Sports Ranking: 'better than', desert, and why the NCAA have got it wrong

Ian Hamilton, University of Warwick

November 5, 2019

1 Abstract

Ranking is a fundamental part of league-based sports competitions. In this paper we focus on official rankings, those produced by tournament organisers, and look at the philosophical underpinnings of quantitative methods applied to the problem of ranking participants. We consider what is meant by 'better than' and desert in this context and in so doing develop criteria that a ranking method should meet. We use these criteria to evaluate common practice and the most recent ranking method mandated by the NCAA for college basketball. We find that the NCAA ranking plausibly violates all relevant criteria.

2 Introduction

Official sports rankings can have significant impacts on stakeholders, for example the financial impact of participation in post-season tournaments or matches, or on the assessment and therefore job security of those responsible for performance, but perhaps most of all the prestige involved for fans, players and coaches. Little however has been written on the philosophical basis for these rankings, and the implications for the criteria they should encapsulate.

Sports tournaments generally consist of two broad types – league and knock-out. This paper focuses on league-based tournaments, both standalone and in the service of ranking participants in order to determine participation

in a consequent knock-out competition. It considers sports where the unit of competition is pairwise, with matches consisting of one competitor playing another. While much of the discussion may be relevant, tournaments where the schedule is determined dynamically, or that are open-ended, rather than confined to a defined period (referred to as the 'season'), are not explicitly considered here. In particular, we are interested in leagues which we shall define as 'incomplete'. We take a complete league to be one where the ranking rule, for example the aggregation of points (wins) over the course of the tournament, is undisputed by the tournament stakeholders. Incomplete leagues occur in two scenarios. First if assessing a complete tournament at an intermediate stage. Second, as commonly occurs in school or college sports, as a result of historical and geographical factors that deny the possibility of complete tournaments. What makes ranking in this context difficult and disputed, and thus produces incomplete tournaments, is the differing schedule strength of participants. The main components of schedule strength being playing stronger (weaker) teams, playing more (fewer) matches, and playing a greater (lesser) proportion of matches at home.

Ranking in US college sports has seen the most active generation of discussion and ideas on incomplete tournaments. Massey (2019) maintains a record of various ranking methods. For college football, as at 4th November 2019, he lists 82 different rankings. A commonly used means of comparing the validity of these rankings is that based on a 'minimum violations' approach. This assesses the quality of the ranking by considering the proportion of matches in which, under the ranking, a lower ranked team beats a higher ranked team. Rankings that achieve a lower proportion are considered better. If considering rankings at the end of the regular season this may be applied retrospectively by using the matches from the regular season itself, or predictively by applying the measure to the outcomes of the post-season matches.

In applying our conclusions, the common practice of taking minimum violations as a measure of rankings is one aspect assessed, but the use of the NCAA Evaluation Tool (NET), devised in conjunction with Google, and first implemented in the 2018/19 season, is also considered. It is mandated by the NCAA to be used by the committee that determines the participants and seedings in the annual NCAA post-season basketball tournament known as 'March Madness'. The NCAA website describes the NET thus: "NET relies on game results, strength of schedule, game location, scoring margin, net offensive and defensive efficiency, and the quality of wins and losses. To

make sense of team performance data, late-season games (including from the NCAA tournament) were used as test sets to develop a ranking model leveraging machine learning techniques. The model, which used team performance data to predict the outcome of games in test sets, was optimized until it was as accurate as possible. The resulting model is the one that will be used as the NET going forward.' (NCAA.com, 2018).

As a definitional note, 'ranking' is used in this document to describe the ordered position of each team i.e. a positive integer from one to the number of teams in the tournament. A 'rating' is defined as a parameter (or a set of parameters) representing the quality (or qualities) of a team, that in some way allows for ranking. Where we refer to 'points' we mean the points awarded to a team due to the result, for example three for a win, one for a draw in modern soccer. We will use 'score' to refer to the in-game accumulations on which a result is based. Round robin tournaments are those where all competitors play all other competitors an equal number of times. In general we will use definitions and language that accommodate sports where the result outcomes are other than binary win/loss or even win/tie/loss. This does at times lead to more cumbersome definitions than the binary case, but will be useful in highlighting an important point and in ensuring the application is as wide as possible.

