summary, and Chen et al. (2016) and Pfeiffer et al. (2012) are perhaps particularly interesting works in the context of CJ. However, much of the work so far is not directly addressable to some of the features of the CJ assessment environment, with the machine learning literature, in particular, tending to be interested in identifying the best items rather than a reliable

rating for all items.

As specific suggestions for future research, questions could include designing a scheduling scheme that:

- best achieves grade categorisation, privileging comparisons close to grade boundaries;
- dynamically optimises with respect to time taken, accounting for the observation that a comparison between two very similar strength items may take much longer than two estimated to have a greater difference in strength;
- takes account of varying judge accuity;
- allows judges to chain judgements, taking in an item that they have previously observed to reduce the overall time taken for judgements;
- allows judges to determine a comparison to be a tie (or to indicate they would have selected a tie were they allowed);
- accounts for the disproportionate influence that a judge providing judgements at the beginning of an adaptively scheduled assessment window may have compared to a judge who provides their judgements later;
- accounts for the time taken to compute and propose subsequent pairings under an adaptive scheme in seeking to maximise efficiency.

Given any particular adaptive scheduling scheme, a credible method of parameter and reliability estimation is required. While Section 4.9 provided some intuition as to how and why the α-adjustment may have performed better than the alternatives examined here in the examples of Sections 4.2-4.8, there is scope for more developed theory on parameter estimation under adaptive schemes, or for empirical work on the applicability of the α-adjustment under different scheduling schemes and with different numbers of items and rounds. Given the sparse data typical of CJ assessments, it is likely that bootstrap methods will remain appealing for both point and error estimation. But, performant adaptive methods are likely to depend on the estimation of item strength for online scheduling and for that task bootstrap methods would be limiting given their computational expense. While it should be noted that the bootstrap methods used here would be readily parallelisable, which may help to mitigate some computational expense, this is unlikely to provide a complete solution.

In this chapter, we have applied the Bradley-Terry model in order to rate and rank the items in terms of their quality based on the pairwise comparisons. As we argued

in Chapters 1 and 3, there can be principled reasons in some contexts for choosing particular models. In this setting, a statistical model offers advantages in being more interpretable than some of the competing alternatives to ranking mentioned in the Introduction to this thesis. However, it is not clear that the round-robin norm is a strong one in the same way it is in the Sports context (as argued for in Chapter 3). As such, it may be that we ought to be prepared to consider other models if they offer pragmatic advantages (see, for example, Glickman and Jensen (2005) and Pfeiffer et al. (2012), where the Thurstone-Mosteller model is applied as it allows computationally cheaper methods to be used for the proposed scheduling stage). In particular, for adaptive schemes, perhaps computationally cheaper spectral methods, such as those discussed in Section 1.7, could be used to approximate Bradley-Terry ratings for intermediate scheduling steps.

Finally it is worth noting that population sizes in CJ assessments are distributed bimodally. Many assessments are small with the number of items in the tens or low hundreds in line with the typical number of students in a particular school or university class or cohort. On the other hand, the national assessments conducted by No More Marking include more than fifty thousand scripts (Wheadon et al., 2020), and this number is growing as their platform becomes more widely adopted. Appropriate answers to some of the further research questions raised here may be dependent on population size and to what degree it is reasonable to make pre-analysis assumptions about the shape of the distribution of item log-strengths as highlighted in Section 4.9.

It is to be hoped that this work provides useful recommendations to improve some current practices in CJ assessment, and a lead into further research. CJ represents a distinct opportunity for statistical researchers in this respect, in being an area that has the potential to combine interesting theory with impactful practice.

Chapter 5

Investigating the 'old boy network' using latent space models

Abstract

This chapter investigates the nature of institutional ties between a group of English schools, including a large proportion of private schools that might be thought of as contributing to the 'old boy network'. The analysis is based on a network of bilaterally-determined school rugby union fixtures. The primary importance of geographical proximity in the determination of these fixtures supplies a spatial 'ground truth' against which the performance of models is assessed. A Bayesian fitting of the latent position cluster model is found to provide the best fit of the models examined. This is used to demonstrate a variety of methods that together provide a consistent and nuanced interpretation of the factors influencing community and edge formation in the network. The influence of homophily in fees and the proportion of boarders is identified as notable, with evidence that this is driven by a community of schools, who have the highest proportion of boarders and charge the highest fees, suggestive of the existence and nature of an 'old boy network' at an institutional level.

5.1 Introduction

'Old boy network' is an English phrase used to refer to the informal system through which men assist other men of a similar socio-economically privileged background, reflected in attending the same school or university. It derives from the term, 'old boy', used to refer to a former pupil at a British 'public school', the confusing name

given to a subset of traditional, and often especially expensive, private schools in England. Generally the old boy network is considered to act at an individual level, but it may be reasonable to think that it will be stronger if and where there are institutional links between schools. However, it is not clear to what degree there are such links and, to the degree that there are, what features of the school might define them. In the English school system, there are a number of distinctions that might be thought to be identifiable in the functioning of such a network, for example private/state, boarding/day, level of fees. In this chapter, latent space models are applied to a network of school rugby union fixtures to investigate whether such functional networks exist and, if they do, what the nature of them may be.

The fixture data provides an interesting data set on which to apply these approaches, for a couple of reasons. First, since one may reasonably expect that a primary consideration in agreeing a fixture would be geographical proximity, then there is a spatial 'ground truth' against which to assess the models considered. Second, there is reason to believe that the set of fixtures may be informative. Fixtures change from year to year but not drastically, and the process of fixtures being scheduled and changed has been taking place for many decades, with the first school rugby union fixture taking place over a hundred years ago. Together these considerations suggest that there has been time for relevant factors to exert an influence such that the observed situation represents a steady state with respect to the schools' current relationships, allowing inferences to be informative.

As with rating, the attempt to detect clusters or communities in networks is a topic that has garnered great attention across a number of areas of academic enquiry including Computer Science, Physics and Statistics. As a result, there are a profusion of available methods (see, for example, Fortunato and Hric (2016) and Javed et al. (2018) for surveys). For the present investigation, where the data have a natural spatial interpretation and there are no definitive membership groups, then latent space models are appealing. Network latent space models position nodes in an unobserved latent space, with the probability of an edge existing between any two nodes being related to the proximity of the two nodes in the latent space. Modelling in this way can allow a number of features common to networks to be captured — transitivity; homophily by attributes; and clustering — as well as often allowing for informative graphical representations.

In the context of networks, transitivity is the phenomenon that two nodes, which each share an edge with the same third node, will have a higher probability of having an edge with each other than a pair that do not. Homophily by attributes describes the greater propensity for an edge to exist when two nodes share observed attributes. For example in a friendship network these could be attributes such as age,

sex, geographical location, or recreational interests. Transitivity and homophily by attributes will both lead to clustering, but it is not uncommon to observe clustering beyond what may be explained by these features. This may be due to homophily by unobserved attributes, self-organisation of actors, the popularity of particular actors, or endogenous attributes such as position in the network (Handcock et al., 2007).

Early latent space network models used multidimensional scaling, and while these captured transitivity and homophily by attributes they were reliant on the arbitrary choice of a distance measure, leading to variable interpretations. Hoff et al. (2002) proposed a stochastic model in which the latent space positions may be estimated through standard statistical techniques. Handcock et al. (2007) extended that model to account for clustering by assuming that the latent positions are drawn from a finite mixture of multivariate normal distributions.

In the present setting, there is reason to believe that there will be clustering beyond that due to transitivity or observed homophily of attributes. First, because there are likely to be factors not captured in available data that account for unobserved homophilies, such as sport orientation within school culture, and social networks between relevant staff. Second, because the fixture information is publicly accessible, which might increase the potential for further clustering through self-organisation; if a school sees that a number of the schools they play have an opponent in common who they do not play they may consider proposing a fixture.

The chapter proceeds in Section 5.2 by discussing the data, with details of its collection, and a brief analysis highlighting some features and inter-relations of the school covariates considered. In Section 5.3, the latent space models of Hoff et al. (2002) and Handcock et al. (2007) are fitted. In Section 5.4, the influence of the different covariates is explored. Section 5.5 provides some concluding remarks.

5.2 Data

5.2.1 Background

The Daily Mail Trophy is an annual tournament between some of the best school rugby teams in England, along with a single school in Wales. In order to qualify for a ranking in the Daily Mail Trophy a school must register for the tournament and compete against a minimum of five other teams in the tournament. There are in total 118 schools included, playing between 3 and 37 matches over the course of the three seasons analysed here. The matches played as part of the tournament are typically only a subset of the matches played by these schools, with other matches taking place as friendlies with non-tournament schools, or as part of a centrally

scheduled knock-out competition. In almost all cases however, if they have played a match against one of the other tournament teams, outside of the centrally scheduled knock-out competition, it would be classed as a Daily Mail Trophy match and would appear in the data set. So the existence or absence of a match with another school is an accurate representation of the bilaterally arranged fixtures within the set of schools in the tournament.

The network to be investigated is based on all matches in the Daily Mail Trophy in the 2015–16, 2016–17 and 2017–18 seasons. Schools who registered but were unable to complete five eligible matches in any given season are maintained in our analysis, despite being excluded from an official ranking. The network being considered is undirected; each node is a school, and each edge has a value equal to the number of seasons, out of the three for which there are data, in which the two schools played each other. This network is considered in relation to school-level data representing variables that could contribute to some observable homophily.

5.2.2 Data collection

The fixtures and results for the Daily Mail Trophy were kindly shared by Schoolsrugby.co.uk, the organiser of the tournament, but are also available at the tournament website (`www.schoolsrugby.co.uk/dailymailtrophy.aspx`). In the analysis that follows it is the number of seasons, out of the three considered, that two teams play each other that are used. There were five instances where teams played each other twice in a season. These are coded the same as if they had played a single time in those seasons.

For each school, data were also collected on:

1. annual fees (Fees)
2. year of foundation (Founded)
3. the number of boys in sixth form (6th Form boys)
4. the proportion of pupils in the school that are boys (Percent boys)
5. the proportion of pupils that are boarders (Percent boarders)
6. whether the school is privately or state-funded (School type)
7. whether the school played one or two terms of rugby (Term type)
8. performance rating of the rugby-playing strength of the school (Rating)

While it would be plausible to consider other variables, these were either not readily available, for example the size of sports bursaries or proportion of pupils from

overseas, or were substantially accounted for by this set of covariates, for example if a school is co-educational throughout or just in sixth form, or not at all, is well captured in the proportion of pupils that are boys and the number of sixth form boys. All covariate data were collected during the first week of June 2019. This means they are not contemporaneous with the fixtures occurring in the period 2015–2018. This is not expected to have materially impacted any conclusions because the covariates are not subject to large year on year changes, .

Fees information was sourced from the schools' own websites. Annual fees were taken to be the minimum fees for full-time education of a pupil in Upper Sixth form, not accounting for bursaries or scholarships. Thus it is zero for state-funded schools, standard day fees for schools that admit day pupils, and minimum boarding fees for schools that admit only boarders.

Year of foundation was generally more readily available on the Wikipedia page for the school than the school's website and so the Wikipedia date was used. This was originally included as it was suspected that it might define a meaningful similarity between schools. But in collecting the data it became clear that the trajectories of schools were very diverse and it was often even difficult to be sure of a definitive foundation date with occurrences of, for example, schools moving geographically, amalgamating, moving from private to state or vice versa, and renaming not uncommon. All of these contributed to subjectivity in the definition of year of foundation. It is included in the analysis however as it provides a useful sanity check to some of the later methods in the degree to which those methods identify year of foundation as a non-informative covariate.

Whether a school was state or privately funded was determined by its classification in the most recent relevant government report (Department for Education, 2019). The source of data on the number of pupils, including the proportions of boys and boarders, and the absolute numbers of sixth form boys, was dependent on the school's status as a private or state school. Pupil numbers for state schools were taken based on the most recent relevant government report (Department for Education, 2019). These did not include a delineation for numbers of sixth form students or of boarders. For this, the most recent available Ofsted report with such data was used (`www.gov.uk/government/organisations/ofsted`). Since these were from previous years and so total numbers differed, the proportions of sixth form students or boarders were assumed to be constant, and the absolute number of sixth form boys was adjusted for the current total number of pupils. The total number of pupils was never materially different from current numbers, and so one may reasonably be confident that these numbers are accurate. However it should be noted that these Ofsted reports were quite commonly from as much as a decade ago.

