Intrinsic Chess Ratings

Kenneth W. Regan *
University at Buffalo

Guy McC. Haworth
University of Reading, UK

February 12, 2011

Abstract

This paper develops and tests formulas for representing playing strength at chess by the quality of moves played, rather than the results of games. Intrinsic quality is estimated via evaluations given by computer chess programs run to high depth, ideally whose playing strength is sufficiently far ahead of the best human players as to be a "relatively omniscient" guide. Several formulas, each having intrinsic skill parameters s for "sensitivity" and c for "competence," are argued theoretically and tested by regression on large sets of tournament games played by humans of varying strength as measured by the internationally standard Elo rating system. This establishes a correspondence between Elo rating and the parameters. A smooth correspondence is shown between statistical results and the century points on the Elo scale, and ratings are shown to have stayed quite constant over time (i.e., little or no "inflation"). By modeling human players of various strengths, the model also enables distributional prediction to detect cheating by getting computer advice during games. The theory and empirical results are in principle transferable to other rational-choice settings in which the alternatives have well-defined utilities, but bounded information and complexity constrain the perception of the utilitiy values.

1 Introduction

Player strength ratings in chess and other games of strategy are based on the results of games, and are subject to both luck when opponents blunder and drift in the player pool. This paper aims to rate players intrinsically by the quality of their decisions, as refereed by computer programs run to sufficient depth. We aim to settle controversial questions of import to chess policy and enjoyment, and then extend the answers to other decision-making activities:

1. Has there been "inflation"—or deflation—in the chess Elo rating system over the past forty years?

2. Does a faster time control markedly reduce the quality of play?

*Dept. of CSE, 201 Bell Hall, University at Buffalo Buffalo, NY 14260-2000; (716) 645-3180 x114, fax 645-3464; regan@buffalo.edu

3. Was Old Master X stronger than modern master Y?

4. Can game scores from tournaments where high results by a player are suspected to result from fraud reveal the extent to which "luck"—or collusion—played a role?

5. Can we objectively support assertions of the kind: this player is only an average master in the openings and middlegames, but plays at super-grandmaster strength in endgames?

The most-cited predecessor study, [Guid and Bratko2006], aimed only to compare the chess world champions, used a relatively low depth (12 ply) of a program Crafty [Hyatt2011] generally considered below the elite, and most importantly for our purposes, evaluated only the played move and Crafty's preferred move, if different. The departure point for our work is that to model probabilistic move-choice to needed accuracy and assess skill at chess, it is necessary to evaluate all of the relevant available moves. At a game turn with ℓ-many legal moves, we can list them $m_0, m_1, \ldots, m_{\ell-1}$ in nonincreasing order of their (composite) evaluations $e(m_0), e(m_1), \ldots, e(m_{\ell-1})$ by a (jury of) computer program(s), and use these as a measure of the moves' utilities. (When the convention of stating scores always from White's point of view is used, for Black moves the evaluations can be negated.) Our work uses the commercial chess program Rybka 3 [Rajlich and Kaufman], which was rated the best program on the CCRL rating list [CCR2010] between its release in August 2007 and the release of Rybka 4 in May 2010.

Guven the evaluations, for each i, $0 \leq i < \ell$, define $\delta_i = e(m_0) - e(m_i)$. A vector

$$\Delta = (\delta_0 = 0, \delta_1, \ldots, \delta_{N-1})$$

is called the *spread* of the top N moves. Often we cap N at a number such as 20 or 50, and if $\ell < N$, we can pad out to length N with values of "$+\infty$" if needed.

The only game-specific information used by our methods is the spread, the actual move played, and the overall evaluation of the position. This minimal information is common to other games of strategy, and our work aspires to handle any decision-making application with bounded rationality in which *utilities* (taken already to include risk/reward tradeoffs) must be converted into *probabilities*.

2 Background and Previous Literature

The Elo rating system (see [Elo2011]) computes an expected score E based on the differences $r_P - r_O$ between the rating of a player P and the ratings of various opponents O, and adjusts r_P according to the difference between the actual score and E. Although different particular formulas for E and (especially) the change to r_P are used by the World Chess Federation (FIDE) and various national federations, they are generally set up so that $r_P - r_O = 200$ means that P must score approximately 75% in games versus O in order to avoid losing rating points. (Under the logistic-curve model used by the United States Chess Federation, USCF, this is close to 76%.)