The paper proceeds in Section 3 with a discussion of 'better than' in a sports ranking context, and of the argument made by Bordner (2016) against the practice of sports ranking. While in places the assertions of Bordner (2016) are directly challenged, this paper largely assumes the norms of complete sports tournaments provide a reasonable basis for ranking in that context, and instead seeks to define what this ought to imply for incomplete tournaments. A brief justification for this assumption is however provided. In section 4 it goes on to discuss desert as it applies here. These discussions together are used to identify four desirable features that a ranking method should include. In Section 5, through considering the relationship between ratings and rankings, an argument is provided for why we may reasonably consider them as necessary criteria. In Section 6 the criteria are used to assess the minimum violations approach and the NCAA's NET ranking system. Some concluding remarks are made in Section 7.

3 Better than

Bordner (2016) argues that the 'commitment to using rankings to determine particpation in tournaments and the awarding of championships is mistaken'. The argument is applicable to complete as well as incomplete tournaments, though examples cited are all from incomplete tournaments in US college sports. The paper argues that ranking depends on three assumptions:

1. All things considered (ATC): there is some one all things considered quality according to which teams can be ordered from best to worst;

2. Transitivity (TR): if A is better than B (all things considered), and B is better than C, then necessarily A is better than C (transitivity);

3. Winner is Better (WIB): winners are all things considered better than their losing opponents.

It goes on to argue that at least one of these must be false and hence ranking in sports tournaments is a flawed occupation. This argument provides a useful starting point for our purpose. We consider these assumptions in turn.

We refute that the first assumption is necessary. It may be reasonable to claim that it would be preferable for a measurable 'all things considered better than' notion to exist, but that does not make the 'all things considered better than' notion necessary. This can be seen by looking at the norms of complete sports tournaments. In round robin league structures the norm is for 'better than' to be defined as 'has accumulated more points (wins) in sufficiently unbiased pairwise competition with a sufficient representation of other competitors', what we will henceforth refer to as statement (R), in recognition of its retrospective nature[1]. There is no claim on this being an 'all things considered better than' notion. An interesting example is the 2015/16 English Premier League football season. Leicester City were the champions. But they were not favourites for the next season, nor would they have been favourites in matches were they to have played any of their near rivals at the end of the season. Indeed they were not favourites in two of their last

[1]Often this is against all competitors. In professional sports in the USA it is not. Here however fixtures are against a wide representation of other competitors with a home-away balance, and teams ranked against each other in regional conferences generally facing the same opposition.

three league matches of that season against Manchester United and Chelsea (football-data.co.uk, 2016). It therefore seems unlikely that Leicester City would have been considered all things considered better than all the other teams in the tournament. No-one disputed who the champions were however, and they were feted worldwide as such.

We would propose that this normative acceptance of the rankings, of which the determination of the winner is a subset, is a sufficient condition for deeming the ranking criteria appropriate, due to the nature of the outcome of the tournament. The primary outcome of a sports ranking is prestige. Even where there are other outcomes, the opportunity for financial gains or further prestige from future (often knock-out) matches, or professional outcomes from job security for example, they are derived from this prestige. People would not pay a premium to watch, or invest attention in the outcome of, a match between the top teams if they did not believe in some meaningful sense that they were the top teams, and a manager would not get the sack for finishing down the rankings if the team would be held in no lower regard for doing so. Prestige exists as a function of the opinions of stakeholders. With the outcome therefore being of a normative type then a ranking method that is accepted by stakeholders may be considered reasonable.

One might accept this argument and agree that it is mistaken to require (ATC) in complete sports tournaments. Even if that is not accepted, given that we are starting from the point of sports tournament norms as reasonable, and we have demonstrated that they are inconsistent with (ATC), as Bordner himself argues, then we cannot require (ATC) of our incomplete sports tournament. One might object that (ATC), or a variant of it, could remain a desirable feature, but it seems unnecessary even as that if we accept, as we discuss later in this section, a modified version of (WIB).

