# Quantifying Depth and Complexity of Thinking and Knowledge

Tamal T. Biswas and Kenneth W. Regan
*Department of CSE*
*University at Buffalo*
*Amherst, NY 14260 USA*
{*tamaltan, regan*}*@buffalo.edu*

Keywords: Decision making, depth of search, chess, item difficulty, judging of learning agents, knowledge representation

Abstract: Qualitative approaches to cognitive rigor and depth and complexity are broadly represented by Webb's Depth of Knowledge and Bloom's Taxonomy. Quantitative approaches have been relatively scant, and some have been based on ancillary measures such as the thinking time expended to answer test items. In competitive chess and other games amenable to incremental search and expert evaluation of options, we show how depth and complexity can be quantified naturally. We synthesize our depth and complexity metrics for chess into measures of difficulty and discrimination, and analyze thousands of games played by humans and computers by these metrics. We show the extent to which human players of various skill levels evince shallow versus deep thinking, and how they cope with 'difficult' versus 'easy' move decisions. The goal is to transfer these measures and results to application areas such as multiple-choice testing that enjoy a close correspondence in form and item values to the problem of finding good moves in chess positions.

## 1 INTRODUCTION

Difficulty, complexity, depth, and discrimination are important and related concepts in cognitive areas such as test design, but have been elusive to quantify. Qualitative approaches are legion: Bloom's taxonomy (Bloom, 1956; Krathwohl et al., 1973; Anderson and Krathwol, 2001), Webb's Depth of Knowledge Guide (Webb, 1997), Bransford et al.'s studies of learning (Bransford et al., 2000; Donovan and Bransford, 2005). Quantitative approaches have mainly either inferred values from performance data, such as results on large-scale tests (Morris et al., 2006; Hotiu, 2006), or have measured ancillary quantities, such as deliberation time in decision field theory (Busemeyer and Townsend, 1993) or estimations of risk (Tversky and Kahneman, 1992).

Our *position* is to approach these concepts by starting in a domain where they can be clearly formulated, cleanly quantified, and analyzed with large data. Then we aim to transfer the formulations, results, and conclusions to domains of wider interest. Our home domain is *competitive chess*, in which the *items* are thousands to millions of positions from recorded games between human players in various kinds of high-level tournaments. Work to date (Chabris and Hearst, 2003; Haworth, 2003; Guid and Bratko, 2006, 2011; Regan and Haworth, 2011) has established solid relationships between quality measures arising from direct analysis of players' move decisions and standard skill assessment metrics in chess, mainly grades of mastery and the Elo rating system. Some prior work (Chabris and Hearst, 2003; Moxley et al., 2012) has extended the correspondence to time available and/or taken for (move) decisions, but this is still short of isolating depth or difficulty as factors.

Our aims are helped by similarities between the tasks of finding an optimal move (or at least a good move) in a chess position and finding the best answer to a multiple-choice question (or at least a good answer in case there are partial credits). There are also mathematical correspondences between the Elo rating system (Elo, 1978; Glickman, 1999) and metrics in Rasch modeling (Rasch, 1961; Andersen, 1973; Andrich, 1978; Masters, 1982; Andrich, 1988; Linacre, 2006; Ostini and Nering, 2006), item-response theory (Baker, 2001; Morris et al., 2006; Thorpe and Favia, 2012), and other parts of psychometrics.

Elo ratings $r_P$ of players $P$ maintain a logistic-curve relationship between the expected score of $P$ over an opponent $Q$ and the rating difference $r_P - r_Q$. A difference of 200 points gives roughly 75% expectation, and this has produced a scale on which 2200 is recognized as "master," the highest few players are over 2800, and many computer *chess engines* are rated well over 3000 even on inexpensive hardware.

The engines can hence act as an objective and authoritative “answer key” for chess positions.

Essentially all engines give values in standard units of *centipawns* and use *iteratively deepened* search. That is, beginning with $d = 1$ (or some other floor value) they search to a basic depth of *d plies* (meaning moves by White or Black, also called half-moves), give values $v_{i,d}$ to each legal move $m_i$ at that depth $d$, and then deepen the search to depth $d+1$. This incremental search can be capped at some fixed maximum depth $D$. Based on depth-to-strength estimates by Ferreira Ferreira (2013) for the Houdini 1.5a engine and matches run by us between it and versions 2.3.1 and 3 of the Stockfish engine used for the results reported here, we estimate depth 19 of the latter (in so-called *Multi-PV* analysis mode) at 2650 $\pm$50.