For privately funded schools, these data were collected from the Independent Schools Council (ISC) website (`www.isc.co.uk`). For boys and for girls, this reports the number of boarders, the number of day pupils, and the number of sixth form pupils. From these the proportion of day pupils, and the proportion of boys are calculated. However, it should be noted that since schools admit pupils from different ages, the relative proportions as they pertain to a consistent age group, say 13–18, are unlikely to be the same. Given the purpose of including these proportions — identifying homophilies — and that no ready alternative was available, then this seems acceptable, as they still provide evidence on the nature of the school. In one case, Ampleforth College, the population information was not available on the ISC website and so the method used was the same as for the state schools, with the most recent Independent Schools Inspectorate report (`www.isi.net`) used, in place of an Ofsted report, to derive the relevant proportions.

In order to determine if the school played one or two terms of rugby, the month of the school's first and last fixtures of the 2018 season were considered, including those fixtures outside the Daily Mail Trophy competition. If these both lay in a single term, then it was interpreted as being a single term rugby-playing school. If there were two or more matches in the second term then it was deemed to be a two term school. If there was a single match outside of first term then previous seasons' fixtures as well as any information from the school's website was used to make the categorisation.

The performance rating used here is described in Hamilton and Firth (2021), applied to an aggregation of the three seasons' results. The method accounts for the varying schedule strengths of participating schools in a manner consistent with the predominant league points system used in rugby union. Taking the projected league points per match were each team to play every other team home and away in a round robin format provides a positive-valued measure of a comparable scale to the other covariates as well as an intuitive interpretation to the rating measure.

For all of these factors, when considering edge covariates the absolute difference is used. The binary variables of school type and term type take value 0 if identical and 1 if different for each pair. The two percentage variables of proportion of boys and proportion of boarders are taken as the absolute difference in these percentages.

Postcodes were taken from the Daily Mail Trophy website and then used via a Google Maps API in order to calculate travel times and distances between schools using the R package googleway (Cooley, 2017) and to project locations onto a relevant map. Distances were calculated as at Saturday 5th October 2019 12pm using the "best guess" methodology, assuming a journey by road. In order to plot schools geographically, latitude and longitude for each postcode was sourced from

`www.freemaptools.com`.

The five continuous variables are scaled to allow for better comparability and interpretability, specifically the following units are used:

Travel Time	hours
Fees	£10,000
Founded	centuries
6th Form Boys	100 boys
Rating	projected league points per match

5.2.3 Exploratory Data Analysis

The two binary covariates of whether a school is privately or state-funded, and whether it plays one or two terms of rugby are compared first. Table 5.1 shows that there is a clear relation, with the private schools substantially more likely to play one term and state-funded schools more likely to play two.

	One term	Two terms
Private	79	17
State	8	14

Table 5.1: Number of schools in the tournament playing one or two terms of rugby and private or state-funded

The other covariates are presented in Figure 5.1. A number of things may be noted. The individual covariate distributions are largely in line with expectations. A number of the covariates are bimodal. This is unsurprising in the case of fees, with one of the modes at zero, or in the case of the percentage of boys, with modes close to 50% and 100%, reflecting a predominance of all-through co-educational and single sex schools, but the bimodal nature to the year of foundation is less intuitive without more historical context. The number of sixth form boys is right-skewed with a mode around 150. The proportion of boarders has a clear mode at just above zero, reflecting a high proportion of day schools and some schools with just a handful of boarders. The remaining proportions are, perhaps not so intuitively, distributed quite evenly up to full boarding status, but without another clearly identifiable mode. Rating is left-skewed with its mode at around 2.7 league points per match, which suggests teams would be expected to share almost one and a half bonus point per match on average. Importantly for our analysis, with the exception of the percentage

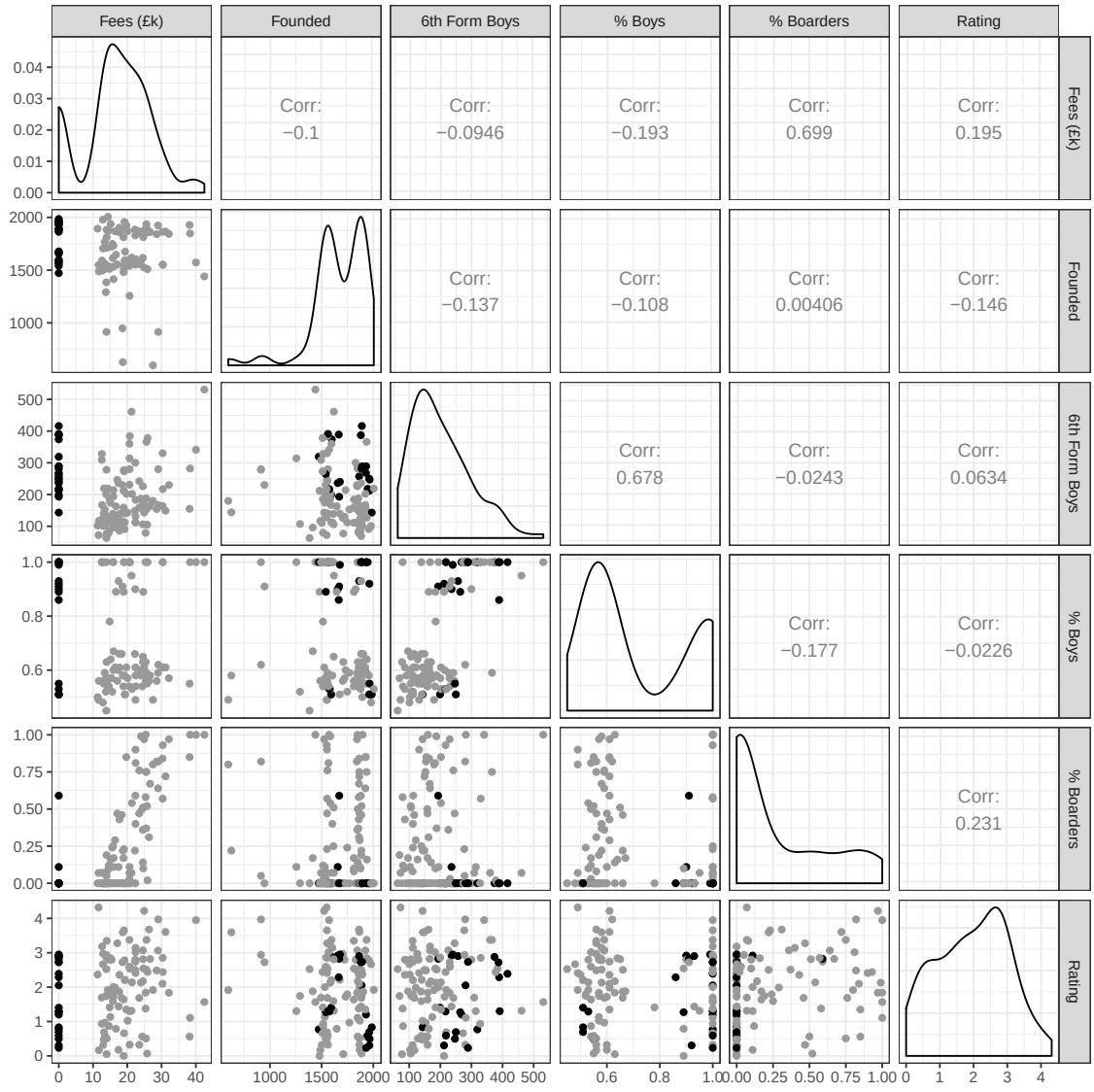

Figure 5.1: Scatterplots of school covariates. Private schools in grey, state-funded schools in black

of boarders and fees, and the proportion of boys and the number of sixth form boys, which have Pearson correlation of 70% and 68% respectively, the continuous covariates all have absolute correlations of less than 25%, which is helpful for being able to discern independent effects. The correlation between the proportion of pupils that are boarders and the fees is not due, as one might first suspect, to boarding fees being higher as they must also account for living expenses, since the minimum upper sixth form fees have been used in all cases, and there are only four pure boarding schools where those minimum fees include boarding fees. In all other cases they represent a day fee. A correlation of 68% between the proportion of pupils that are boys and the number of sixth form boys is less surprising. Here shading has been used in the scatterplots to differentiate state-funded and private schools. Apart from the self-explanatory difference in fees, these also show that within this set of schools the state schools are more likely to have high numbers of sixth form boys, to be single sex (or close to), and to be day schools rather than boarding. We might therefore expect some confounding of these factors in later analyses. These charts were also looked at with a differentiation based on the number of terms of rugby played. This highlighted similar features, as one might expect from Table 5.1, but less strongly.

5.3 Latent Space Model

5.3.1 Model specification

Given the three seasons of fixture data, a binomial latent space model is fitted here, the general form of which is

$$P(A; \boldsymbol{Z}, \boldsymbol{x}, \boldsymbol{\beta}) = \prod_{i<j} \binom{3}{A_{ij}} \mu_{ij}^{A_{ij}} (1-\mu_{ij})^{3-A_{ij}} \tag{1}$$

$$\text{logit}(\mu_{ij}) = \beta_0 + \sum_{k=1}^{p} x_{ijk}\beta_k - d(\boldsymbol{Z_i}, \boldsymbol{Z_j}). \tag{2}$$

Here $A = [A_{ij}]$ is the symmetric adjacency matrix of fixtures with A_{ij} equal to the number of seasons out of the three in which teams i and j played each other, x_{ijk} is the kth edge covariate for teams i and j, β_k is the coefficient for the kth covariate, $\boldsymbol{Z_i}$ is the latent position for team i, and $d(\boldsymbol{Z_i}, \boldsymbol{Z_j})$ is the Euclidean distance between teams i and j in the latent space. This is the binomial version of the model proposed by Hoff et al. (2002).

It is worth noting that a sociality parameter is not included. In this context a sociality parameter would describe the propensity for a particular school to have

matches. It is not included here because the fixtures represented are an incomplete set of fixtures for the participating teams, with teams generally playing additional matches, against teams outside of the tournament, as friendlies or as part of other competitions. While these data are not fully available, the schedule for a sample of teams has been inspected and generally they have played a similar total number of matches, so observing a team to have higher degree within the network is not a reflection of that team having a higher propensity to play matches in general. Including such effects could therefore, for example, misleadingly diminish the extent to which one might infer a lower homophily from the absence of a fixture in the case of teams with lower degree. On the other hand, the current rules of the tournament encourage teams to play as many other tournament participants as possible, since they are awarded bonus points merely for playing matches in the tournament, independent of result. Based on conversations with the tournament organisers and observations of historic results, it is likely that there will be a difference in motivation between teams in their desire to be competitive in the tournament, and some may actively seek to schedule more tournament matches. However there are only a small number of teams, consistent across years, for whom the tournament is a goal in itself. For the vast majority, they enter simply because they can, as a by-product of their standard fixture list, which is largely similar from year to year. So while this effect could be argued to be a genuine sociality effect, within the context of just this network of fixtures, it is likely to apply to only a small minority of teams. Therefore on balance, the distorting impact of inclusion is considered to be more of a danger than that of exclusion and so no sociality effect is included. All models are fitted using the latentnet package in R (Krivitsky and Handcock, 2008).

5.3.2 Hierarchical Clustering

Initially the parameters of the model represented in equations (1)–(2) are estimated through the method of maximum likelihood and with no covariates included.