Turning to the second assumption. Transitivity is definitional for ranking and so is required. It follows immediately from (R), and assuming we can find a method that achieves transitivity then it also makes redundant in this context discussions of incomparability and incommensurability.

Bordner also objects to the third assumption. He considers various versions of this statement, and notes some of the objections that we elucidate below in allowing a relaxation of the strongest one. But ultimately he rejects all of the versions considered. In particular, he rejects what he refers to as 'Overall WIB', which is a form of statement (R), as leading to the repugnant conclusion that a higher ranked team may have lost in all head-to-head matches to a lower ranked team. We reject that this is a repugnant

conclusion at all. First, upsets are a common feature in sport, and even where one had two teams of the same ability one would expect a result, with many sports not allowing for tied outcomes. This is so much a part of sport that the refrain "may the best team win" is a commonplace trope on the eve of matches. So there is a level of arbitrariness to individual results that is expected, and often relished. Second, even if one takes a deterministic approach and denies the possibility of such arbitrariness, one may still point out that this is an outcome under a particular set of external variables. For example, weather conditions and officiating decisions may be considered to have an impact on a single match outcome and are beyond the attributes of the teams.

However there is something important here that we should desire of a ranking method, that 'winning is all things considered better than losing', or in language that respects non-binary outcome sports 'gaining more points is all things considered better than gaining fewer', or even more precisely 'under equivalent conditions, gaining more points is all things considered better than gaining fewer', what we will henceforth refer to as statement (A). This final clarification is in order to have a condition that is not compromised by relative schedule strength. Statement (A) is in a sense foundational to sport, where the goal is clearly defined. It also allows us to avoid additionally including a variant of (ATC) since this preference for winning over losing within the ranking method is necessarily expressed through the 'better than' notion applied, and thus (WIB) effectively acts as a requirement on the 'better than' notion used, and so is, in a sense, itself a variant of (ATC).

It should be noted though that (A) potentially places a preeminence to the result outcome as reflected by points (wins) to the exclusion of other factors, perhaps most notably score. But where the stakeholders of sports feel the need to reflect something beyond result outcome in the 'better than' comparison then this may be done through adjusting the points system. Examples of this include ice hockey, where a point is awarded to the losing team if the loss takes place in overtime; rugby union, where a losing bonus point is awarded for losing within a certain score margin and a try bonus is awarded for scoring a certain number of tries; and county championship cricket where batting and bowling bonus points exist independent of result outcome. This formulation also depends on accepting the points norm of the relevant sport. The reasonableness of those points norms is a separate question, that is not discussed here, other than to note that in major sports these are not a significant point of contention, and that they may be changed

if they are not widely accepted.

We have relied heavily, particularly in refuting Bordner's first assumption, on the norms of complete sports tournaments. It is worth considering these in more detail then, especially since our objective is to understand incomplete tournaments where these norms do not, by definition, exist.

A strong claim on the complete sports tournament norm would be that a league-based sports tournament is normatively one where the 'better than' quality is as defined in (R). That is to say that if we were to construct a new league-based sports tournament our goal should not be to define a pragmatic 'better than' notion encapsulating various desirable features, but to construct a tournament where 'better than' is defined by (R), because that is what a league-based sports tournament is. There are perhaps four pieces of evidence for believing the stronger norm applies:

1. ubiquity – the round-robin format is used worldwide and in almost all sports at some level;

2. uniformity – almost everywhere it is used it is of a similar format with a sufficiently fair distribution of opposition, often every team playing every other team, with (depending on the seriousness of the competition) structural controls for variables that may meaningfully influence the outcome e.g. venue and neutral officiating;

3. undisputedness – the outcomes of round robin tournaments are very rarely disputed as they relate to the structure;

4. uncertainty – 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.