Taking care to begin with an empty hash table for each position in each game, we use Stockfish’s values $v_{i,d}$ for $1 \leq d \leq D = 19$ to quantify our key concepts. Our measures are weighted so that values of poor moves have little effect, so we could effectively bound the number of legal moves at $\ell = 50$. We consider moves ordered so that $v_{1,D} \geq v_{2,D} \geq \cdots \geq v_{\ell,D}$ at the highest depth, but of course the highest value $v_d^*$ for $d < D$ might equal $v_{i,d}$ where $i > 1$. We actually work in terms of the differences $v_d^* - v_{i,d}$, and in order to reflect that differences matter less when one side has a large advantage, we further *scale* them by defining

$$\delta_{i,d} = \int_{x=v_{i,d}}^{x=v_d^*} \frac{1}{1+a|x|} dx.$$

Here the constant $a$ might be engine-dependent but we fix $a = 1$ since we used only two closely-related Stockfish versions. Cases where $v_d^*$ is positive but $v_{i,d}$ is negative (meaning that move $m_i$ is an error leading from advantage to disadvantage) are handled by doing the integral in two pieces. All $\delta_{i,d}$ values are nonnegative, and are 0 for the optimal move at each depth and any other moves of equal value. The key idea of *swing* is exemplified by these two cases:

- A move $m_i$ *swings up* if $v_{i,d} < v_{j,d}$ for some other moves $m_j$ at low depths $d$, but $v_{i,d} \geq v_{j,d}$ for (almost) all other $m_j$ for depths $d$ at or near the maximum analyzed depth $D$.
- The move *swings down*—and intuitively is a “trap” to avoid—if it has one of the highest values at low depths, but is markedly inferior to the best move $m_1$ at the highest depth: $v_{i,D} \ll v_{1,D} = v_D^*$.

It is expected in the former case that $v_{i,D} > v_{i,d}$ for lower depths $d$, and in the latter that $v_{i,D} \ll v_{i,d}$, so that a swinging move changes its absolute value, but it is its value *relative to other moves* that is primarily assessed.

# 2 METRICS AND RATINGS

At each depth $d$, the chess program produces an ordered list $L_d$ of moves and their values. Comparing these lists $L_d$ for different $d$ involves standard problems in preference and voting theory, with the twist that high values from poor moves have diminished weight. We speak of *rating aggregation* rather than *rank aggregation* because the values of each move, not just the ordinal ranks, are important.

We postulate that *swing* should be a signed quantity in centipawn units that pertains to an individual move option, while *complexity* should be nonnegative and dimensionless and pertain to a position overall. Swing should reflect a bulk comparison of $L_d$ for low $d$ versus high $d$, while complexity can be based on how $L_d$ changes to $L_{d+1}$ in each round of search. Thus for complexity we may employ some divergence measure between ordered sequences $X = (x_i)$, $Y = (y_i)$ and sum it up over all $d$. Whereas common voting and preference applications give equal weight to all choices, we wish to minimize the effects of appreciably sub-optimal moves.

Any anti-symmetric difference function $\mu(x_i, x_j)$ gives rise to the *generalized Kendall tau coefficient*

$$\tau_{X,Y} = \frac{\sum_{i,j} \mu(x_i,x_j)\mu(y_i,y_j)}{||\mu_X|| \cdot ||\mu_Y||}, \qquad (1)$$

where $||\mu_X|| = \sqrt{\sum_{i,j} \mu(x_i,x_j)^2}$ and $||\mu_Y||$ is defined similarly. Then always $-1 \leq \tau_{X,Y} \leq +1$, with $+1$ achieved when $Y = X$ and $-1$ when $Y = -X$. If $\mu$ is homogeneous, so that $\mu(cx_i, cx_j) = c'\mu(x_i, x_j)$ where $c'$ depends only on $c$, then $\tau_{X,Y}$ becomes scale-invariant in either argument: $\tau_{X,cY} = \tau_{cX,Y} = \tau_{X,Y}$.

The usual difference function $\mu(x_i, x_j) = x_i - x_j$ is linear, and also invariant under adding a fixed quantity to each value. It is not, however, invariant under augmenting the lists with irrelevant alternatives having low ratings. We swap these properties by employing

$$\mu(x_i, x_j) = \frac{x_i - x_j}{x_i^2 + x_j^2}$$

instead. When either $x_i$ or $x_j$ is large, say of order $K$ representing a poor move, then $\mu(x_i, x_j)$ will have order at most $1/K$. Assuming that the same move is poor in $Y$, the augmentation will add terms of order only $1/K^2$ to the numerator and denominator of (1), yielding little change. This naturally confines attention to *reasonable* moves at any juncture. We define the *complexity* $\kappa(\pi)$ of a position $\pi$, for $d$ ranging from the minimum available depth $d_0$ to $D-1$, by:

$$\kappa(\pi) = 1 - \frac{1}{D-1} \sum_{d=1}^{D-1} \tau_{L_d, L_{d+1}}.$$

Notice that high agreement ($\tau$ always near 1) flips around to give complexity $\kappa$ near 0. The definition of complexity might be modified by weighting higher depths differently from lower depths.