It was suspected that geographical proximity would be a primary driver of the propensity for a match to occur, and therefore of model distance. Figure 5.2 plots the pairwise Euclidean distances, calculated using the latent space positions, against the pairwise estimated travel times. This shows a strong relationship, with a Pearson correlation of 74%. Travel time is used here as this would seem to be a more relevant motivating condition for a fixture than geographical distance, but substantially similar results are found when using geographical distance. The comparison between latent space distance and travel time may also be used as a means of testing the choice to use a two dimensional latent space. When fitting with three dimensions

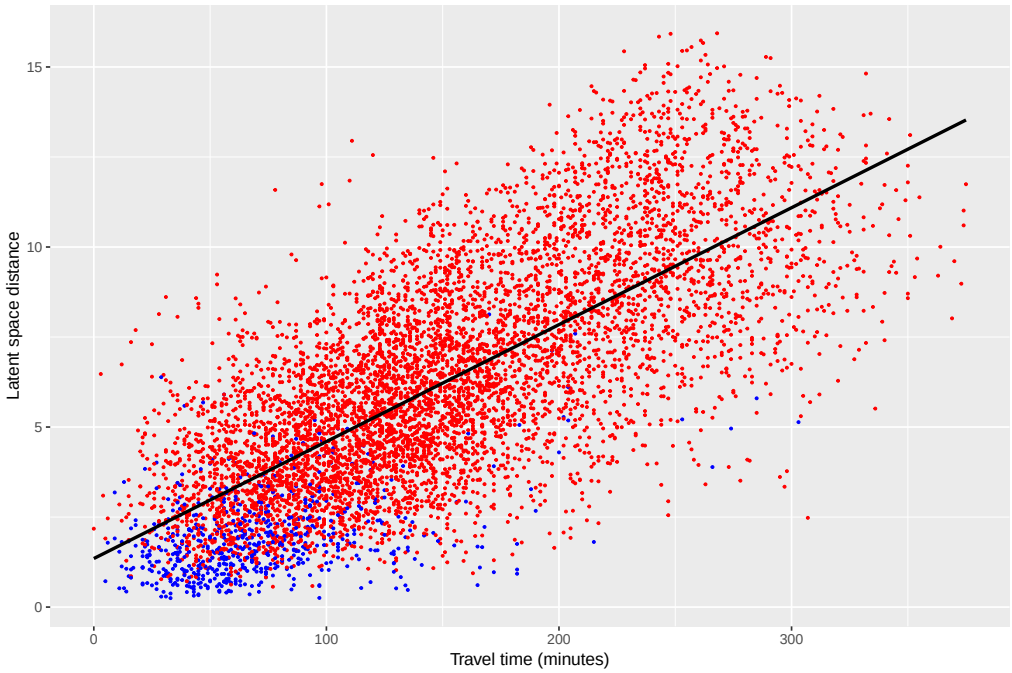

Figure 5.2: Scatterplot of latent space distance against travel time. School pairs who do not play each other during the three seasons are in red, pairs who play each other at least once are in blue. OLS regression line is shown in black.

the correlation increased only by 1%, and goodness of fit measures based on the posterior predictive distribution of degree and minimal geodesic distance (Krivitsky and Handcock, 2008) showed no clear improvement with increased dimension, strongly suggesting that modelling in two dimensions, with the representational benefits it brings, is a reasonable choice. As such, all further models will be applied using a two dimensional latent space.

While the fit is reasonable, there seems to be a notable skew to the residuals. Looking at the residual plot in Figure 5.3 it can be seen that this is driven by two groups. The first has high travel time and considerably lower latent space distance than the linear regression would suggest, the second low travel time but considerably higher latent space distance than the regression would suggest.

It might be supposed that the former group is likely to be due to the requirement for geographically extreme teams to travel longer distances in order to complete a sufficient number of matches in the tournament, so that travel time for them has a different level of consideration than for teams with greater geographical proximity

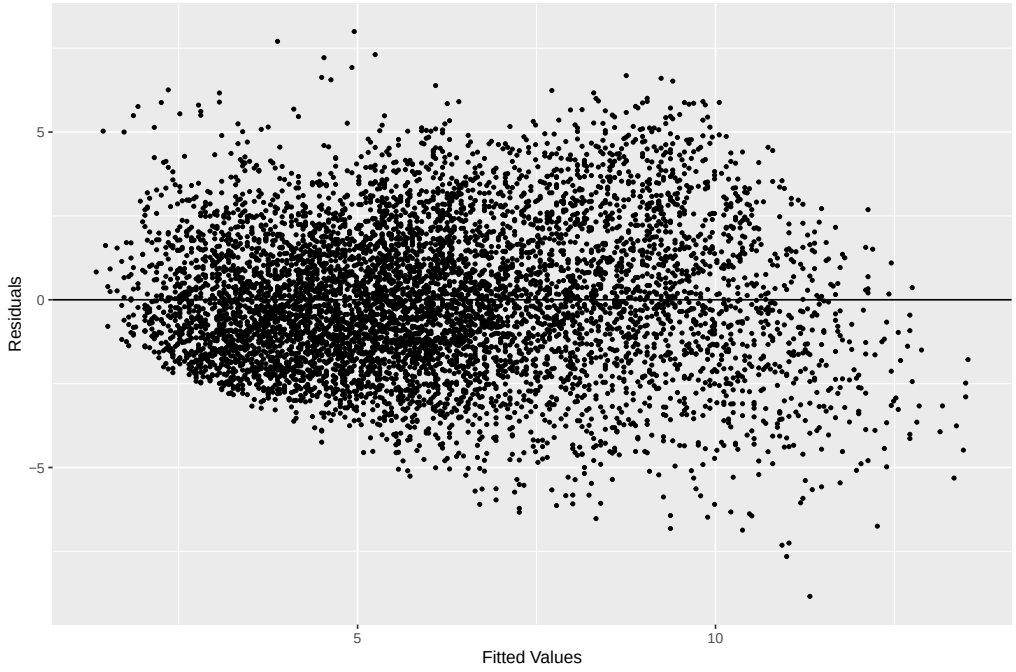

Figure 5.3: Scatterplot of residuals from OLS fit of latent space distance against travel time

Figure 5.4: Location maps for the teams featuring in the extreme residuals. Size of dot represents the number of times that the team appears in the relevant set of pairs. Left hand chart includes pairs where $d < \hat{d} - 4$, and right hand where $d > \hat{d} + 4$, where d is the latent space distance and $\hat{d}$ the expected latent space distance based on the linear regression with travel time.

to other schools. However, Figure 5.4 suggests this does not account for the entire effect. While the most northerly and southerly teams, as well as the single team in East Anglia do stand out in the figure on the left, teams from all over the country are represented and the third highest weighted is King Edward's, Birmingham in the middle of the country, suggesting there is something else at work here. The other group is made up of geographically proximate teams with a large latent space distance. Again it is perhaps to be expected that many of these are in the more densely geographically clustered south east, but some of the most westerly teams also feature strongly. Even in the case of the south east teams, it remains unexplained as to why these particular teams have this greater latent space distance. This will be examined further in Section 5.4 when the influence of the other edge covariates is investigated.

The geographical implications of model distance may also be considered by investigating community detection in relation to the geographical location of the schools. Numerous methods of clustering could be applied given the latent space distance matrix. As an example, by applying a hierarchical clustering with complete linkage the dendrogram presented in Figure 5.5 is obtained.

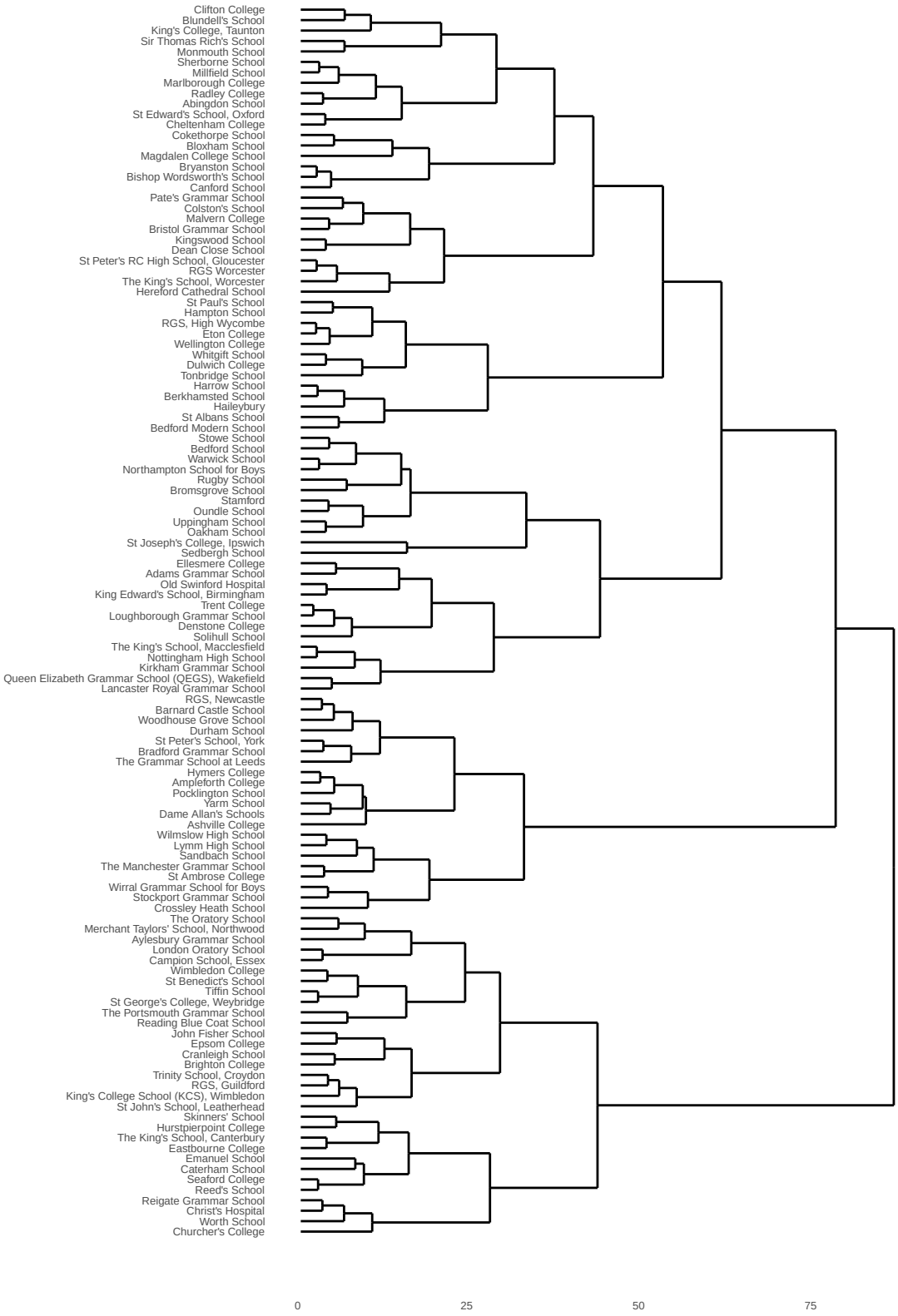

Figure 5.5: Dendrogram of hierarchical clustering of schools by latent space

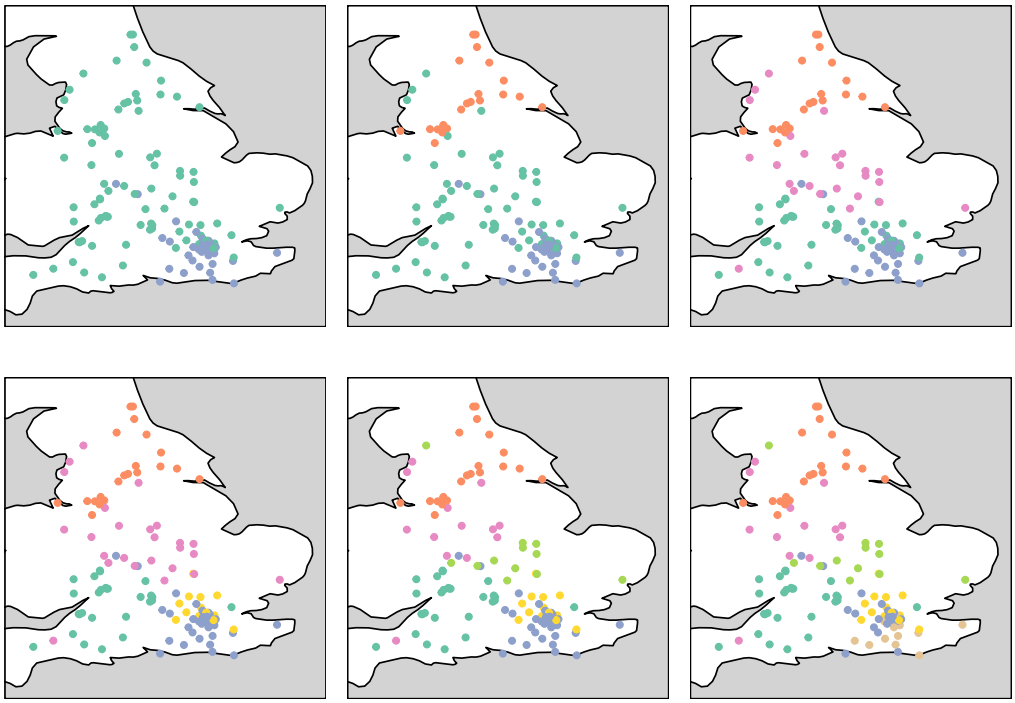

Figure 5.6: Communities of size $G = 2, 3, 4, 5, 6, 7$ based on complete linkage hierarchical clustering of latent space distances

Inspection of the dendrogram suggests that five communities could be an appropriate partition with a number of communities converging at a distance of around 45. In Figure 5.6, communities are plotted, based on the dendrogram, for $G = 2, 3, 4, 5, 6, 7$, where G is the number of groups, in order to show the geographical detection ability with different numbers of groups.