Since only rating differences matter there is no absolute meaning to the numbers produced, but considerable effort has been expended by FIDE and national federations to maintain the strength levels indicated by particular numbers stable over time. For four decades all World Champions and their closest contenders have held ratings in the neighborhood of 2800, while 2650 is the most commonly cited threshold for the unofficial "super-grandmaster" status, 2500 is typical of Grandmasters, 2400 of International Masters, 2300 of FIDE Masters (indeed, awardees of the the last three titles must have current ratings above those floors), while 2200 is often designated "master" by national federations. The USCF then uses 2000 as the threshold for "Expert," 1800 for "Class A," officially down to 1000 for "Class E" which has also been referred to as the strength of "bright beginners." Hence we refer to 200 points as a "class unit."

By fitting to these itemized skill levels, our paper continues work on *reference fallible agents* in the game context [Reibman and Ballard1983, Korf1987, Korf1988, Haworth2003, Haworth and Andrist2004, Andrist and Haworth2005]. The aim going beyond these papers, and beyond the results reported in this preliminary work, is to fit probabilities and confidence intervals for move section by thus-calibrated agents. The present work establishes that a reference-fallible model is supported by data taken on a far larger scale than previous studies, and using a more-ramified model than [DiFatta, Haworth, and Regan2009, DiFatta, Haworth, and Regan2010] which are based on Bayesian inference.

3 Basic Model

Our key abstract concept is the probability $p_i = p_i(s, c, \ldots)$ of a player P of skill level corresponding to parameters $s, c, \ldots$ choosing move m_i at a given turn t. We wish to fit p_i as a function of the spread Δ_t for that turn. We make three basic modeling assumptions:

(a) A player's choices at different game turns are independent, even for successive game turns.

(b) There is a relation between the probability p_i of selecting the i-th move and the probability p_0 of choosing an optimal move that depends simply and mainly on the value δ_i alone.

(c) For players at all skill levels, the relation has the form $r(p_i, p_0) = g(\delta_i)$, where r is a ratio, g is a *continuous, decreasing* function of δ_i, and g depends only on the parameters $s, c, \ldots$ for the skill level.

Assumption (3) is *de rigeur* for the whole enterprise of regarding p_i as a function of information for the given turn t alone. It also enables us to multiply probabilities from different turns, and add probabilities to project (for instance) the total number of matches to a computer's first choices over a sequence of moves. We justify it further below.

Note that (3) does not assert that p_i depends on δ_i alone. The quality of alternative moves must factor into the probability of selecting move m_i somehow. Our assumption (3) asserts what seems to be the simplest and weakest dependence on alternatives,

however, saying all their effect is bundled into p_0. We assert no other dependence on δ_j for $j \neq i$. In (3) we allow a smaller dependence on the overall evaluation e, but only to down-weight or filter out cases where e is extreme—i.e. for poor moves or when one side is clearly winning—as detailed in Section 5.

The function r is said to define the *model*, and the functions $g = g_{s,c,..}(\delta)$ are curves used to fit the model. The models and curves are normalized so that $g(\delta_0) = g(0) = 1$. Plausible models include:

1. *"Shares":* r is p_i/p_0, so $p_i = p_0 \cdot g(\delta_i)$.
2. *"Powers":* r is $\log(1/p_0)/\log(1/p_i)$, so $p_i = p_0^{1/g(\delta_i)}$.
3. $r(p_i, p_0) = \frac{p_i \log(1/p_0)}{p_0 \log(1/p_i)} = g(\delta_i)$.

Another way of describing Model (1) is that the curve value $s_i = g_{s,c,..}(\delta_i)$ is the "share" of move m_i, and its probability p_i is the ratio of s_i to the sum of the shares, $S = \sum_{i=0}^{\ell-1} s_i$. Hence the name "Shares." In Model (2), the curve represents an "exponential decay" of probability in going from an optimal move to an inferior one.

Justification of the Assumptions

Assumption (3) is intuitively false when a sequence of move choices constitutes a single *plan*. For example, if White plays 20.Nh1-g3 and the best square for the Knight is d4, then White is humanly likely to follow with 21.Ng3-e2 and 22.Ne2-d4. However, this is one of several places where our modeling requires only that assumptions be "usually approximately true," and where the degree of compliance is ascertainable from the data.

