Introductory note to 1928

Mark E. Glickman*
Department of Health Policy and Management
Boston University School of Public Health

Zermelo's 1928 paper on measuring participants' playing strengths in chess tournaments is a remarkable work in the history of paired comparison modeling. Apart from several contemporary papers by Thurstone (1927a, 1927b, 1927c), Zermelo's paper was an isolated excursion into paired comparison methods that was far ahead of its time. After this paper, the field remained mostly dormant for about 25 years until seminal publications by Bradley and Terry (1952) and Mosteller (1951a, 1951b, 1951c) led to increased interest in paired comparison methodology. Subsequently, work in paired comparison methods has remained an active area of research since the 1950's, and with paired comparison applications to large data sets including marketing research problems involving choice preference among consumers, the field continues to be active. Scholarly work in this area arguably led to popularizing paired comparison methods in the form of competitor rating systems. For example, in the United States, a rating system due to Elo (1978) for chess players was implemented in the early 1960s (supplanting a fairly crude system developed in the 1950s) which had as a basis the model appearing in Zermelo's paper. The system was adapted to ensure computational simplicity for the analysis of large collections of players whose abilities were changing over time. Elo's system eventually was adopted by the World Chess Federation

*Address for correspondence: Center for Health Quality, Outcomes & Economics Research , Edith Nourse Rogers Memorial Hospital (152), Bldg 70, 200 Springs Road, Bedford, MA 01730, USA. E-mail address: mg@bu.edu. Phone: (781) 687-2875. Fax: (781) 687-3106.

in the early 1970s and is still used today to rate chess players in international tournaments. Given that the basis of Elo's system can be traced back to Zermelo, tournament chess players both in the United States and elsewhere are therefore indirect inheritors of Zermelo's work on rating chess players.

Interestingly, despite sharing the same probability model and developing nearly identical numerical algorithms, the prominent paired comparison researchers from the 1950s, as well as later authors including Elo, may not have known about Zermelo's paper, as they did not cite it. The earliest citation in paired comparison literature that I could find was by Good (1955) who limits mention to a property of his own model being shared by that of Zermelo. Only in Herbert David's (1988) monograph documenting the history of paired comparison methods do we begin to get a sense for the profundity of Zermelo's work.[1]

Zermelo's paper is concerned mostly with estimating the relative strength of chess players in imbalanced designs, that is, tournaments in which each player does not compete against every other the same number of times. To address the problem, Zermelo introduces a probability model for game outcomes as a function of the players' unknown strengths. The model assumes that chess games result in only wins and losses.[2] Letting A_r and A_s be competitors r and s, the model is given by

$$\mathrm{P}(A_r \text{ defeats } A_s) = \frac{u_r}{u_r + u_s} \tag{1}$$

where u_r and u_s are the unknown playing strengths that are to be estimated. Perhaps unfairly to Zermelo, the model in (1) is now commonly called the Bradley-Terry model, based on the

[1]To add to the confusion in recognizing his work, both Good (1955) and David (1988) cite Zermelo's paper as 1929 rather than 1928.

[2]Zermelo does not explicitly consider a draw (tie) as an outcome in his model, but in footnote 3 he implies that he can reframe the optimization problem by doubling the number of wins and losses, and then counting draws as one win and one loss. A similar approach was advocated by Glickman (1999). It is worth pointing out that direct application of Zermelo's approach will result in overly optimistic standard error estimates of the playing strengths, as the sample size has doubled.

thorough treatment by Bradley and Terry in their 1952 paper. Ebbinghaus and Peckhaus (2007) in fact suggest referring to (1) as the Zermelo-Bradley-Terry model. There is no known evidence that this model for chess strength has been published prior to Zermelo's paper. These days, it is more convenient and common to work with a reparameterized version of this model. Setting $v_r = \ln u_r$ for player A_r, the model can be rewritten as

$$\mathrm{P}(A_r \text{ defeats } A_s) = \frac{\exp(v_r)}{\exp(u_r) + \exp(u_s)} = \frac{1}{1 + \exp(-(v_r - v_s))} \tag{2}$$