Smead (2019) argues, based on social choice theory, that numerous rankings are available and may be used, and in particular he discusses the Ranked Pair procedure. This might suggest that it is just a matter of historical habit that the round robin tournament has become so prominent. However there are two reasons to doubts this. First there is the geographical ubiquity, where sports with independent histories use the same methodology. Second it is notable that these features do not pertain in other related situations where ranking is frequently performed. For example, multiple political

election methods are applied globally around which there are high levels of dispute. Likewise in sports where the unit of comparison is not pairwise a diverse range of methods are used. Compare for example the aggregation of results in track and field heptathlon/decathlon to Olympic climbing to golf to Formula1 to speedway. There is little uniformity.

A weak claim on the sports tournament norm would be that defining a 'better than' quality by (R) is a pragmatic choice by stakeholders of that tournament that encapsulates various desirable features potentially in a compromised form that will result in the normative acceptance that, as we have discussed, is a requirement in itself. In the previous discussion two of these features have already been identified: transitivity and (A). It is clearly necessary to have both of these. For example, were we to have transitivity alone we could rank based on the average height of the team or the alphabetical order of their names; or if we had (A) alone we could define 'better than' by 'has gained more points in head-to-head matches'[2].

This discussion thus provides two features desirable for a sports ranking method applied to incomplete tournaments: that the ranking method encapsulates a 'better than' notion that is transitive; and that it is consistent with (A). These are met by (R). If the strong form of normativity for round robin tournaments is accepted then these desirable features become redundant. But we face a problem in either applying the strong or weak norms by employing (R), since the 'sufficient representation of other competitors' of (R) definitionally does not exist in incomplete tournaments. A consideration of this problem helps us to identify two further desirable features.

4 Desert

It is useful to consider two 'better than' notions that meet the criteria of transitivity and consistency with (A). The first is that which we have identified and discussed previously: 'has accumulated more points (wins) in sufficiently unbiased pairwise competition with a sufficient representation of other competitors' (R). The other is: 'would be expected to accumulate more points

[2]Temkin ((Temkin, 2012, p.470)) raises the problem of identification of an appropriate competitor set as an objection to using the aggregation of pairwise comparisons for broader general reasoning. Here we consider that the set of competitors is not disputed, which is generally, at least to any meaningful degree, the case in sports tournaments, and so we do not explicitly add it as a condition, but one could readily do so.

(wins) in sufficiently unbiased pairwise competition with a sufficient representation of other competitors', what we will henceforth refer to as statement (P) in recognition of its predictive nature. These highlight another question that must be addressed, that of the preference for a retrospective versus a predictive approach. In the context of an incomplete tournament, (R) is not available since schedule strengths differ and thus the requirement for 'sufficient representation of other competitors' cannot be met, so it would be attractive if we could find conditions under which (P) could be used.

At first sight things do not look hopeful. Feldman (1995) identifies numerous opinions that would appear to negate a predictive approach when considering desert: '... desert considerations are always past oriented. When talking about desert, we are evaluating certain actions which have already happened.' (Sadurski, 1985, p.117); 'Desert can be ascribed to something or someone only on the basis of characteristics possessed or things done by that thing or person. That is, desert is never simply forward-looking' (Kleinig, 1971, p.73); 'Desert looks to the past or at most to the present whereas incentive and deterrence are forward looking notions...' (Barry, 1965, p.111); 'Desert judgements are justified on the basis of past and present facts about individuals, never on the basis of states of affairs to be created in the future. Desert is a "backward-looking" concept ...' (Miller, 1979, p.93). Feldman himself rejects these arguments but provides four arguments of his own for rejecting a predictive approach when discussing pre-punishment. The discussion of desert in the context of pre-punishment does not generally translate across though. There are at least three important differences. First, the nature of intention, which was particularly important in the exchange between New and Smilansky (New (1992), Smilansky (1994), New (1995)) on the subject, is different. We may assume that a team predicted to win does intend to win, but so do all the other teams. An analogy to New's objection to the free will argument, that it does not respect the competitor's intent and therefore that element of their free will, does not therefore apply. Second there is the asymmetry of considering rewards and punishments, with the burden of proof generally being considered higher for the latter. So a probabilistic determination in sports ranking may be acceptable in a way that it would not be where the outcome is punishment. Third, not pre-punishing maintains a status quo, whereas in the ranking context we are assuming that it is not an option to maintain the status quo of no ranking. We could also look at the sports norm for rejecting (P). The round robin makes no claim on future outcomes and is just an aggregation of past results. The case against

a predictive approach thus seems strong.