5.3.3 Latent Position Cluster Model

Handcock et al. (2007) provided a model with the extended feature that the latent positions are drawn from a finite mixture of multivariate normal distributions. That is,

$$\boldsymbol{Z_i} \overset{i.i.d.}{\sim} \sum_{g=1}^{G} \lambda_g \text{MVN}_d(\mu_g, {\sigma_g}^2 I_d). \tag{3}$$

Two methods of estimation were proposed in Handcock et al. (2007). The first

method is a two-stage maximum likelihood procedure. In the first stage, the latent positions are estimated in the same way as with the model in Section 5.3.1. In the second stage, a maximum likelihood estimate for the group membership is found conditional on the latent positions calculated in the first step using an EM algorithm (Dempster et al., 1977). This provides a quick and simple estimation. However, by not estimating latent position and clustering simultaneously, it loses information from the clustering that may be useful in the determination of the latent positions. The second method is a fully Bayesian estimation using Markov chain Monte Carlo sampling. In this chapter, this is fitted based on a burn-in period of 10,000 iterations and a sample run of 1,000,000 iterations of which every fiftieth was sampled, giving a sample size of 20,000.

Figures 5.7 and 5.8 present the results from the two methods. In Figures 5.6, 5.7, and 5.8, communities are identified such that the number of schools remaining in the same community as in the previous clustering (as represented by a particular colour) is maximised. The different colours representing different regions in the three Figures is thus a result of different evolutions of the community detection in each case. The fittings using the Gaussian clustering, both based on the two-stage MLE and the MCMC, appear to show better geographic separation than did the hierarchical clustering, particularly when looking at schools close to the west coast. The MCMC clustering arguably shows a better separation in the London area, though there is substantial agreement in the community membership up to $G = 5$.

For the purposes of further investigation it is useful to select a single number of communities, G, with which to work. Graphical inspection of Figures 5.7 and 5.8 suggests that between three and five communities may fit best, given the clear geographical separation they evidence. Handcock et al. (2007) suggest the use of a Bayesian Information Criterion (BIC) for the purpose of selecting the number of communities when using the MCMC fitting. Figure 5.9 shows this BIC for each value of G. Experimentation with different specifications of the algorithm showed the BIC value to be somewhat unstable. As such this is taken to be indicative rather than definitive and so taking into account both the BIC and the geographical separation $G = 4$ is chosen.

In Figure 5.10 the travel time in minutes is plotted against the latent space distance. Due to the nature of the two-stage MLE fit, the left-hand chart is the same as appears in Figure 5.2. It is repeated here next to that from the MCMC estimation to aid comparison. The most notable feature is that the effect of applying the simultaneous fit of latent position and cluster of the MCMC estimation is to reduce and also compress latent space distances particularly for the large latent space distances. The interquartile range of the residuals, for example, is thus more

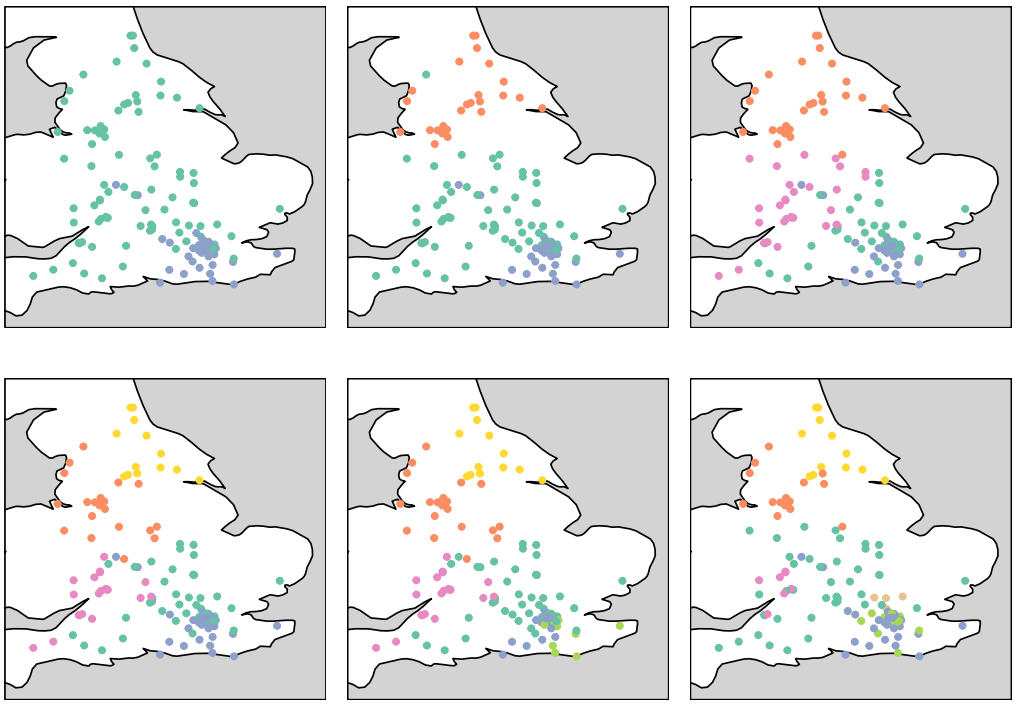

Figure 5.7: Communities of size $G = 2, 3, 4, 5, 6, 7$ based on two-stage MLE

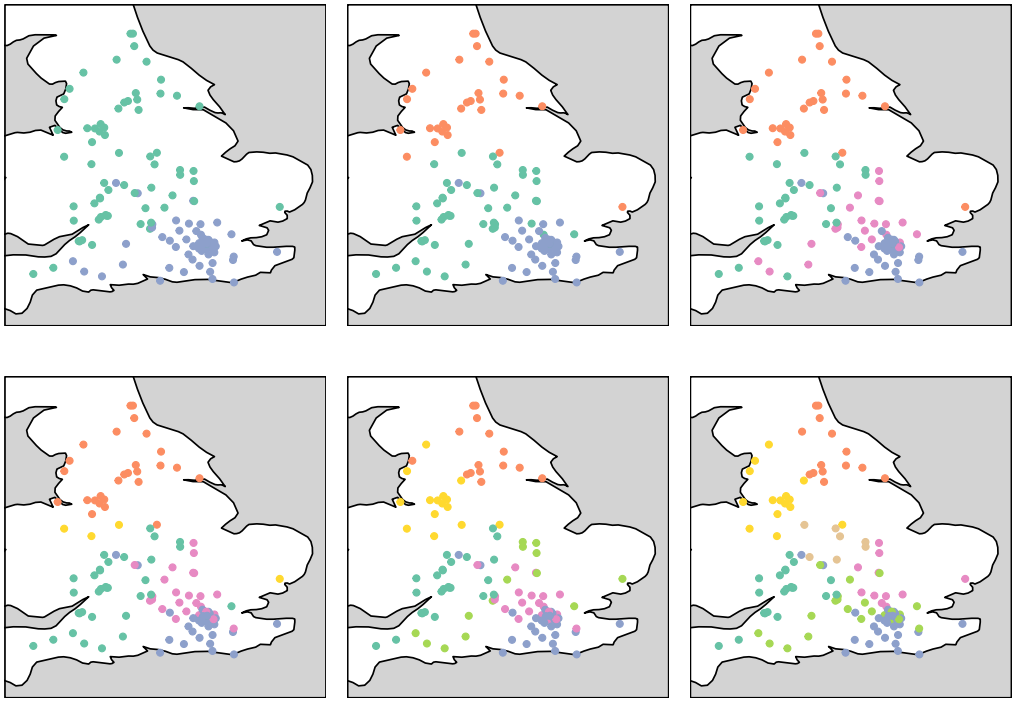

Figure 5.8: Communities of size $G = 2, 3, 4, 5, 6, 7$ based on MCMC parameter estimation

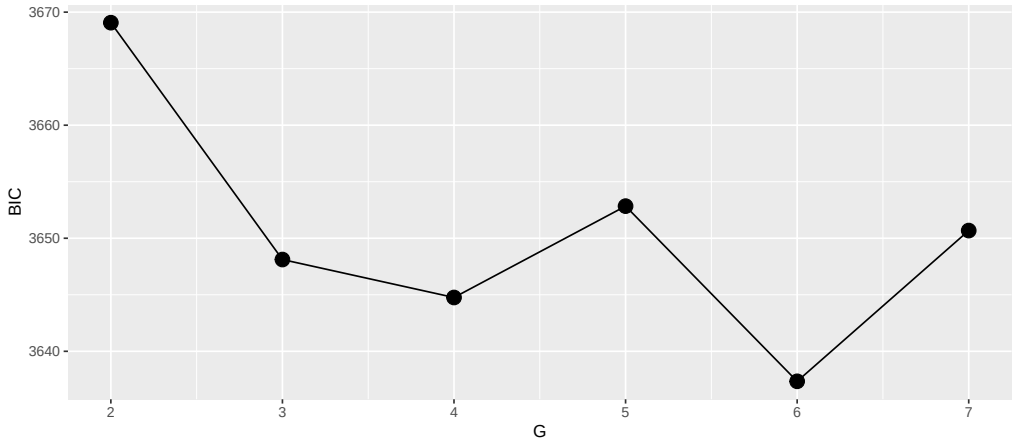

Figure 5.9: Bayesian Information Criterion for different numbers of groups, G.

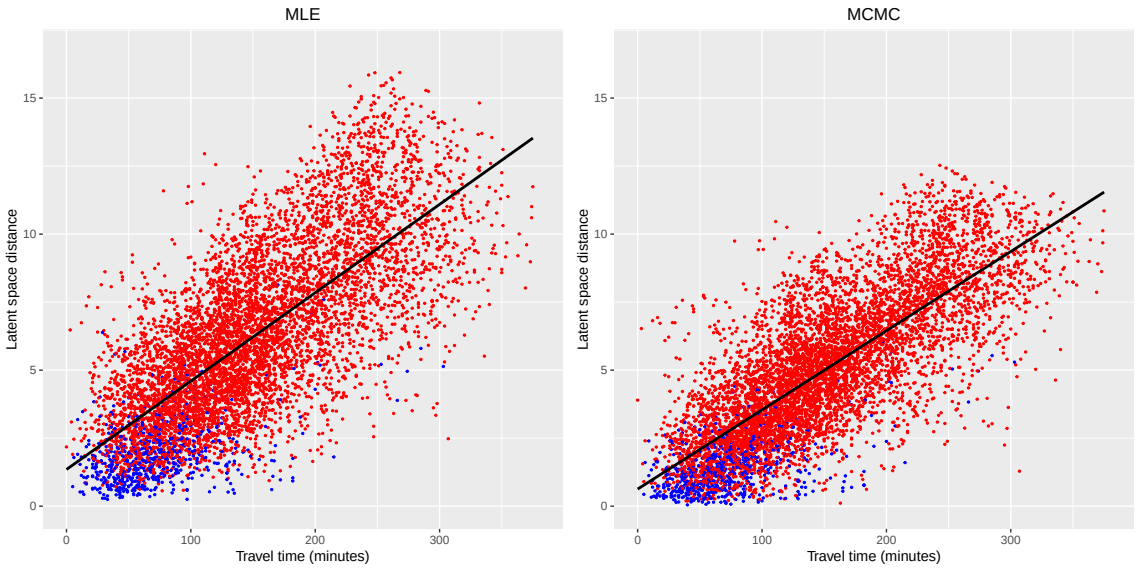

Figure 5.10: Scatterplot of latent space distance against travel time for two-stage MLE and MCMC fits with $G = 4$. School pairs who do not play each other during the three seasons are in red, pairs who play each other at least once are in blue. Linear regression line is shown in black.

than 25% lower when compared to the two-stage MLE fit. The Pearson correlation is slightly higher at 78% compared to 74% from the two-stage MLE fit, and the intercept is closer to zero. This suggests that the assumption of community membership usefully constrains the fit. The community memberships are substantially similar with 108 of the 118 schools belonging to equivalent groups. However based on the seemingly improved latent space fit, the MCMC estimation with $G = 4$ is the model selected for further analysis.