Assumption (3) tacitly assumes that r itself is monotone in p_i, so that holding p_0 constant, the probabilities $p_i(s, c, \ldots)$ are also decreasing functions of δ_i, for all fixed skill levels. How reasonable is this? It is easy to devise positions with an attractive but inferior move that most human players would choose, where the best move may—or may not—be found by our computer jury at the reference depth, thus falsifying (3) for that position. However, the "Shares" model extends naturally to assert

$$p_i = p_0 \cdot \sum_d w_d g(\delta_{d,i}),$$

and the others can be extended similarly. Then weaker players can be recognized as those with higher values of w_d for lower d, and their fitted ensembles of curves may be non-monotone for fixed higher values of d. Although we do not fit such w_d parameters in this work, and use evaluations only at the given fixed depth 13 ply of Rybka 3, there is reason to believe from our results that we have gotten a representative "slice" of this spectrum for a fairly wide variety of players. Thus we have two defenses to objections on (3) that stay within our basic modeling assumptions, and we can use goodness-of-fit data to justify both it and (3).

A third kind of objection is that important chess-specific information has been left out, such as received evaluations of chess openings (especially gambits whose compensation is not recognized quickly enough by engines), player styles, tournament situations, time-pressure during specific games, and more. To this we answer that the

very force of our model is its *chess-neutrality*, and that a statistical yardstick can be objective only when it does not adapt to any player style or situation internally. The most we allow are ideas of putting external weights on the significance of the models' results according to situation, such as weighting games at rapid time controls less.

To justify this answer, and turn away objections similar to those leveled at the Guid-Bratko study, we need internal means of measuring the effectiveness of our simple model, and a good positive result from such measurements. First we describe the kinds of curves to fit.

4 Possible Move-Choice Formulas

We regard the two parameters already called "s" and "c" above as organic, with a common meaning across models and curves. The s parameter represents a conversion from the hundredths-of-a-pawn units of δ_i into the dimensionless quantity δ_i/s used in all of our curves. The smaller s, the greater the ratio when δ_i is moderate, thus lowering the projected probability of the i-th move. Hence s governs a player's ability to discriminate moderately inferior moves, so we call it the *sensitivity*. We use the symbol s, and divide rather than multiply, because it equals or scales with the standard deviation of several of the curves when the curves themselves are viewed as distributions.

The parameter c appears as an exponent of δ/s, directly or with some intervening terms. Intuitively it governs how often a player avoids moves in the range the player discriminates as inferior, and avoids poor moves overall. Hence we regard it as a notion of *competence*.

Additional parameters to fit may come from fitting linear combinations of curves, and will come as weights over ply-depths in the full model with "swing." However we regard s and c, together with the model and curve family, as determining the *shape* of skill at chess. Salient families of curves g, all normalized to make $g(0) = 1$, that we have considered are:

- Inverse-exponential curves:

$$\textit{invexp}(\delta) = e^{-(\delta/s)^c}.$$

- Inverse-polynomial curves:

$$\begin{aligned} \textit{ipa}(\delta) &= \frac{1}{1 + (\delta/s)^c}, \qquad \text{or} \\ \textit{ipb}(\delta) &= \frac{1}{(1 + \delta/s)^c}. \end{aligned}$$

- Logistic function-related curves:

$$\begin{aligned} \textit{secha}(\delta) &= \frac{2}{(e^{(\delta/s)^c} + e^{-(\delta/s)^c})} \qquad \text{or} \\ \textit{sechb}(\delta) &= \frac{4}{(e^{(\delta/s)^c} + 2 + e^{-(\delta/s)^c})}. \end{aligned}$$

All of these curves were found to give similar results, largely owing to the way they approximate each other near the origin. We standardized our results on the inverse-exponential family.

5 Data Methodology and Experiments

Two large sets of data were taken, the former acting as a control for the latter. The former comprises approximately 150,000 games of chess, including every match for the world championship, every qualifying match for the championship, the top round-robin tournaments from London 1851 onward, large selections from the Chess Olympiads, every USSR/Russia and USA Championship, and a host of other kinds of chess competition including rapid, blitz, correspondence, and computer play. These were scripted by running Rybka 3 in so-called *Single-PV* mode, whereby it computes a full evaluation only for the preferred move, to reported depth 13-ply in the Arena chess GUI [Blume2010], which logs the evaluation of each move and other information automatically to a text file.

These runs re-created a somewhat simpler form of the Guid-Bratko experiment [Guid and Bratko2006]. Similar to there, each game was begun on move 9 since earlier moves are often repeated and are considered part of common opening theory. Moves where one side was already established as being more than 3 pawns ahead according to Rybka 3 were discarded; [Guid and Bratko2006] used a similar 2-pawn cutoff for Crafty run to 12 ply, but 3 pawns was felt better for Rybka 3 owing to its deeper search. The percentage of moves on which the player and Rybka 3 agreed, called the move-match percentage (mm), were tallied for each player in each event. The difference in evaluation in cases where the player chose a sub-optimal move according to Rybka 3, summed and averaged over all moves, comprised our version of the Guid-Bratko "Average Difference" (ad) statistic.