which is the standard logistic cumulative distribution function evaluated at $v_r - v_s$. In fact, letting $\boldsymbol{x} = (x_1, \ldots, x_n)$ be a vector of length n (the number of players in the tournament) with $x_r = 1$, $x_s = -1$, the remaining elements of $\boldsymbol{x}$ set to 0, and $\boldsymbol{v} = (v_1, \ldots, v_n)$, the model can be rewritten as

$$\text{logit } \mathrm{P}(A_r \text{ defeats } A_s) = \boldsymbol{x}'\boldsymbol{v} \tag{3}$$

where $\boldsymbol{x}'\boldsymbol{v} = \sum_{r=1}^{n} x_r v_r$,[3] and logit $p = \ln\left(\frac{p}{1-p}\right)$. Equation (3) is simply a logistic regression model with a linear predictor containing unknown coefficients $\boldsymbol{v}$, a member of the class of generalized linear models (McCullagh and Nelder, (1989)). Understanding this connection provides greater insights into the model properties and numerical methods in Zermelo's paper. It is interesting that Zermelo simply asserts the model in (1) as the *de facto* choice, as the contemporaneous model of Thurstone assumed a Gaussian cumulative distribution function rather than Zermelo's logistic distribution function.

The optimization problem on which Zermelo focuses involves maximizing the joint probability of the tournament game results as a function of the u_r. This is precisely the method of maximum likelihood estimation, which is arguably still the most common approach in the 21st Century to fitting probability models. It is likely that Zermelo's familiarity with

[3]This notation for the inner product is conventional in statistics literature.

statistical mechanics and its reliance on the principle of maximum entropy (which is intricately connected to maximum likelihood estimation) that may have guided his choice. At the time of Zermelo's paper, maximum likelihood estimation was in its infancy, with the formal development appearing in a series of papers by R. A. Fisher between 1912 and 1925. While Zermelo's statement of the optimization problem does not rely on the details of Fisher's development, it is unclear whether Zermelo was aware of Fisher's foundational work.

The focus of the first half of Zermelo's paper is the decomposition of tournaments into disjoint collections of players to establish the optimization of the likelihood function (that is, Zermelo's function $\Phi(u_1, \ldots, u_n)$) under the constraints that the u_r are non-negative and that $\sum_{r=1}^{n} u_r \leq 1$. The key point is that tournaments can be decomposed uniquely into a set of irreducible partial tournaments, or "prime" tournaments, with the following property: For a fixed prime tournament, every tournament player not in the prime tournament has either lost all games, won all games, or did not play against any player in the prime tournament. Zermelo then creates an ordering of the prime tournaments such that for the jth partial tournament C_j, every player in a partial tournament earlier in the ordering did not defeat a player in C_j, and every player later in the ordering did not lose to a player in C_j. The main theorem of the paper is that for tournaments with a decomposition that can be ordered in this way, maximizing the likelihood involves essentially maximizing factors in the likelihood product corresponding to the prime partial tournaments. The second part of the theorem connects optima across the prime partial tournaments: The ratio of optimized strength parameters between players A_j and A_k in two different ordered prime tournaments C_j and C_k where players in C_j are dominated by the players in C_k goes to 0 in the limit. It is important to note that the theorem only applies to tournament decompositions in which an ordering exists among all the prime partial tournaments, which is not always guaranteed.

As Zermelo acknowledges, tournaments can exist in which prime partial tournaments are incomparable; for example, when the players of C_j have not competed against any player in C_k, and (say) both C_j and C_k are dominated by every other prime partial tournament.

Zermelo's analysis, particularly with the comparison of strengths between prime tournaments, is more detailed than that of later authors. Ford (1957), an important article that re-derives a subset of Zermelo's results without ostensibly being aware of his paper (and, for that matter, of Bradley and Terry, (1952)), focuses on estimating strength in tournaments that are irreducible. Ford asserts the condition for the irreducibility of a tournament in a clean way: In every possible partition of players into two non-empty subsets, some player in the second set has defeated at least one player in the first set. Note that this definition of irreducibility coincides with Zermelo's for partial tournaments containing at least two players. The paper then demonstrates that this condition is sufficient for a unique optimum of Zermelo's likelihood function. By contrast to Ford's paper, which focuses only on irreducible tournaments, Zermelo considers more general tournaments that can be decomposed into irreducible prime partial tournaments, possibly with only one player in the partial tournament. In practice, however, optimization only makes sense when analyzing each irreducible prime tournament separately.