But what if the prediction embodied in (P) is based entirely on the teams' performances? In that context while (P) is explicitly predictive it is also based solely on things that the teams have done, and so appears to be retrospective at the same time. The prediction in this case may be thought of as just a function for evaluating what has happened, so (P) may be accommodated even if we grant that desert is inherently backward-looking. In the context of the round robin tournament norm, we may consider it a prediction on the points for that season, and we have chosen to structure the tournament to have the results in order to make that prediction easier for a full round robin. So there may be grounds for considering (P) after all. However even those who dispute the idea of a temporal nature to desert grant that there are still likely to be epistemic objections (New, 1992, p.35). As has been noted previously, the outcome here, reward, is likely not to carry the same burden of proof as the one New is discussing, punishment, and so there may be reason to believe that more uncertainty would be permissible. Whether this is the case or not, if the prediction is going to suffice for this purpose it must be generally acceptable to stakeholders who will be granting the reward in the form of prestige.

The practical problems of prediction can be summarised by what variables should be included, and on what data it will be applied. The question of variables can be usefully broken down into three main parts. First, there is a temporal issue. For example, suppose we have a full round robin league and team A win their first six matches and lose their final five, whereas team B lose their first six and win their final five. (R) will place team A above team B, but (P) could very well rank team B above team A. Second, there is the issue of what variables beyond the match result outcomes (win, draw, loss etc.) may be included in producing the ranking. For example, suppose we have two teams in a full round-robin tournament, team A has a record of won nine and lost one, and team B has a record of won eight and lost two. They have played each other twice with one win apiece. Team A's wins have all come by the narrowest score margins possible, and their loss (to team B) by a very wide score margin. Team B's wins have all come by a wide score margin, and their two losses (one to team A), by the narrowest score margin possible. Applying (R) would place team A above team B, but team B may well be ranked higher than team A under (P). Third, there is the issue of respecting the points system of the sport. Suppose we have a soccer tournament, operating under the convention of three points for a win and

one for a draw. We exclude all other variables in our model, including time and margin of victory. Suppose team A finishes with a record of six wins, two draws and eleven losses, and team B with a record of nineteen draws. Then (R) would place team A above team B, but it would seem quite plausible that a ranking with team B above team A would be produced under (P).

These objections can be eliminated for a round robin tournament if the evaluation of (P) for a single season is based solely on aggregated points in that season. With no other variables available the ranking produced would be the same under (P) as under (R), and (P) then conforms to our complete sports tournament norm. Implicit in limiting the ranking to aggregated points is that we should not control for anything in our prediction of the incomplete tournament that is not controlled for in the structure of the complete sports tournament. So if for example team A were ranked highest but only because team B lost matches in freak weather conditions, it would be outside the complete sports tournament norm to claim that a future season would have normal weather and so team B should be champions in the present season, just as it would be inconsistent to consider score margin, or the order of result outcomes. In the incomplete tournament setting, this means that we should control for those variables which are controlled for by the analogous complete tournament structure and only for those variables, and in a way that is consistent with that complete tournament structure.

For example, one of the main biases that will need to be controlled for in the incomplete tournament is venue. It is important to note that the double-header round robin will only control for a venue bias if that bias is symmetric, that is if the (negative) effect on team A playing away at team B is equal to the (negative) effect on team B playing away at team A. This could be the case if there is a fixed effect from playing away from home or if it is a function of the distance travelled for example, but would not be the case, if it were, for example, a function of the difference between the number of fans of team A and team B present or some other team specific effect. Any method under (P) should therefore control for venue, but only under the assumption that the effect is symmetric.