When the average difference was plotted against the overall evaluation e of a given position for the player to move, it was found that the former scaled markedly up with $|e|$. The effect was so pronounced that in many kinds of chess events, the same players when judged ahead by only 0.5 pawns showed a 60-70% higher average-difference than when the position was judged very close to dead-even. Since it seemed strange to infer that players in such cases were performing 60–70% worse, it was decided to institute a scale correction. This was done by integrating a line differential $\ell(x)$. Thus if $e = +0.50$ for a White move turn, but the move m_i played was judged a 1.25 pawn error, the $\delta(m_i)$ value was recorded as the integral of $\ell(x)$ from -0.75 to $+0.50$. It was found that a simple proportional scale correction equalized the global ad statistics in relation to e fairly well.

The main data set comprised games in which both players were within 10 Elo rating points of one of the "milepost" values: 2700, 2600, 2500, ..., run under standard time controls in individual-player round-robin or small-Swiss tournaments. Team events and large-Swiss (meaning more than six times as many players as rounds) tournaments were excluded. Games were taken from three time periods: 2006–2009, 1991–1994, and 1976–1979. These games were evaluated to depth 13 ply in 50-PV mode; since most positions have fewer than 50 legal moves, and all but a trace with more legal

moves have fewer than 50 remotely sensible ones, this guaranteed full evaluation and consideration of alternatives to the played move and/or the preferred move. Each set had at least 5,000 moves that were not subject to the 3-pawn cutoff or discarding for "repetitions," while the largest set had just over 25,000 moves, so the relative sizes were reasonably consistent. In each case we ran *all* available games meeting the description, from two major commercial game collections marketed by ChessBase Gmbh. and OpeningMaster.com, so as to avoid any bias in selection.

For 1976–1979 it was possible to find relaible and large-enough game sets only for the 2300 through 2600 mileposts, while 1991–1994 made it possible to add 2700 and 2200. Since the World Chess Federation has expanded its FIDE rating system to players below 2200 in recent years, for 2006–2009 it was possible to find enough games down to 1600. The ranges around mileposts were expanded from ± 10 to ± 15 or ± 20 for some of the lower sets.

For each non-discarded move of each game, evaluation and spread data was processed from the Arena analysis logs into the following format:

```
Move played: 12.Ng5
Engine move: 12.e5
Eval end-13: +0.63
Delta = (0.00,0.04,*0.12,0.12,0.39,...)

Move played: 12...h6
Engine move: 12...h6
Eval end-13: +0.56
Delta = (*0.00,1.05,1.06,1.20,1.85,...)

Move played: 13.Nxf7+
Engine move: 13.Nh3
Eval end-13: +0.57
Delta = (0.00,0.10,2.40,2.42,2.46)*

Move played: 13...Rxf7
Engine move: 13...Rxf7
Eval end-13: -2.20
Delta = (*0.00,9.30,9.42,---,---,---,..)
```

In this hypothetical example, the White move 13.Nxf7+ (giving check) might be classed as a "blunder," and the $\star$ after the closing) signifies that it wasn't among the top ten moves. At the next move the $---$ marks indicated that Black had only 3 legal replies; one could pad "Delta" with a large value such as 5.00 or 10.00 instead. Since Black's 13...Rxf7 capturing the Knight is a forced move (else Black loses Queen for Knight as hinted by $\delta_2 = 9.30$ and $\delta_3 = 9.42$ for the two other legal moves), it has a near-total share, and the turn itself has entropy near zero, a non-critical move. Since the evaluation after 13...Rxf7 shows an imbalance above 2.00, this last datum might be discarded anyway.

6 Fitting Methodology

If all spread tuples were the same $\Delta = (0, \delta_2, \ldots, \delta_N)$, or if we had a large-enough set of nearly-equal tuples to form a histogram, fitting the results to a curve would be relatively simple. Let $f_1, f_2, \ldots, f_N, f_{N+1}$ be the observed frequencies of which indexed move in the spread was played, with f_{N+1} standing for "move played not in the top N" and hopefully negligibly small. Then given a curve $g_{s,c}(\delta)$ and distance measure μ, such as $\mu(x, y) = |x - y|^2$ for least-squares, we could compute the fit score $S_{q,c} =$

$$\mu(f_1, 1/S) + \mu(f_2, g_{s,c}(\delta_2)/S) + \cdots + \mu(f_N, g_{s,c}(\delta_N)/S),$$

where $S = 1 + g_{s,c}(\delta_2) + \cdots + g_{s,c}(\delta_N)$. In the case of equal spreads this yields the same best-fit s, c as maximum-likelihood estimation.