A feature of the mixture model is that it allows us to view the probability of each community membership for each node. Figure 5.11 provides a graphical interpretation of these probabilistic community memberships. Perhaps the most notable feature is the difference in delineation of the communities. What might be referred to as the northern community is most clearly delineated, with all but two members having greater than 75% probability of being part of that community. On the other hand, the two central communities are substantially more indeterminate with the southernmost (that represented in the chart in pink) having no member with a greater than 75% probability of being part of the community. The schools in and

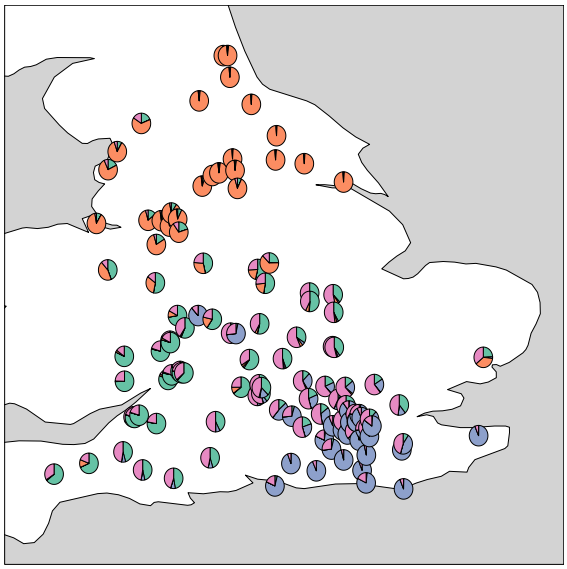

Figure 5.11: Probability of membership of each community.

around London show perhaps more delineation than one might expect given the short distances. Due to the geographical proximity of a number of schools, some are obscured in the chart. This does not alter the interpretation when the distributions of these schools are known. It could also be noted that the most geographically isolated school, St Joseph's Ipswich, has a substantial probability of being in three of the four different groups.

An alternative way of considering the link between the latent and geographical spaces is presented in Figure 5.12. This suggests that the two latent space dimensions are picking up something like a polar coordinate system with a radial coordinate eminating from a centre in the south-east (as seen on the size scale), and an angular coordinate going from the south to the north east (as seen in the colour scale), an observation which is consistent with the probabilistic clustering seen in Figure 5.11, as well as the evolution of the community detection seen in Figures 5.7 and 5.8.

5.4 A consideration of covariates

5.4.1 Relative Importance

The analysis so far has shown a clear link between the travel time and the propensity for the existence of fixtures, but this is to be expected and it is desirable to consider

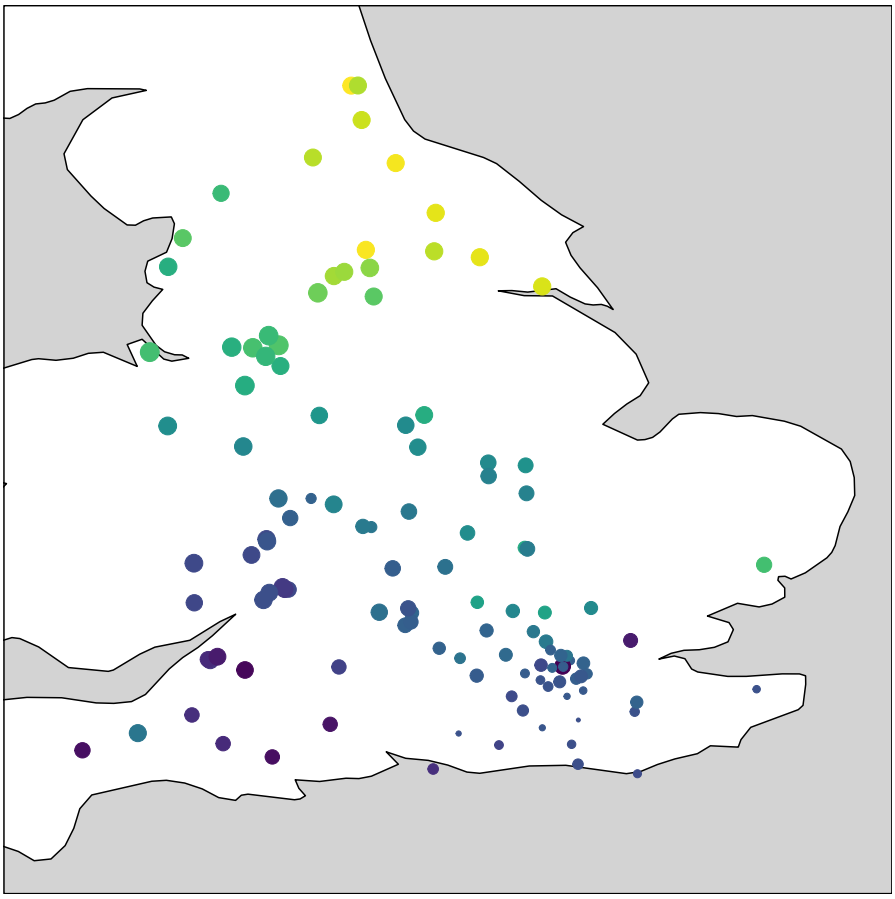

Figure 5.12: Map of schools with colour and size representing first and second latent space coordinate respectively for each school, based on MCMC estimation of latent position cluster model with $G = 4$.

how travel time compares to the other covariates in its importance for fixture propensity. Grömping (2015) provides a thorough overview of various measures of relative importance, and the factors to consider in selecting one. These are presented in the context of linear regression. Some of these measures may be extended to wider families of models, and, in particular, Thomas et al. (2008) proposed relative importance measures for logistic regression that are consistent with some of the features of the measure axiomatically justified by Pratt (1987) for linear regression.

In the present case, the unreliability and the computational expense of the coefficient estimates obtained when the model is fitted with all covariates simultaneously are considerable hindrances to applying relative importance measures directly to the model. This is especially so since a number of the measures require at least $p!$ computations, where p is the number of regressor variables, in order to average over the permutations of regressor variables. One alternative is instead to use the pairwise latent space distances estimated by the selected model. A linear regression of these distances against the edge covariates can then be considered. Doing so provides an estimate for the relative importance of the covariates in the determination of pairwise latent space distance. One reason to be cautious of this approach is that it may be viewed as a measure of relative importance of the covariates only in so much as latent space distance is a good measure of the propensity for a fixture to exist. However if the conclusions drawn from the analysis prove to be corroborated by other approaches then we may have more confidence about the meaningfulness of the results.

Two specific measures are considered here. Both have the desirable and intuitive properties of providing: independence from the order of the regressors in the model; scale invariance; and a proper decomposition of the model variance (R^2) for any orthogonal regressor subgroups. The first (which we will refer to as 'Pratt') is the one advocated by Pratt (1987)[1]. It provides an axiomatic justification, based on appealing ideas of symmetry, for using $b_k\rho_k$ to assess relative importance, where b_k is the standardised coefficient for the kth regressor variable, x_k, and ρ_k is the correlation of x_k with the independent variable y, in this case the latent space distance. The principal objection to this approach is that it can produce a negative measure, which is not clearly interpretable. The second (LMG, after the original authors) is the method originally due to Lindeman et al. (1980), which was also influentially and independently advocated by Kruskal (1987). For each regressor, this takes the marginal increase of explained variance from the addition of the regressor to the model, and takes the mean of these values over all regressor order permutations of

[1]The paper itself is somewhat hard to find but Pratt gives a clear lecture on the subject available at `https://www.youtube.com/watch?v=EzLQkAH5g3A`

the model. Taking the notation of Grömping (2015), denote the explained variance and the sequential additional variance respectively as

$$\begin{aligned} \text{evar}(S) &= \text{var}(y) - \text{var}(y \mid x_j; j \in S), \\ \text{svar}(M \mid S) &= \text{evar}(M \cup S) - \text{evar}(S). \end{aligned}$$

Then we can define, without loss of generality, the measure for the first regressor as

$$\text{LMG}(1) = \frac{1}{p!} \sum_{\pi} \text{svar}(\{1\} \mid S_1(\pi)),$$

where π are the regressor permutations and $S_1(\pi)$ the set of regressors preceding regressor 1 in permutation π. The model has an R^2 of 62%, and the proportion of that coming from each covariate is shown in Table 5.2.

Variable	Pratt	LMG
Travel time	98.09%	97.77%
Percent Boarder	0.88%	0.88%
Fees	0.72%	0.63%
Percent Boys	0.09%	0.08%
Boys	0.08%	0.12%
School type	0.05%	0.30%
Rating	0.04%	0.14%
Term type	0.02%	0.02%
Founded	0.01%	0.04%

Table 5.2: Proportion of explained variance (R^2) attributable to each covariate in linear regression of pairwise latent space distance against covariates. Ordered in descending order of relative importance under the Pratt measure.

It is perhaps unsurprising based on the previous analysis that travel time is dominant, but the degree to which this is the case is nevertheless notable. The other covariates have negligible relative importance in comparison to travel time, but relatively they suggest that fees and the proportion of boarders may have greater influence. The two measures of relative importance assessed here, Pratt and LMG, are very substantially consistent, with Pratt giving slightly higher values for travel time and fees and lower for school type and rating.

5.4.2 Covariate inclusion

An alternative means to investigate the influence of the different covariates is by considering models with each covariate included individually,

$$\text{logit}(\mu_{ij}) = \beta_0 + x_{ijk}\beta_k - d(\boldsymbol{Z_i}, \boldsymbol{Z_j}), \qquad (k = 1, \dots p), \tag{4}$$

where the index k represents the edge covariates. Table 5.3 presents the parameter estimates based on the posterior mean of the coefficient, as well as a statistic representing $q = 2 \times \min(P(\beta_k > 0), P(\beta_k < 0))$, where β_k is the coefficient under consideration, and the Bayesian Information Criterion (BIC) for each model, which may be compared with a BIC of 3647 for the model with no covariates included. Comparing individually in this way thus allows us to consider the importance of a particular covariate based on the change in BIC, and the q-value statistic.

	$\hat{\beta}_k$	q	BIC
Travel Time	-1.523	$< 10^{-15}$	3316
Fees	-0.368	$< 10^{-15}$	3618
Percent boarder	-0.860	$< 10^{-15}$	3630
Term type	-0.309	0.0016	3642
6th Form boys	-0.199	0.0039	3642
School type	-0.370	0.0017	3643
Founded	-0.038	0.0473	3646
Percent boys	-0.756	0.0070	3649
Rating	-0.057	0.3239	3651

Table 5.3: Coefficient, q-value and BIC when model fitted with individual additional covariates. Ordered in increasing BIC.

The BIC estimates are consistent with what was seen in the relative importance analysis in highlighting travel time as the dominant covariate. In that fees and the proportion of boarders are the only others that have a notably lower BIC and also q-value, these are also somewhat consistent with the previous results. With the exception of rating, the q-values are all low, indicating that it is likely that these factors are influencing the propensity for a fixture to exist. This is perhaps surprising in the case of year of foundation, though the q-value there is less conclusive.

Alternatively we may choose to fit the model with all covariates, as in equation 5.

$$\text{logit}(\mu_{ij}) = \beta_0 + \sum_{k=1}^{p} x_{ijk}\beta_k - d(\boldsymbol{Z_i}, \boldsymbol{Z_j}), \tag{5}$$

The results are presented in Table 5.4. The computation of this showed greater sensitivity to the specification of the algorithm, so we might be cautious about the results. They do however show broad agreement with previous analyses, and where there are discrepancies between the coefficient estimates when taken together and individually these may reasonably be thought of as being as a result of the confounding of covariates. For example, the correlation between the percentage of boarders and fees provides an explanation for the smaller absolute impact from boarding when the covariates are considered together.