The second half of Zermelo's paper concentrates on obtaining the solution to the optimization problem. He first proves that, for a balanced design (in which each player competes against every other the same number of times) in an irreducible tournament, the total score for each player is in the same rank order as the optimized playing strengths. Ford (1957) independently derives this result, and Bühlmann and Huber (1963) demonstrate that in fact the model assumed by Zermelo is the only linear paired comparison model (that is, linear in the $v_r = \ln u_r$) for which ranking according to the total score is equivalent to ranking

according to the maximum likelihood estimates. Zermelo then derives an iterative numerical algorithm to solve the optimization problem for irreducible tournaments. The method is the same one described by Bradley and Terry (1952), though credit clearly is due Zermelo. The algorithm, however, has been shown to have slow convergence (Dykstra, (1956)), particularly with poorly chosen initial values. The modern treatment for optimizing Zermelo's model is to recognize it as a logistic regression and use Fisher scoring (see McCullagh and Nelder, 1989, p. 42) to perform the optimization, which is quite fast.

One interesting detail of Zermelo's development is how he addresses the non-identifiability of the player strengths. The probability model in (1) implies that if $(u_1^*, \ldots, u_n^*)$ is a solution to the optimization problem, then so is $(au_1^*, \ldots, au_n^*)$ for $a > 0$, provided that $\sum_{r=1}^n au_r \leq 1$. Thus, without an additional constraint, the solution is not unique. As Zermelo demonstrates, this indeterminacy does not affect the decomposition theorem, but the actual optimization requires a norm restriction on the u_r. Zermelo's extra constraint is to fix $\sum_{r=1}^n u_r$ at a constant (unity, in the development, and then 100 in his example). The modern way to add a constraint is to reparameterize the model as in (3), and to assume a constraint on $\sum_{r=1}^n v_r$; in other words, to fix the product of the u_r rather than the sum of the u_r. For example, it is conventional to assume that $\sum_{r=1}^n v_r = 0$ and then to estimate only $v_1, v_2, \ldots, v_{n-1}$, in which case the logistic regression model in (3) changes by the deletion of a term in the linear predictor. Specifically, recognizing that $v_n = -(v_1 + \cdots + v_{n-1})$, we have

$$\begin{aligned} \text{logit P}(A_r \text{ defeats } A_s) &= \boldsymbol{x}'\boldsymbol{v} = \left(\sum_{r=1}^{n-1} x_r v_r\right) + x_n v_n \\ &= \left(\sum_{r=1}^{n-1} x_r v_r\right) + x_n(-v_1 - \cdots - v_{n-1}) \\ &= \tilde{\boldsymbol{x}}'\tilde{\boldsymbol{v}} \end{aligned} \tag{4}$$

where $\tilde{\boldsymbol{v}} = (v_1, \ldots, v_{n-1})$, and $\tilde{\boldsymbol{x}} = (x_1 - x_n, x_2 - x_n, \ldots, x_{n-1} - x_n)$. This logistic regression

can then be fit via Fisher scoring to obtain unique maximum likelihood estimates, assuming the tournament is irreducible in the sense of Ford (1957).

Zermelo, at the end of his paper, applies the numerical algorithm to the analysis of playing strengths of participants in the New York Masters' Tournament of 1924, which was won by Emmanual Lasker (the world chess champion from 1894 through 1921). Lasker, who was also an accomplished mathematician, had written a letter to Zermelo in 1929 expressing appreciation of his work on measuring chess strength, pointing out that this paper was the first to use probability theory for this problem (see Ebbinghaus and Peckhaus (2007), p. 149). Lasker's performance in the 1924 tournament was arguably the last great chess performance of his career before he effectively retired from playing.