To define the *swing* of a move $m_i$ we use a simple sum of scaled differences in value between depth $d$ and the highest depth $D$, rather than average or otherwise weight them over $d$:

$$sw(m_i) = \sum_{d=1}^{D} (\delta_{i,d} - \delta_{i,D}).$$

This is a signed quantity—if positive it means that the value of move $m_i$ "swings up", while negative means it "swings down"—in the manner of falling into a trap. The overall "swinginess" of a position $\pi$, however, is a non-negative quantity. It is convenient first to define it between any two depths $d$ and $e$:

$$s_{d,e}(\pi) = \sum_{i=1}^{\ell} |\delta_{i,d} - \delta_{i,e}|.$$

For overall swing it is expedient to dampen the effect of moves for which $\delta_{i,d}$ is large. Unlike the case with Kendall tau, we want to dampen a difference $|\delta_{i,d} - \delta_{i,e}|$ only if *both* values are large. We also wish to divide by a dimensionless quantity, in order to preserve the centipawn units of swing. Hence we postulate a scaling factor $c$ that might depend on the chess program, and divide by an exponential function of the harmonic mean of the deltas divided by $c$:

$$\nu(\delta, \delta') = \exp\left(\frac{-2\delta\delta'}{c(\delta + \delta')}\right).$$

Since this paper uses only one chess program, we again take $c = 1$. Thus we define the damped overall swing between depths $d$ and $e$ by:

$$s^*_{d,e}(\pi) = \sum_{i=1}^{\ell} \nu(\delta_{i,d}, \delta_{i,e})|\delta_{i,d} - \delta_{i,e}|.$$

Then the swing *at* depth $d$ is given by $s^*_{d,d+1}(\pi)$, while the aggregate swing to the highest depth is defined by

$$S(\pi) = \sum_{d=1}^{D-1} s^*_{d,D}.$$

We employ weighted versions of this to define our key concepts. We desire the measure of *difficulty* to be in units of *depth* rather than centipawns. Our idea is that a position is deeper, hence more difficult, if most of the swing occurs at higher depths. It is OK to multiply it by the complexity since that is dimensionless.

Accordingly, we first define the *relative depth* $\rho$ to be the depth below which half of the swing has occurred. For this we add up the swing from each depth to the next, rather than the swing relative to the highest depth. With respect to nonnegative weights $w(d)$ summing to 1, define

$$\Sigma(\pi) = \sum_{d=1}^{D-1} w(d) s^*_{d,d+1}(\pi).$$

We used $w(d) = d$ normalized by $\sum_{d=1}^{D-1} d$. Then, letting $\Sigma_e(\pi)$ be the sum up to $e$ rather than $D-1$, define

$$\rho(\pi) = \max\{e : \Sigma_e(\pi) \geq \frac{1}{2}\Sigma(\pi)\} - \psi,$$

where the adjustment $\psi$ term for the indicated $e$ is

$$\psi = \frac{\Sigma_e(\pi) - \frac{1}{2}\Sigma(\pi)}{w(e-1)s^*_{e-1,e}(\pi)}.$$

Finally, we stipulate that the *analyzed difficulty* of a position $\pi$ is given by

$$Diff(\pi) = \kappa(\pi) \cdot \rho(\pi).$$

For calculating the discrimination we use the relative depth of the position. We evaluate the mean $\alpha_l$ and standard deviation $\sigma_l$ of $s^*_{d,D}$ values where $d \in (1, e-1)$ $(e = \rho(\pi))$ and mean $\alpha_r$ and standard deviation $\sigma_r$ of $s^*_{d,D}$ values where $d \in (e, D-1)$.

The discrimination parameter $\Psi$ of the position $\pi$ can then be evaluated as:

$$\Psi(\pi) = (\alpha_l - \alpha_r)(\frac{\alpha_l}{\sigma_l} + \frac{\alpha_r}{\sigma_r})\frac{\sum_{i,j} w_{i,j}(s^*_{i,D} - s^*_{j,D})^2}{\sum_{i,j} w_{i,j}}.$$

The weights $w_{i,j} = 1/(j-i)$ where $i \in [l]$ and $j \in [r]$ ensures more emphasize to the depths near the difficulty of the position while calculating discrimination.