In this context however it is perhaps more intuitive to interpret the impact of the covariates individually as in the models represented by equation (4), in the sense that knowing nothing else about the schools in each case, one would expect the odds of a fixture to decrease by 78% for every additional hour of travel between schools, by 31% for every £10,000 per annum difference in fees, and by 58% for a fully boarding school (100% boarders) playing a fully day school (0% boarders) as compared to a match between two schools with the same proportion of boarders.

	$\hat{\beta}_k$	q
Travel Time	-1.552	$< 10^{-15}$
Fees	-0.363	$< 10^{-15}$
Percent boarder	-0.317	$< 10^{-15}$
Term type	-0.283	$< 10^{-15}$
6th Form boys	-0.252	$< 10^{-15}$
Percent boys	-0.734	$< 10^{-15}$
Founded	-0.042	0.0478
School type	$+0.154$	0.1541
Rating	-0.002	0.9613

Table 5.4: Coefficient q-values when model fitted with all covariates. Ordered in increasing q-value.

5.4.3 Graphical inspection

Another method of investigation that the model allows is by inspection of the latent space graphically. Given the dominance of travel time, it is useful to control for it in considering the other covariates. Thus the following model is considered with G=4 based on BIC values.

$$\text{logit}(\mu_{ij}) = \beta_0 + x_{ij\text{Travel Time}}\beta_{\text{Travel Time}} - d(\boldsymbol{Z_i}, \boldsymbol{Z_j}), \tag{6}$$

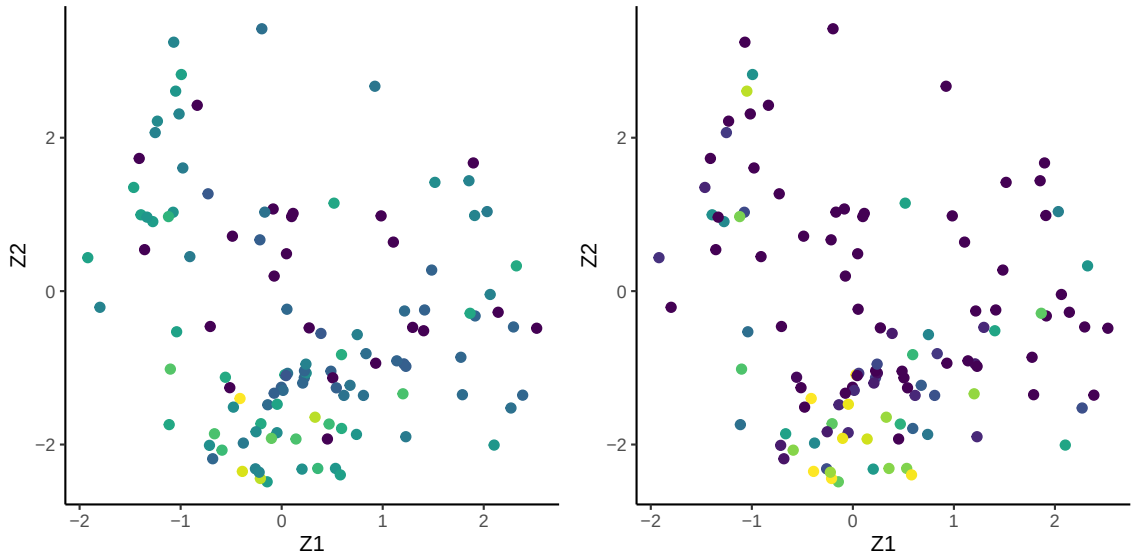

Figure 5.13: Latent space plots based on model with Travel Time controlled for and with G=4. Colour scale representing Fees on left hand side, and % Boarders on right hand side.

In Figure 5.13, the colours are used to show the fees and proportion of boarders for each school. As was noted in Section 5.2, these covariates are strongly linked and this is evident in the plots. It can be seen that while schools with lower fees and low proportions of boarders are distributed quite evenly throughout the latent space, those with higher fees and a high proportion of boarders are to be found disproportionately often in a cluster at the bottom of the latent space. This effect is perhaps clearer for the proportion of boarders. This suggests that there is a small community of schools with a higher proportion of boarders and higher fees, who are more likely to play each other, and that it is this that was driving the greater relative importance of these covariates noted in Sections 5.4.1 and 5.4.2.

5.5 Concluding Remarks

The latent space models demonstrated a number of desirable features in the context of addressing the question of the nature of institutional ties. They showed a good ability to detect geographical communities in the data set, and the spatial nature of the output provided a natural means of interpreting the result in terms of pairwise distances. It was possible to employ a variety of methods in order to analyse the relative importance of the different school homophilies in the propensity for a fixture to exist — by using the methods of Pratt (1987) and Lindeman et al. (1980) to provide quantitative estimates based on latent space distance; by inspection of $q = 2 \times \min(P(\beta_k > 0), P(\beta_k < 0))$ and BIC on fitting models with covariates included; and by using graphical assessment based on latent space plots as in Figure 5.13. These produced a consistent interpretation. The latent position cluster model of Handcock et al. (2007) was found to usefully constrain the latent space model such that it appeared to provide a better fit. It also usefully facilitated a means of assessing the strength of community attribution at a school level by producing the probabilities of each school being a member of each community as seen in Figure 5.11. It is notable that, while a relatively high number of iterations was chosen to be run based on the diagnostics, most of these conclusions were substantially unchanged when running off the default settings of burn-in of 10,000, with 40,000 sampling iterations of which every tenth was sampled giving a sample size of 4,000. The BIC values and coefficient ranges were however notably more affected by changes in other algorithm specifications when run on a lower number of simulations. The latent space models thus performed well in providing quantitative and qualitative insights into the nature of communities and the influential elements of edge formation in this network.

Perhaps unsurprisingly, travel time was found to be the dominant factor in contributing to the prevalence of fixtures. The degree of this dominance is notable,

representing 98% of R^2 based on the two relative importance measures employed. The proportion of boarders and fees were identified as the next most important factors, with the dependency driven by a community of schools with the highest proportion of boarders and the highest fees. It should be noted that the data represent a self-selecting group of the top rugby-playing schools. It seems not unreasonable to expect that were a full set of data available then the impact of some of the other covariates would take on a greater importance. For example, over a wider data set with more state schools and a greater range of abilities, perhaps fees, school type or rating would take on greater relative importance. However, for the question at hand, it is still notable that rather than clustering seeming to correlate with, for example, rating, as meritocrats might have liked to believe, or with school type as others might have suspected, it was a school's status as a majority-boarding school, with a high level of fees, that produced a detectable clustering effect.

There should be caution in drawing social conclusions from this. The analysis might be seen as consistent at an institution level with a traditional understanding of the old boy network, where relationships within a socio-economically privileged group are preferentially advanced. However, the relative importance of the measured homophilies were found to be very small in comparison to travel time and, in particular, school type (private or state-funded) seemed to have little relative importance. To the degree that there was an identifiable effect of schools with a high proportion of boarders and high fees having a greater propensity for fixtures, the tournament organisers offer an explanation. They identify that it is common for schools to seek opponents who can provide fixtures for all the teams they wish to field. The measurable variables of high proportion of boarders and high fees may be indicative of schools where there is a greater expectation that boys will represent the school at rugby union. The mutual desire to be able to field fixtures for all their teams, with as many as five teams in each year group, may lead to these schools having a higher propensity for fixtures. Thus, the greater propensity for these relationships might be argued to be meeting a functional need rather than being based on some prejudicial exclusion. On the other hand, others may argue that even if the motivation is functional rather than prejudicial, this would still have the effect of increasing contact between the schools and the pupils within them due to a commonality of culture that was not shared by other schools representing different demographics, and this may contribute to what gets referred to as the old boy network.

Bibliography

Ali, I., Cook, W. D., and Kress, M. (1986). On the minimum violations ranking of a tournament. *Management Science*, 32(6):660–672.

Altman, A. and Tennenholtz, M. (2005). Ranking systems: the Pagerank axioms. In *Proceedings of the 6th ACM conference on Electronic Commerce*, pages 1–8.

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, 2(3):451–462.

Arrow, K. J. (1963). *Social choice and individual values*. Yale University Press.

Attrill, M. J., Gresty, K. A., Hill, R. A., and Barton, R. A. (2008). Red shirt colour is associated with long-term team success in English football. *Journal of Sports Sciences*, 26(6):577–582.

Audley, R. (1960). A stochastic model for individual choice behavior. *Psychological Review*, 67(1):1.

Baker, D. (2014). Death to the RPI. `https://www.theonlycolors.com/2014/2/26/5444872/death-to-the-rpi`, accessed March 31, 2021.

Baker, S. G. (1994). The multinomial-Poisson transformation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 43(4):495–504.

Barker, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton electron plasma. *Australian Journal of Physics*, 18(2):119–134.

Barrada, J. R., Olea, J., Ponsoda, V., and Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *British Journal of Mathematical and Statistical Psychology*, 61(2):493–513.

Barrada, J. R., Olea, J., Ponsoda, V., and Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement*, 34(6):438–452.

Barrow, M. (2014). Conkers. A favourite children's game. `http://projectbritain.com/conkers.html`, accessed May 18, 2022.

Bartholomew, S. R. and Jones, M. D. (2021). A systematized review of research with Adaptive Comparative Judgment (ACJ) in higher education. *International Journal of Technology and Design Education*, pages 1–32.

Bartholomew, S. R. and Yoshikawa, E. (2018). A systematic review of research around Adaptive Comparative Judgment (ACJ) in K-16 education. *Council on Technology an Engineering Teacher Education: Research Monograph Series*.

Berker, Y. (2014). Tie-breaking in round-robin soccer tournaments and its influence on the autonomy of relative rankings: UEFA vs. FIFA regulations. *European Sport Management Quarterly*, 14(2):194–210.

Bertoli-Barsotti, L., Lando, T., and Punzo, A. (2014). Estimating a Rasch model via fuzzy empirical probability functions. In *Analysis and modeling of complex data in behavioral and social sciences*, pages 29–36. Springer.

Binmore, K. (2008). *Rational decisions*. Princeton University Press.

Bisson, M.-J., Gilmore, C., Inglis, M., and Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education*, 2(2):141–164.

Block, H. D. and Marschak, J. (1960). Random orderings and stochastic theories of responses. In *Contributions to probability and statistics*. Stanford University Press Stanford, CA.

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182.

Bordner, S. S. (2016). 'All-things-considered', 'better-than', and sports rankings. *Journal of the Philosophy of Sport*, 43(2):215–232.

Bradley, R. A. (1965). Another interpretation of a model for paired comparisons. *Psychometrika*, 30(3):315–318.

Bradley, R. A. (1976). Science, statistics, and paired comparisons. *Biometrics*, 32(2):213–239.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Bramley, T. (2007). Paired comparison methods. In Newton, P., Baird, J.-A., Goldstein, H., Patrick, H., and Tymms, P., editors, *Techniques for monitoring the comparability of examination standards*, pages 246–294. London: Qualifications and Curriculum Authority.

Bramley, T. (2015). Investigating the reliability of adaptive comparative judgment. *Cambridge Assessment, Cambridge*, 36.

Bramley, T. and Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 26(1):43–58.

Broome, J. (1991). *Weighing goods: equality, uncertainty and time*. Basil Blackwell.

Broome, J. (2004). Weighing lives. *OUP Catalogue*.

Brown, F. (2018). Modelling player contribution to team success in professional football. Master's thesis, Department of Statistics, University of Warwick. Unpublished.

Bühlmann, H. and Huber, P. J. (1963). Pairwise comparison and ranking in tournaments. *The Annals of Mathematical Statistics*, 34(2):501–510.

Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de corrélation. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 29, pages 77–183.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, pages 273–304.

Chang, R. (2014). *Making comparisons count*. Routledge.

Chebotarev, P. Y. (1994). Aggregation of preferences by the generalized row sum method. *Mathematical Social Sciences*, 27(3):293–320.

Chebotarev, P. Y. and Shamis, E. (1998). Characterizations of scoring methods for preference aggregation. *Annals of Operations Research*, 80:299–332.

Chen, C. and Smith, T. M. (1984). A Bayes-type estimator for the Bradley-Terry model for paired comparison. *Journal of statistical planning and inference*, 10(1):9–14.