The development of models for estimating competitor playing strength has come a long way since Zermelo's paper, but Zermelo's work was remarkable in the level of detail in which it laid the foundation for paired comparison methods. Once Zermelo's approach caught on some 25 years later mostly through the work of Bradley and Terry and their contemporaries, a variety of interesting extensions were explored that are noteworthy. For example, models were developed for outcomes that include ties and other degrees of partial preferences. An extension to the model of Zermelo that has become standard is the incorporation of an (unknown) advantage for playing white in chess, which is more generally known as an "order" effect (in the sense that the probability of a preference between two objects may depend on the order in which they are presented). A detailed synopsis of recent developments in paired comparison modeling appears in the monograph by David (1988).

It is worth pointing out that recent approaches via Bayesian modeling (see, for example, Leonard (1977), and Glickman (1999)) in which a proper prior distribution is assumed for the player strengths avoid the difficulties connected with optimizing over Zermelo's prime

decomposition of tournaments. In the Bayesian framework, the results of a tournament simply revise the prior distribution of playing strengths to a posterior distribution, regardless of whether subsets of players have not competed against each other, or if any player has won or lost all of his games. These occurrences are problematic for Zermelo's (and in general the maximum likelihood estimation) approach. The Bayesian approach, however, still has important connections to the groundwork laid out by Zermelo in his analysis of the likelihood function, so much of Zermelo's development is unquestionably relevant today.

While seeming to have little influence immediately following its publication, Zermelo's paper has had a long-lasting impact. His work is now regularly cited in papers on paired comparison models, and his name is now immortalized in connection with rating competitors in games and sports. For example, statistician David Marcus, who constructed a rating system applicable for tournament table tennis (Marcus (2001)), paid tribute by developing Windows software called "Zermelo" that organizes and runs table tennis tournaments. Even American football rankings[4] have been attached to Zermelo. The resurgence of interest in Zermelo's paper on measuring playing strength over recent years is the appropriate recognition for an impressive piece of work.

References

Bradley, Ralph A. and Terry, Milton E. (1952). "The rank analysis of incomplete block designs. I. The method of paired comparisons." *Biometrika*, **39**, 324-45.

Bühlmann, Hans and Huber, Peter J. (1963). "Pairwise comparison and ranking in tournaments." *Annals of Statistics*, **34**, 501-10.

David, Herbert A. (1988). The method of paired comparisons, 2nd ed. Oxford University

[4]See the web site `http://www.ezfootballrankings.com`, where the "ez" in the domain name stands for "Ernst Zermelo."

Press: New York.

Dykstra, Otto (1956). "A note on the rank analysis of incomplete block designs – applications beyond the scope of existing tables." *Biometrics*, **12**, 301-6.

Ebbinghaus, Heinz-Dieter, and Peckhaus, Volker (2007). Ernst Zermelo: An approach to his life and work. Springer-Verlag: Berlin.

Elo, Arpad E. (1978). The rating of chessplayers, past and present. Arco Publishing: New York.

Ford, Lester R. Jr. (1957). "Solution of a ranking problem from binary comparisons." *American Mathematical Monthly*, **64(8)**, 28-33.

Glickman, Mark E. (1999). "Parameter estimation in large dynamic paired comparison experiments." *Applied Statistics*, **48**, 377-94.

Good, Irving J. (1955). "On the marking of chess-players." *Mathematical Gazette*, **39**, 292-6.

Leonard, Thomas (1977). "An alternative Bayesian approach to the Bradley-Terry model for paired comparisons." *Biometrics*, **33**, 121-32.

Marcus, David J. (2001). "New table-tennis rating system." *The Statistician*, **50**, 191-208.

McCullagh, Peter and Nelder, John A. (1989). Generalized Linear Models. Chapman and Hall: London

Mosteller, Frederick (1951a) "Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations." *Psychometrika*, **16**, 3-9.

Mosteller, Frederick (1951b) "Remarks on the method of paired comparisons: II. The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed." *Psychometrika*, **16**, 203-6.

Mosteller, Frederick (1951c) "Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed." *Psychometrika*, **16**, 207-18.

Thurstone, Louis L. (1927a). "A law of comparative judgment." *Psychological Review*, **34**, 273–86.

Thurstone, Louis L. (1927b). "Psychophysical analysis." *American Journal of Psychology*, **38**, 368–89.

Thurstone, Louis L. (1927c). "The method of paired comparisons for social values." *Journal of Abnormal and Social Psychology*, **21**, 384-400.