Our first of two main datasets comprised all recorded games in standard round-robin[1] tournaments in 2006–2009 between players each within 10 Elo of a "milepost" value. The mileposts used were Elo 2200, 2300, 2400, 2500, 2600, and 2700. The second comprised all 900 games of the 2013 World Blitz (WB) Championship, which was held in Khanty-Mansiysk, Russia, and distinguished by giving an accurate record of the moves of every game. This form of blitz, 3 minutes per game plus an increment of 2 seconds per move, is comparable to the historical "5-minute" form of blitz, and gives markedly less time than the minimum 90 minutes plus 30 seconds per move of the "milepost" games. Our idea was to test whether the blitz games were played at an identifiably lower level of depth. The average rating of the 60 WB players was 2611.

[1] "Small Swiss" events with up to 64 players over 9 rounds were also included.

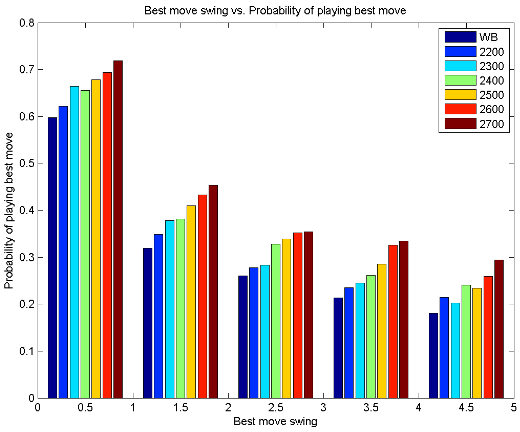


Figure 1: Frequency of playing engine moves with different swing values

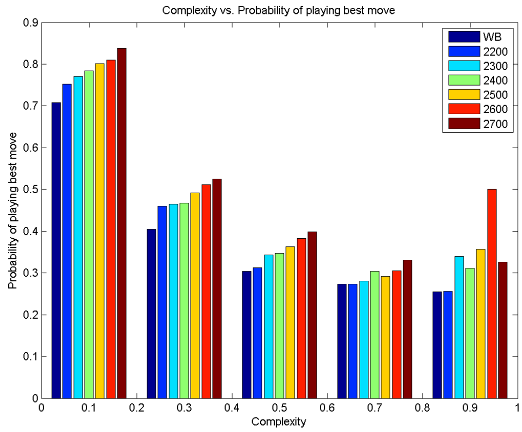


Figure 2: Frequency of playing engine moves for position with various complexity

# 3 RESULTS

Our results show that the raw factor of *swing* makes a large impact on the ability of players at all levels to find the optimal move $m_1$ identified (at the highest depth) by the engine, and that this carries forward to our more-refined difficulty and discrimination measures. The WB games seemed to function as if they were a rating level below 2200, most often in the range 1800 to 2100.

Table 1 gives the total moves (TM) and times with the engine's move played (EMP) for each of five intervals of swing values $sw(m_1)$, and Figure 1 graphs the frequencies of $m_1$ being played in each case. The plot clearly indicates that high-swing moves are "tricky" for players to find—the players more often chose inferior moves. The phenomenon is consistent with players of any Elo ratings, where higher rated players are slightly less tricked by the swing values. This feature is more prominent in the blitz tournament. Quick decision making often leads to pick inferior moves, or where the virtue of the engine move was not obvious at lower depths.

In our implementation, we rank the possible moves at any particular position based on the order provided by the chess engines. Often the first move listed by the engine shows less swing, and make it attractive choice for the players from the beginning. Earlier studies show that players often chose the first move listed by the engine 58% of the time whereas the second move is chosen only 42% of the time. Table 2 shows that in fact the first listed move often has much lower swing with comparison to the other tied moves. This is true for players across any ability level.

Figure 2 represents the probability of playing the best move for positions of various complexity. The probability gets monotonically decreased. The random noise seen at positions with higher complexity is due to insufficient number of samples(see Table 3).