Chen, X., Jiao, K., and Lin, Q. (2016). Bayesian decision process for cost-efficient dynamic ranking via crowdsourcing. *The Journal of Machine Learning Research*, 17(1):7617–7656.

Christodoulou, D. (2017). *Making good progress?: The future of assessment for learning*. Oxford University Press.

Christodoulou, D. (2022). *How does Year 2 writing attainment in February 2022 compare with 2021 & 2020?* `https://medium.com/blog-nomoremarking-com/how-does-year-2-writing-attainment-in-february-2022-compare_with-2021-2020-e1ba54dce19`, accessed May 16, 2022.

Coleman, B. J., DuMond, J. M., and Lynch, A. K. (2010). Evidence of bias in NCAA tournament selection and seeding. *Managerial and Decision Economics*, 31(7):431–452.

Colley, W. (2002). *Colley's bias free college football ranking method*. PhD thesis, Princeton University Princeton, NJ, USA.

Cooley, D. (2017). googleway: Accesses google maps APIs to retrieve data and plot maps. *R package version*, 2(0).

Cox, D. R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, 29(2):357–372.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202.

Crompvoets, E. A., Béguin, A. A., and Sijtsma, K. (2020). Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics*, 45(3):316–338.

Crompvoets, E. A. V., Béguin, A. A., and Sijtsma, K. (2021). Pairwise comparison using a Bayesian selection algorithm: Efficient holistic measurement. *PsyArXiv preprint psyarxiv.com/32nhp*.

Cucuringu, M. (2016). Sync-rank: Robust ranking, constrained ranking and rank aggregation via eigenvector and SDP synchronization. *IEEE Transactions on Network Science and Engineering*, 3(1):58–79.

Daniels, H. E. (1969). Round-robin tournament scores. *Biometrika*, 56(2):295–299.

Darmois, G. (1935). Sur les lois de probabilitéa estimation exhaustive. *CR Acad. Sci. Paris*, 260(1265):85.

David, H. A. (1988). *The method of paired comparisons*. Charles Griffin, London, second edition.

Davidson, R. R. and Beaver, R. J. (1977). On extending the Bradley-Terry model to incorporate within-pair order effects. *Biometrics*, pages 693–702.

Davidson, R. R. and Farquhar, P. H. (1976). A bibliography on the method of paired comparisons. *Biometrics*, pages 241–252.

Davidson, R. R. and Solomon, D. L. (1973). A Bayesian approach to paired comparison experimentation. *Biometrika*, 60(3):477–487.

Davies, B., Alcock, L., and Jones, I. (2020). Comparative judgement, proof summaries and proof comprehension. *Educational Studies in Mathematics*, 105(2):181–197.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

De Bacco, C., Larremore, D. B., and Moore, C. (2018). A physical model for efficient ranking in networks. *Science Advances*, 4(7):eaar8260.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Department for Education (2019). Schools, pupils and their characteristics: January 2019. https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2019.

Dittrich, R., Hatzinger, R., and Katzenbeisser, W. (1998). Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(4):511–525.

D'Arcy, J. (1997). Comparability studies between modular and non-modular syllabuses in GCE Advanced level biology, English literature and Mathematics in the 1996 summer examinations. In *Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE*.

Efron, B. (1992). Bootstrap methods: Another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Finn, S. (2009). In defense of the playoff system. *Journal of the Philosophy of Sport*, 36(1):66–75.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

Firth, D. (2022). *Maths. Football. That's all.* `https://alt3.uk/`, accessed September 16, 2022.

Firth, D. and De Menezes, R. X. (2004). Quasi-variances. *Biometrika*, 91(1):65–80.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge University Press.

Floridi, G. and Lauderdale, B. E. (2022). Pairwise comparisons as a scale development tool for composite measures. *Journal of the Royal Statistical Society Series A*, 185(2).

Fogel, F., d'Aspremont, A., and Vojnovic, M. (2014). Serialrank: Spectral ranking using seriation. *Advances in Neural Information Processing Systems*, 27.

football-data.co.uk (2016). *Premier League 2015/2016 Results and Historical Odds*. `https://www.football-data.co.uk/englandm.php`, accessed November 4, 2019.

Ford Jr, L. R. (1957). Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8P2):28–33.

Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1–44.

Frobenius, G. (1912). Über matrizen aus nicht negativen elementen. *Sitzungsberichte der Königliche Akademie der Wissenschaften, Berlin*, 23:456 – 477.

Geyer, C. J. (2020). Stat 8054 lecture notes: Exponential families. `https://www.stat.umn.edu/geyer/8054/notes/expfam.html`, accessed October 4, 2022.

Geyer, C. J., Wagenius, S., and Shaw, R. G. (2007). Aster models for life history analysis. *Biometrika*, 94(2):415–426.

Glickman, M. E. (2013). Introductory note to 1928 (= 1929). In *Ernst Zermelo-collected works/Gesammelte Werke II*, pages 616–671. Springer.

Glickman, M. E. and Jensen, S. T. (2005). Adaptive paired comparison design. *Journal of Statistical Planning and Inference*, 127(1-2):279–293.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, pages 423–453.

Goffin, R. D. and Olson, J. M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science*, 6(1):48–60.

González-Díaz, J., Hendrickx, R., and Lohmann, E. (2014). Paired comparisons analysis: An axiomatic approach to ranking methods. *Social Choice and Welfare*, 42(1):139–169.

Good, I. J. (1955). On the marking of chess-players. *The Mathematical Gazette*, 39(330):292–296.

Good, I. J. et al. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, 34(3):911–934.

Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(2):137–152.

Gumbel, E. J. (1958). *Statistics of extremes*. Columbia University Press.

Haberman, S. J. (2004). Joint and conditional maximum likelihood estimation for the Rasch model for binary responses. *ETS Research Report Series*, 2004(1):i–63.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1):23.

Hamilton, I. and Firth, D. (2021). Retrodictive modelling of modern rugby union: Extension of Bradley-Terry to multiple outcomes. *arXiv preprint arXiv:2112.11262*.

Han, C. (2022). Assessing spoken-language interpreting: The method of comparative judgement. *Interpreting*, 24(1):59–83.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354.

Handfield, T. (2016). Essentially comparative value does not threaten transitivity. *Thought: A Journal of Philosophy*, 5(1):3–12.

Handfield, T. and Rabinowicz, W. (2018). Incommensurability and vagueness in spectrum arguments: options for saving transitivity of betterness. *Philosophical Studies*, 175(9):2373–2387.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Heldsinger, S. and Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, 37(2):1–19.

Henery, R. J. (1986). Interpretation of average ranks. *Biometrika*, 73(1):224–227.

Henriet, D. (1985). The Copeland choice function: An axiomatic characterization. *Social Choice and Welfare*, 2(1):49–63.

Herbrich, R., Minka, T., and Graepel, T. (2006). Trueskill™: A Bayesian skill rating system. *Advances in Neural Information Processing Systems*, 19.

Hill, R. A. and Barton, R. A. (2005). Red enhances human performance in contests. *Nature*, 435(7040):293–293.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Holmes, S., Black, B., and Morin, C. (2017). *Marking reliability studies 2017*. Ofqual.

Hooper, D. and Whyld, K. (1996). *The Oxford companion to chess*. Oxford University Press, USA.

Humphry, S. M. and Heldsinger, S. (2019). A two-stage method for classroom assessments of essay writing. *Journal of Educational Measurement*, 56(3):505–520.

Javed, M. A., Younis, M. S., Latif, S., Qadir, J., and Baig, A. (2018). Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications*, 108:87–111.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4):620.

Jech, T. (1983). The ranking of incomplete tournaments: A mathematician's guide to popular sports. *The American Mathematical Monthly*, 90(4):246–266.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Joe, H. (1988). Majorization, entropy and paired comparisons. *The Annals of Statistics*, pages 915–925.

Jones, I. and Alcock, L. (2012). Summative peer assessment of undergraduate calculus using adaptive comparative judgement. *Mapping University Mathematics Assessment Practices*, pages 63–74.

Jones, I. and Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, 39(10):1774–1787.

Jones, I., Inglis, M., Glimore, C., and Hodgen, J. (2013). Measuring conceptual understanding: The case of fractions. In Lindmeier, A. and Heinz, A., editors, *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education (PME 37)*, volume 3, pages 113–120. International Group for the Psychology of Mathematics Education (IGPME).

Jones, I. and Sirl, D. (2017). Peer assessment of mathematical understanding using comparative judgement. *Nordic Studies in Mathematics Education*, 22(4).

Jones, I. and Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation*, 47:93–101.

Jones, K. (2018). The five factors behind the NCAA's NET ranking system. https://www.si.com/college/2018/11/04/college-basketball-rankings-net-system-explain, accessed October 4, 2022.

Jorgensen, R. (2022). Algorithms and the individual in criminal law. *Canadian Journal of Philosophy*, 52(1):61–77.

Kelly, F. P. (1979). *Reversibility and stochastic networks*. Cambridge University Press.

Kendall, M. G. (1955). Further contributions to the theory of paired comparisons. *Biometrics*, 11(1):43–62.

Kenne Pagui, E. C., Salvan, A., and Sartori, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika*, 104(4):923–938.

Knapp, C. (2007). Trading quality for quantity. *Journal of Philosophical Research*, 32:211–233.

Kolmogorov, A. (1936). Zur Theorie der Markoffschen Ketten. *Mathematische Annalen*, 112(1):155–160.

Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409.

Kosmidis, I. (2014). Bias in parametric estimation: reduction and useful side-effects. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(3):185–196.

Kosmidis, I. (2020). brglm2: Bias reduction in generalized linear models. *R package version 0.6*, 2:635.

Kosmidis, I. and Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika*, 96(4):793–804.

Kosmidis, I. and Firth, D. (2011). Multinomial logit bias reduction via the Poisson log-linear model. *Biometrika*, 98(3):755–759.

Kosmidis, I. and Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika*, 108(1):71–82.

Kosmidis, I., Kenne Pagui, E. C., and Sartori, N. (2020). Mean and median bias reduction in generalized linear models. *Statistics and Computing*, 30(1):43–59.

Krivitsky, P. N. and Handcock, M. S. (2008). Fitting position latent cluster models for social networks with latentnet. *Journal of Statistical Software*, 24.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204–213.

Kruskal, W. (1987). Relative importance by averaging over orderings. *The American Statistician*, 41(1):6–10.

Laming, D. (2003). *Human judgment: The eye of the beholder*. Cengage Learning EMEA.

Langville, A. N. and Meyer, C. D. (2012). *Who's# 1?: the science of rating and ranking*. Princeton University Press.

Laslier, J.-F. (1997). *Tournament solutions and majority voting*, volume 7. Springer Verlag.

Lehmann, E. L. (1953). The power of rank tests. *The Annals of Mathematical Statistics*, pages 23–43.

Leonard, T. (1977). An alternative Bayesian approach to the Bradley-Terry model for paired comparisons. *Biometrics*, pages 121–132.

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement*, 5(1):95–110.

Linacre, M. (2022a). Facets. `https://www.winsteps.com/facets.htm`, accessed June 18, 2022.

Linacre, M. (2022b). Paired comparison of objects. `https://www.winsteps.com/facetman/pairedcomparisons.htm`, accessed June 18, 2022.

Lindeman, R. H., Merenda, P., and Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Scott, Foresman and company.

Lindeman, R. L. (1942). The trophic-dynamic aspect of ecology. *Ecology*, 23(4):399–417.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005.

Lindley, D. V. (1972). *Bayesian statistics: A review*. SIAM.

Loland, S. (2013). *Fair play in sport: A moral norm system*. Routledge.

Lord, F. M. (1959). Statistical inferences about true scores. *Psychometrika*, 24(1):1–17.

Luce, R. and Suppes, P. (1965). Preference, utility, and subjective utility. *Handbook of Mathematical Psychology, III, New York: Wiley*, pages 249–409.

Luce, R. D. (1959). *Individual choice behavior*. Wiley, New York.

Luce, R. D., Ng, C., Marley, A., and Aczél, J. (2008). Utility of gambling I: entropy modified linear weighted utility. *Economic Theory*, 36(1):1–33.

Massey, K. (1997). Statistical models applied to the rating of sports teams. *Bluefield College*, 1077.

Massey, K. (2019). *Massey Ratings*. `https://www.masseyratings.com/`, accessed November 4, 2019.