With heterogeneous spreads, however, the estimation is trickier. Maximum-likelihood estimation can still be applied to obtain the best $s, c, \ldots$ for a given curve or hybrid g, but this alone does not judge whether g has the right "shape" across a range of δ.

To this end we devised a "percentiling" method. Given a curve $g_{s,c}(\delta)$, let q additionally stand for a percentile. For each point (q, s, c) in a fine-enough grid, say stepping q by 0.05 from 0 to 1, s by 0.02 from 0 to 0.70, and c by 0.20 from 1 to 5, we iterate through each spread tuple $\Delta_t = (0, \delta_2, \ldots, \delta_N)$. For each i, $1 \le i \le N$, compute the probabilities $p_i = g_{s,c}(\delta_i)/S_t$, where $S_t = \sum_i g_{s,c}(\delta_i)$. Let i_t be the index of the played move. Define $p^- = \sum_{j=1}^{i_t - 1} p_j$ and $p^+ = p^- + p_{i_t}$, giving the endpoints of the predicted probability interval of the played move. Then:

- If $p^+ \le q$, call the tuple "up."
- If $p^- \ge q$, call the tuple "down."
- If $p^- < q < p^+$, so that the prediction for the played move straddles the q-th percentile, count the tuple as being $|q - p^-|/p_{i_t}$ up, and $|q - p^+|/p_{i_t}$ down.

Finally define $R^q_{s,c}$ to be the percentage of "up" tuples. Given a distance measure μ as above, the score now becomes

$$S_{s,c} = \sum_q \mu(R^q_{s,c}, q).$$

A low score indicates a good fit across a range of percentiles for the curve $g_{s,c}(\delta)$.

Note that for a spread Δ with one clearly-indicated best move, say with $\delta_2 = 1.50$, the predicted range for most s, c will span beyond the 90th percentile. Suppose the best move is played, as predicted. For $q = 0.30$, say, the tuple will count as (roughly) one-third up, two-thirds down. It may seem counter-intuitive for a result that confirms a prediction to give an overall "down" score, but the prediction that is actually tested by our method is not the individual move but the overall frequency of hits above/below a given percentile. Nor is it necessary to estimate the proportion of the cumulative distribution of $g_{s,c}(\delta)$ to the left and right of 0.30 in the spanned range—the straight one-third/two-thirds division is correct. In effect we have converted from the "δ scale" to the percentile scale, with the effect that instead of plotting data points for a horizontal δ-axis and fitting $g_{s,c}(\delta)$, we fit the derived percentile curve(s) instead.

7 Results

A two-parameter model such as ours is trickier to fit, especially when the parameters trade strongly off against each other. A statistical analyzing program written in C++ carried out the two-dimensional minimization needed to implement the above fitting method. It was found that while s varied from 0.07 to 0.16 and beyond, the c value stayed between 0.430 and 0.545. Accordingly we did a simple linear fit of the c values for 2006–2009, getting intervals coincidentally highly close to 0.007, and then used these to compute "normalized" fitted s-values for each rating milepost. The results, the predicted and actual move-match and average-difference statistics, and a measure of the quality of the fit, are shown in the following table.

2006–2009

Elo	s	c	c_{fit}	s_{fit}	mm_p/mm_a	ad_p/ad_a	Q_{fit}
2700	.078	.503	.513	.080	56.2/56.3	.056/.056	.009
2600	.092	.523	.506	.089	55.0/54.2	.063/.064	.041
2500	.092	.491	.499	.093	53.7/53.1	.067/.071	.028
2400	.098	.483	.492	.100	52.3/51.8	.072/.074	.016
2300	.108	.475	.485	.111	51.1/50.3	.084/.088	.044
2200	.123	.490	.478	.120	49.4/48.3	.089/.092	.084
2100	.134	.486	.471	.130	48.2/47.7	.099/.102	.034
2000	.139	.454	.464	.143	46.9/46.1	.110/.115	.065
1900	.159	.474	.457	.153	46.5/45.0	.119/.125	.166
1800	.146	.442	.450	.149	46.4/45.4	.117/.122	.084
1700	.153	.439	.443	.155	45.5/44.5	.123/.131	.065
1600	.165	.431	.436	.168	44.0/42.9	.133/.137	.129