This discussion highlights two additional desirable features. First, that the ranking should be dependent on the results of the present season alone. This has perhaps been implicit in our considerations so far but as we will see when we consider the NET ranking it is useful to make it explicit. Second, that factors controlled for in adjusting for differing schedule strength should be those variables which are controlled for by the analogous complete

tournament structure and only for those variables, and in a way that is consistent with that structure. One of the corollaries of this feature that we have leant on in our reasoning so far is that the ranking produced by the method when applied to the results from a round robin tournament should match the ranking that we would get from applying (R).

We thus have four desirable features for our ranking method:

1. If team A is better than team B, and team B is better than team C, then team A is better than team C. (transitivity)

2. Under equivalent conditions, gaining more points is all things considered better than gaining fewer. (A)

3. There should be dependence on the present season's results alone

4. Factors controlled for in adjusting for differing schedule strength should be those variables which are controlled for by the analogous complete tournament structure and only those variables, and in a way that is consistent with the analogous complete tournament.

5 Ratings vs Rankings

The discussion has proceeded so far on the basis of identifying desirable features. This was because while the case has been made for their desirability, apart from for transitivity, a case has not been made for their necessity. It may be that some or all of them are unachievable, at least mutually. If this were to be the case then we would be faced with a choice between relaxing one, some or all of the features and being able to still produce a ranking, or insisting that they should be viewed as necessary criteria and thus rejecting a ranking altogether, as Bordner does. In both scenarios they would remain desirable, but under the latter choice we would have determined that their desirability was strong enough to consider them necessary. We thus next consider their mutual achievability. If we can show that there exists at least one method by which they may be achieved then we may consider them necessary criteria, since no other method that does not achieve them could then be considered as desirable.

Here the distinction between ratings and rankings becomes useful. A rating was defined as a parameter (or a set of parameters) representing the

quality of a team, that in some way allows for ranking. Therefore a set of ratings will always carry at least as much information as a set of rankings of the same size, as by definition it will allow a ranking, but can also in addition convey information on the relative closeness of any two competitors, even if we restrict it to being a single parameter. Thus we should always prefer the production of ratings. There is no theoretical reason not to allow the rating to be multi-dimensional. It is interesting to note that doing so would allow for the non-transitive cyclical triad of team A is expected to beat team B, team B is expected to beat team C, team C is expected to beat team A, but still allows us to construct a ranking under (P) that would meet the requirement of transitivity. There may however be reasons of statistical practice not to do this since the number of matches may not justify the notably higher number of parameters from doing so.

Perhaps the most obvious way in which the calculations for (P) may be made would be by choosing the ratings such that a pairwise combination of any two ratings will define the probabilities of the relevant result outcomes for that pair of teams. We may then impose the condition that the expected number of points gained by each team using these probabilities is equal to the actual number of points gained. This may be extended to the incomplete tournament by conditioning on the fixtures played. So that we have that the expected number of points gained given the fixtures played is equal to the actual points gained. Statistically tractable models of this nature may readily be built, and they address all the identified features. Thus there is no reason to relax the requirement for these features in any ranking method and they may be thought of as necessary criteria instead of simply desirable features.

It is perhaps worth considering here a popular rating, known as the Ratings Percentage Index (RPI), which was the one replaced by NET. RPI explicitly seeks to take account of the strength of opposition faced by providing a score conventionally composed of

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

This fails to meet criteria 4 in two ways. First it fails to take into account home advantage, though some versions attempt to adjust for this. Second if we consider the simplest possible round robin tournament, with two teams

playing each other once, then both will have an RPI score of 0.5, and will thus be tied, whereas under (R) the winner of the match would be ranked higher. This is an indication of a common criticism, that RPI overweights the strength of the opposition compared to the outcome of matches.