Maystre, L. and Grossglauser, M. (2015). Fast and accurate inference of Plackett-Luce models. In *Advances in Neural Information Processing Systems*, pages 172–180.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall.

Merrick, J. R., Van Dorp, J. R., Mazzuchi, T., Harrald, J. R., Spahn, J. E., and Grabowski, M. (2002). The Prince William Sound risk assessment. *Interfaces*, 32(6):25–40.

Mikhailiuk, A., Wilmot, C., Perez-Ortiz, M., Yue, D., and Mantiuk, R. (2020). Active sampling for pairwise comparisons via approximate message passing and information gain maximization. *arXiv preprint arXiv:2004.05691*.

Molenaar, I. W. (1995). Estimation of item parameters. In *Rasch Models*, pages 39–51. Springer.

Moon, J. W. (1968). *Topics on tournaments in graph theory*. Holt, Rinehart and Winston.

Moon, J. W. and Pullman, N. (1970). On generalized tournament matrices. *SIAM Review*, 12(3):384–399.

Mosteller, F. (1951). Remarks on the method of paired comparisons. I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16:6–9.

NCAA (2018). *The NET explained.* `https://www.ncaa.com/news/basketball-men/article/2018-11-26/net-explained-ncaa-adopts-new-college-basketball-ranking`, accessed November 4, 2019.

NCAA (2020). *College basketball's NET rankings, explained.* `https://www.ncaa.com/news/basketball-men/article/2020-07-12/college-basketballs-net-rankings-explained`, accessed March 31, 2021.

NCAA (2021). Where the 2021 NCAA bracket stands, 2 days before Selection Sunday. `https://www.youtube.com/watch?v=g5TXz1PPNMw`, accessed March 31, 2021.

Nebel, J. M. (2018). The good, the bad, and the transitivity of better than. *Noûs*, 52(4):874–899.

Negahban, S., Oh, S., and Shah, D. (2012). Iterative ranking from pair-wise comparisons. *Advances in Neural Information Processing Systems*, 25:2474–2482.

Negahban, S., Oh, S., and Shah, D. (2017). Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287.

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment. *Assessment in Education: Principles, Policy & Practice*, 21(2):205–220.

Nitzan, S. and Rubinstein, A. (1981). A further characterization of Borda ranking method. *Public Choice*, 36(1):153–158.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3(1):1–18.

Oberfeld, D., Hecht, H., Allendorf, U., and Wickelmaier, F. (2009). Ambient lighting modifies the flavor of wine. *Journal of Sensory Studies*, 24(6):797–832.

O'Donovan, P., Lībeks, J., Agarwala, A., and Hertzmann, A. (2014). Exploratory font selection using crowdsourced attributes. *ACM Transactions on Graphics (TOG)*, 33(4):92.

Ofqual (2013). Introduction to the concept of reliability. `https://www.gov.uk/government/publications/reliability-of-assessment-compendium/introduction-to-the-concept-of-reliability`, accessed October 4, 2022.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The Pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Pakaslahti, A. (2019). The use of head-to-head records for breaking ties in round-robin soccer contests. *Journal of the Philosophy of Sport*, 46(3):355–366.

Parfit, D. (2011). *On what matters*, volume 1. Oxford University Press.

Paul, R. J. and Wilson, M. (2015). Political correctness, selection bias, and the NCAA basketball tournament. *Journal of Sports Economics*, 16(2):201–213.

Penrose, R. (1955). A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge University Press.

Pfeiffer, T., Gao, X., Chen, Y., Mao, A., and Rand, D. (2012). Adaptive polling for information aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26.

Phelan, G. C. and Whelan, J. T. (2017). Hierarchical Bayesian Bradley-Terry for applications in major league baseball. *arXiv preprint arXiv:1712.05879*.

Pinot de Moira, A., Wheadon, C., and Christodoulou, D. (2022). The classification accuracy and consistency of comparative judgement of writing compared to rubric-based teacher assessment. *Research in Education*, 113(1):25–40.

Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12(5):297–312.

Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 32, pages 567–579. Cambridge University Press.

Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2):157–170.

Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment in Education: principles, policy & practice*, 19(3):281–300.

Pollitt, A. (2015). On "reliability" bias in ACJ: Valid simulation of adaptive comparative judgement. *Cambridge Exam Research, Cambridge, England*.

Pollitt, A. and Elliott, G. (2003). Monitoring and investigating comparability: A proper role for human judgement. In *Invited paper, QCA comparability seminar, Newport Pagnall*. Qualifications and Curriculum Authority, London.

Pratt, J. W. (1987). Dividing the indivisible: using simple symmetry to partition variance explained. In Pukkila, T. and Puntanen, S., editors, *Proceedings of the Second Tampere International Conference in Statistics*, pages 245–260. University of Tampere, Finland.

Pummer, T. (2018). Spectrum arguments and hypersensitivity. *Philosophical Studies*, 175(7):1729–1744.

Qizilbash, M. (2005). Transitivity and vagueness. *Economics & Philosophy*, 21(1):109–131.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society: Series B*, 11:68–84.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43(3/4):353–360.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rachels, S. (1998). Counterexamples to the transitivity of better than. *Australasian Journal of Philosophy*, 76(1):71–83.

Rangel-Smith, C. and Lynch, D. (2018). Addressing the issue of bias in the measurement of reliability in the method of adaptive comparative judgment. In *PATT36 international conference. Research & Practice in Technology Education: Perspectives on Human Capacity and Development*, pages 378–388.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. *Danish institute for Educational Research*.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 321–333.

Reinig, B. A. and Horowitz, I. (2019). Analyzing the impact of the NCAA selection committee's new quadrant system. *Journal of Sports Analytics*, 5(4):325–333.

Revuelta, J. and Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 35(4):311–327.

Robitzsch, A. (2021). A comprehensive simulation study of estimation methods for the Rasch model. *Stats*, 4(4):814–836.

Robitzsch, A. and Robitzsch, M. A. (2022). Package 'sirt'.

Rubinstein, A. (1980). Ranking the participants in a tournament. *SIAM Journal on Applied Mathematics*, 38(1):108–111.

Selby, D. (2020). *Statistical modelling of citation networks, research influence and journal prestige*. PhD thesis, Department of Statistics, University of Warwick. Unpublished.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Slater, P. (1961). Inconsistencies in a schedule of paired comparisons. *Biometrika*, 48(3/4):303–312.

Slutzki, G. and Volij, O. (2005). Ranking participants in generalized tournaments. *International Journal of Game Theory*, 33(2):255–270.

Slutzki, G. and Volij, O. (2006). Scoring of web pages and tournaments—axiomatizations. *Social Choice and Welfare*, 26(1):75–92.

Smead, R. (2019). Sports tournaments and social choice theory. *Philosophies*, 4(2):28.

Stern, H. (1990). A continuum of paired comparisons models. *Biometrika*, 77(2):265–273.

Stern, H. (1992). Are all linear paired comparison models empirically equivalent? *Mathematical Social Sciences*, 23(1):103–117.

Stigler, S. M. (1994). Citation patterns in the journals of statistics and probability. *Statistical Science*, pages 94–108.

Stirzaker, D. (1999). *Probability and random variables: A beginner's guide*. Cambridge University Press.

Stob, M. (1984). A supplement to "A mathematician's guide to popular sports". *The American Mathematical Monthly*, 91(5):277–282.

Strang, A., Abbott, K. C., and Thomas, P. J. (2020). The network HHD: Quantifying cyclic competition in trait-performance models of tournaments. *arXiv preprint arXiv:2011.01825*.

Strimel, G. J., Bartholomew, S. R., Jackson, A., Grubbs, M., and Bates, D. G. M. (2017). Evaluating freshman engineering design projects using adaptive comparative judgment. In *2017 ASEE Annual Conference & Exposition*.

Stuart-Fox, D. M., Firth, D., Moussalli, A., and Whiting, M. J. (2006). Multiple signals in chameleon contests: designing and analysing animal contests as a tournament. *Animal Behaviour*, 71(6):1263–1271.

Su, Y. and Zhou, M. (2006). On a connection between the Bradley-Terry model and the Cox proportional hazards model. *Statistics & Probability Letters*, 76(7):698–702.

Sugden, R. (1985). Why be consistent? A critical analysis of consistency requirements in choice theory. *Economica*, 52(206):167–183.

Suits, B. (1978). *The Grasshopper: Games, Life and Utopia*. Boston: D. R. Godine. reprinted, Peterborough: Broadview Press (2014).

Temkin, L. S. (1987). Intransitivity and the mere addition paradox. *Philosophy & Public Affairs*, pages 138–187.

Temkin, L. S. (1996). A continuum argument for intransitivity. *Philosophy & Public Affairs*, 25(3):175–210.

Temkin, L. S. (2014). *Rethinking the good: Moral ideals and the nature of practical reasoning*. Oxford University Press.

Thomas, D. R., Zhu, P., Zumbo, B. D., and Dutta, S. (2008). On measuring the relative importance of explanatory variables in a logistic regression. *Journal of Modern Applied Statistical Methods*, 7(1):4.

Thomas, T. (2021). Are spectrum arguments defused by vagueness? *Australasian Journal of Philosophy*, pages 1–15.

Thompson, W. and Singh, J. (1967). The use of limit theorems in paired comparison model building. *Psychometrika*, 32(3):255–264.

Thomson, W. et al. (1996). Consistent allocation rules. Technical report, University of Rochester-Center for Economic Research (RCER).

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review*, 34(4):273.

Thurstone, L. L. (1927b). The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, 21(4):384.

Thurstone, L. L. (1927c). Psychophysical analysis. *The American Journal of Psychology*, 38(3):368–389.

Tisi, J., Whitehouse, G., Maughan, S., and Burdett, N. (2013). *A review of literature on marking reliability research*. National Foundation for Educational Research, Slough, UK.

Torres, C. R. and Hager, P. F. (2005). Competitive sport, evaluation systems, and just results: The case of rugby union's bonus-point system. *Journal of the Philosophy of Sport*, 32(2):208–222.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., and De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: Examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*, 26(1):59–74.

Van Den Brink, R. and Gilles, R. P. (2000). Measuring domination in directed networks. *Social Networks*, 22(2):141–157.

Verhavert, S. (2018). *Beyond a mere rank order. The method, the reliability and the efficiency of Comparative Judgemnent*. PhD thesis, Universiteit Antwerpen. Unpublished.

Verhavert, S., Bouwer, R., Donche, V., and De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 26(5):541–562.

Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2018). Scale separation reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement*, 42(6):428–445.

Verhavert, S., Furlong, A., and Bouwer, R. (2022). The accuracy and efficiency of a reference-based adaptive selection algorithm for comparative judgement. In *Frontiers in Education*, page 553. Frontiers.

Vojnović, M. (2015). *Contest theory: Incentive mechanisms and ranking methods*. Cambridge University Press.

Voorhoeve, A. (2008). Heuristics and biases in a purported counterexample to the acyclicity of 'better than'. *Politics, Philosophy & Economics*, 7(3):285–299.

Voorhoeve, A. (2013). Vaulting intuition: Temkin's critique of transitivity. *Economics & Philosophy*, 29(3):409–423.

Voorhoeve, A. and Binmore, K. (2006). Transitivity, the sorites paradox, and similarity-based decision-making. *Erkenntnis*, 64(1):101–114.

Vredenburgh, K. (2022). The right to explanation. *Journal of Political Philosophy*, 30(2):209–229.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3):427–450.

Wei, T.-H. (1952). *Algebraic foundations of ranking theory.* PhD thesis, University of Cambridge.

Wheadon, C. (2015a). *Analysing comparative judgement data in R.* `https://blog.nomoremarking.com/analysing-comparative-judgement-data-in-r-19ac5924602a`, accessed June 7, 2022.

Wheadon, C. (2015b). The opposite of adaptivity? `https://blog.nomoremarking.com/the-opposite-of-adaptivity-c26771d21d50`, accessed September 27, 2022.

Wheadon, C., Barmby, P., Christodoulou, D., and Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy & Practice*, 27(1):46–64.

Whelan, J. T. (2017). Prior distributions for the Bradley-Terry model of paired comparisons. *arXiv preprint arXiv:1712.05311*.

Wobus, J. (2007). Krach ratings. `http://sports.vaporia.com/krach.html`, accessed October 4, 2022.

Yellot, J. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144.

Zermelo, E. (1929). Die Berechnung der Turnier-ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460.