1991–1994

2700	.079	.487	.513	.084	55.2/54.9	.058/.060	.043
2600	.092	.533	.506	.087	55.3/54.6	.064/.063	.042
2500	.098	.500	.499	.092	54.3/53.8	.068/.069	.013
2400	.101	.484	.492	.103	52.3/51.9	.077/.079	.016
2300	.116	.480	.485	.117	51.0/50.3	.088/.091	.031
2200	.122	.477	.478	.122	49.7/48.7	.092/.098	.058

1976–1979

2600	.094	.543	.506	.087	53.8/53.0	.062/.061	.038
2500	.094	.512	.499	.091	53.2/52.5	.067/.068	.032
2400	.099	.479	.492	.103	52.3/51.7	.076/.079	.020
2300	.121	.502	.485	.116	50.9/50.0	.088/.090	.070

Our first major conclusion is that there is a fairly smooth relationship between the players' Elo rating and the intrinsic quality of the moves as measured by the chess program and the fitting. Moreover, the final s_{fit} values obtained are nearly the same for the corresponding entries of all three time periods. Since a lower s indicates higher skill, we conclude that *there has been little or no "inflation" in ratings over time—if anything there has been deflation!* This runs highly counter to conventional wisdom, but is predicted by population models on which rating systems have been based [Elo2011].

8 Conclusions and Further Directions

We have demonstrated that quality of move choice can be ascertained based on intrinsic measures rather than the results of games. We have essentially fitted only the first-moments of the respective skill levels. The next main task is to obtain projected confidence intervals from the percentile-fitting method, based on its similarity to Bernoulli trials, and then test whether they are accurately populated by our large amounts of data. The result will be a model of move choice that is capable of testing allegations of whether players—of a given skill level—have been agreeing with computer evaluations more than chance would warrant. Another use for our model will be a better simulation of human players of these skill levels, especially being faithful to their observed tendency to make markedly inferior choices ("blunders"). Finally, insofar as our methods involve almost no information specific to chess, they should be transferable to other domains.

References

[Andrist and Haworth2005] Andrist, R., and Haworth, G. 2005. Deeper model endgame analysis. *Theoretical Computer Science* 349:158–167.

[Blume2010] Blume, M. 2010. Arena Chess GUI.

[CCR2010] 2010. CCRL rating lists.

[DiFatta, Haworth, and Regan2009] DiFatta, G.; Haworth, G.; and Regan, K. 2009. Skill rating by bayesian inference. In *Proceedings, 2009 IEEE Symposium on Computational Intelligence and Data Mining (CIDM'09), Nashville, TN, March 30–April 2, 2009*, 89–94.

[DiFatta, Haworth, and Regan2010] DiFatta, G.; Haworth, G.; and Regan, K. 2010. Performance and prediction: Bayesian modelling of fallible choice in chess. In *Proceedings, 12th ICGA Conference on Advances in Computer Games, Pamplona, Spain, May 11–13, 2009*, volume 6048 of *Lecture Notes in Computer Science*, 99–110. Springer-Verlag.

[Elo2011] 2011. Elo rating system - Wikipedia.

[Guid and Bratko2006] Guid, M., and Bratko, I. 2006. Computer analysis of world chess champions. *ICGA Journal* 29(2):65–73.

[Haworth and Andrist2004] Haworth, G., and Andrist, R. 2004. Model endgame analysis. In *Advances in Computer Games*, volume 135, 65–79. Kluwer Academic Publishers, Norwell MA.

[Haworth2003] Haworth, G. 2003. Reference fallible endgame play. *ICGA Journal* 26:81–91.

[Hyatt2011] Hyatt, R. 2011. Crafty chess engine.

[Korf1987] Korf, R. 1987. Real-time single-agent search: first results. In *Proceedings, 6th International Joint Conf. on Artificial Intelligence*.

[Korf1988] Korf, R. 1988. Real-time single-agent search: new results. In *Proceedings, 7th International Joint Conf. on Artificial Intelligence*.

[Rajlich and Kaufman] Rajlich, V., and Kaufman, L. Rybka 3 chess engine. http://www.rybkachess.com.

[Reibman and Ballard1983] Reibman, A., and Ballard, B. 1983. Non-minimax strategies for use against fallible opponents. In *proceedings, Third National Conference on Artificial Intelligence (AAAI-83)*.