6 Implications for current practice

The minimum violations measure is very commonly used in comparing ranking methods. A 'violation' is defined as a result where a higher ranked competitor is beaten by a lower ranked competitor. The measure prefers ranking methods that produce the lowest number of violations. By comparing ranking methods by a minimum violations measure one advantages rankings produced by a minimum violations approach itself, that is a ranking that has been optimised to minimise the number of violations. An example can be used to demonstrate how this is undesirable. Consider a full round robin league consisting of teams A-E, who play each other exactly once at a neutral venue. Let us also assume two points for a win and one for a draw. The results are summarised in Table 1. Team A draws with B, beats C and D, and loses to E. Team B beats C and D and loses to Team E. Team C draws with D but beats E, Team D beats E.

Team	B	C	D	E
A	$\sim$	A	A	E
B		B	B	E
C			$\sim$	C
D				D

Table 1: Results table showing winners of matches between any two teams. Draws shown as $\sim$

So a conventional ranking table would be as in Table 2. [3] The minimum violations ranking by contrast would have placed C first or last, but never in the middle. This violates criteria 1 and 4. They violate criterion 1, transitivity, by producing a non-unique ranking, where C may be better than A and B (D and E), or A and B (D and E) may be better than C,

[3]Note that this ranking would not change under the convention in modern soccer of three points for a win and one for a draw, just that the points would change to 7,7,6,4,4 instead of 5,5,4,3,3 respectively.

Position	Team	W	D	L	Points
1=	A	2	1	1	5
1=	B	2	1	1	5
3	C	2	0	2	4
4=	D	1	1	2	3
4=	E	1	1	2	3

Table 2: Conventional ranking

and violates criterion 4 by producing a ranking that disagrees with (R) in a round robin tournament.

It is not totally transparent what NET does and how it defines 'better than' but the description provided on the NCAA website allows some conclusions to be drawn. Explicitly it relies on 'net offensive and defensive efficiency, and the quality of wins and losses' which are outside of the simple points (wins) reliance, and since these are not factors that would be controlled for in any comparable round robin this would seem to violate criterion 4. Additionally, in training the machine learning algorithm, a dependency is created on previous seasons' activity that compromises the dependence on just the present season's results, violating criterion 3. This also may violate criterion 2, (A). In order to see this, suppose that in previous seasons a high net offensive and defensive efficiency rating was found to be the best indication of post season success, then in the present season a team who had a higher efficiency rating, due to some very wide victories against a few opponents, could be ranked higher by NET than a team who had played exactly the same opposition and had more wins but had a lower net efficiency rating.

It may even be the case that NET has a dependence on minimum violations approaches. NET was reportedly developed after consultation with the developers of 'various prominent indices' (NCAA.com, 2019b). But given the prominence of the minimum violations approach, it is plausible that these would have been determined to be prominent based significantly on their success under the minimum violations measure. This description from the NCAA website hints at this being the case (italics ours):"The primary component of the NET is the TVI, a results-based factor that considers the strength of the opponent and the location of the game. If you beat a team that you're *expected to beat*, then it doesn't do as much for your ranking. Losing to teams that you were *expected to beat* will hurt your ranking." (NCAA.com, 2019a). The binary nature of the phrase 'expected to beat' fits better the description

of minimum violations approaches than other alternatives. This is somewhat speculative, but if this were the case, then NET may also be leaning heavily on a method that violates criterion 1, transitivity.

7 Concluding remarks

We have noted in this discussion that there exists a ranking method in incomplete sports tournaments that does meet the identified criteria. One criticism of that approach is the potential complexity. Implementing a method in line with that suggested in our discussion in Section 5 is likely to involve computational methods not transparent to the typical stakeholder. It was not clear that transparency was a fifth desirable feature, since it is hard to conceive of a method that meets our four criteria in a round robin setting but not transparency. But it seems reasonable that depending on the tournament scenario it may well be that transparency is something that is valued. As we noted, since the reward is prestige-based then the acceptance of stakeholders is vital and that may be conditional on a degree of transparency, and so there may be scenarios where it is undesirable to implement this method directly. Even if that is the case, the method can still be useful. First it gives clear guidance for what should and should not be considered within a simplified more transparent ranking model. Second it may be used to assess the quality of more transparent approaches, with better methods yielding rankings more similar to this ranking method. Third if transparency is included as an explicit criterion then there will need to be some compromise on at least one of the criteria and perhaps transparency is the best one to compromise on.

An alternative criticism might be the opposite of this, presumably that supported by proponents of NET, that it is not complicated enough. That is to say that it could be argued that additional complexity will be required to form a ranking in the situation of an incomplete tournament in any case, and so once we are considering more complex approaches then our sensitivity to increasing the complexity in any number of ways, e.g. including new variables and basing the ranking on even less transparent machine learning approaches, may be lessened since transparency has already had to have been sacrificed. But this violates the complete sports tournament norm, and if that is to be done then that is something stakeholders should be agreeable to, having understood the issues.

Throughout this paper we have argued for basing any approach on the

norms established by complete sports tournaments. It is worth noting that the applicable norm in any specific situation will be a function of time, location and stakeholder community, and while currently one of the strengths of the round robin norm is its apparent global ubiquity and uniformity, it would be possible for local norms, or even the global one, to change. Perhaps a new generation of stakeholders familiar with the less simplistic and more dynamic ratings of eSports, or with being rated in a predictive sense by online engines, may be more comfortable with a norm other than the round robin, or with a relaxation of specific features such as a lack of transparency.

But as it stands, the arguments presented here provide grounds for objecting to both a standard method applied for assessing quantitative rankings in the form of minimum violations, and, perhaps more importantly, for the official procedures of the NCAA. It is interesting to wonder how the NCAA came to make their decision to use a predictive machine-learning based approach. It seems likely to have been partly a result of the lack of a consistent voice amongst the stakeholder community about how this should be done. Perhaps a coherent analysis from a philosophical perspective can assist in this. Our suggestion is that the NCAA rethink their use of NET and instead move to an approach meeting the criteria identified here. This should be accompanied by a careful explanation to stakeholders of the principled justification for the alternative approach consistent with the discussion presented. Organisers of other tournaments, for example school leagues of super-regional size, may also want to consider the conclusions presented here and their own ranking methods.

References

Barry, B. (1965). *Political Argument*. Humanities Press.

Bordner, S. S. (2016). 'all-things-considered,''better-than,'and sports rankings. *Journal of the Philosophy of Sport*, 43(2):215–232.

Feldman, F. (1995). Desert: Reconsideration of some received wisdom. *Mind*, 104(413):63–77.

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.

Kleinig, J. (1971). The concept of desert. *American Philosophical Quarterly*, 8(1):71–78.

Massey, K. (2019). *Massey Ratings*. https://www.masseyratings.com/, accessed November 4, 2019.

Miller, D. (1979). *Social Justice*. OUP Oxford.

NCAA.com (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.com (2019a). *Get to know the NET rankings — and what they mean for the NCAA tournament*. https://www.ncaa.com/news/basketball-men/article/2019-02-07/net-rankings-ncaa-tournament-what-they-mean, accessed November 4, 2019.

NCAA.com (2019b). *NET rankings: What to know about college basketball's new tool to help select the NCAA tournament field*. https://www.ncaa.com/news/basketball-men/2018-08-22/net-rankings-what-know-about-college-basketballs-new-tool-help, accessed November 4, 2019.

New, C. (1992). Time and punishment. *Analysis*, 52(1):35–40.

New, C. (1995). Punishing times: reply to smilansky. *Analysis*, 55(1):60–62.

Sadurski, W. (1985). *Giving Desert Its Due: Social Justice and Legal Theory*. Springer.

Smead, R. (2019). Sports tournaments and social choice theory. *Philosophies*, 4(2):28.

Smilansky, S. (1994). The time to punish. *Analysis*, 54(1):50–53.

Temkin, L. S. (2012). *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford University Press, USA.