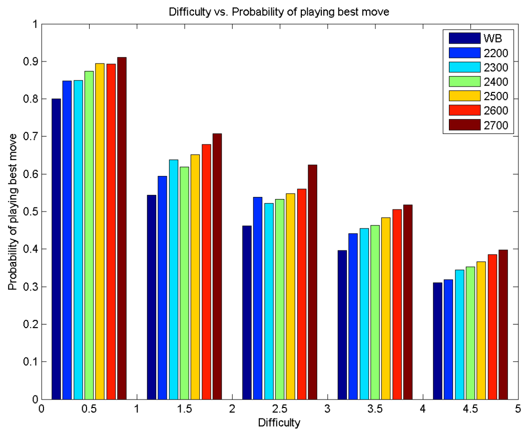


Figure 3: Frequency of playing engine moves for position with various difficulty

Figure 3 demonstrates difficulty and best-move probability for various positions. The figure clearly shows that players of all calibers could find the best move when the position is easy, but less than 50% of the time when the difficulty lies between 4 and 5. Table 4 shows the distribution of data across various difficulty levels. Figure 4 shows a similar but lesser effect for our measure of discrimination.

# 4 CONCLUSION AND PROSPECTS

We have defined quantitative measures for qualitative concepts of depth, difficulty, complexity, and discrimination. The definitions are within a specific

Table 1: Best move and number of total moves played various swing

| Level | Swing < 1 | | $1 \leq$ Swing < 2 | | $2 \leq$ Swing < 3 | | $3 \leq$ Swing < 4 | | $4 \leq$ Swing $\leq 5$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | #EMP | TM | #EMP | TM | #EMP | TM | #EMP | TM | #EMP | TM |
| WB | 22,785 | 38,150 | 3,523 | 11,025 | 1,542 | 5,920 | 697 | 3,267 | 825 | 4,566 |
| 2200 | 4,954 | 7,967 | 812 | 2,331 | 364 | 1,308 | 184 | 781 | 220 | 1,025 |
| 2300 | 8,081 | 12,164 | 1,413 | 3,740 | 575 | 2,030 | 303 | 1,235 | 350 | 1,731 |
| 2400 | 8,878 | 13,536 | 1,575 | 4,127 | 754 | 2,296 | 340 | 1,301 | 493 | 2,049 |
| 2500 | 7,203 | 10,620 | 1,374 | 3,351 | 691 | 2,040 | 332 | 1,162 | 383 | 1,635 |
| 2600 | 3,252 | 4,689 | 701 | 1,619 | 323 | 918 | 165 | 507 | 213 | 823 |
| 2700 | 2,823 | 3,927 | 596 | 1,315 | 261 | 737 | 144 | 430 | 208 | 708 |

Table 2: Swing for Tied moves

| Level | Any # of Tied Moves | | | 2 Tied Moves | | | 3 Tied Moves | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | First | Second | #Moves | First | Second | #Moves | First | Second | Third | #Moves |
| WB | 1.168 | 1.903 | 12,163 | 1.383 | 2.073 | 6,580 | 1.246 | 2.051 | 2.646 | 1,661 |
| 2200 | 1.263 | 1.912 | 2,447 | 1.430 | 2.056 | 1,391 | 1.141 | 2.005 | 2.265 | 329 |
| 2300 | 1.310 | 2.063 | 3,731 | 1.530 | 2.233 | 2,120 | 1.333 | 2.061 | 2.782 | 516 |
| 2400 | 1.330 | 1.964 | 4,270 | 1.555 | 2.178 | 2,514 | 1.211 | 1.841 | 2.449 | 569 |
| 2500 | 1.380 | 2.154 | 3,309 | 1.538 | 2.302 | 1,981 | 1.318 | 2.127 | 2.782 | 476 |
| 2600 | 1.408 | 2.216 | 1,607 | 1.533 | 2.349 | 974 | 1.244 | 2.338 | 3.083 | 242 |
| 2700 | 1.558 | 2.106 | 1,255 | 1.737 | 2.292 | 755 | 1.411 | 2.011 | 2.614 | 210 |
| Overall | 1.273 | 1.989 | 28,782 | 1.477 | 2.163 | 16,315 | 1.260 | 2.043 | 2.645 | 4,003 |

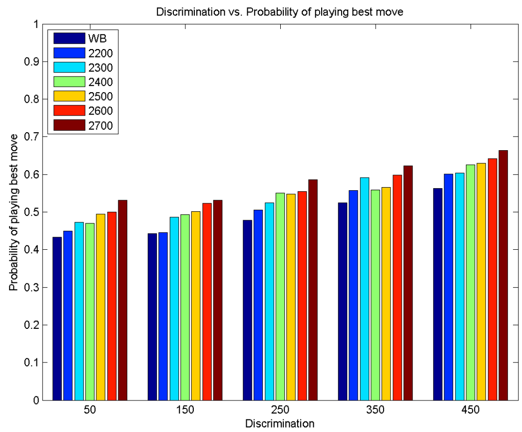


Figure 4: Frequency of playing engine moves for position with various discrimination

model of decision making at chess, but use no feature of chess apart from utility values of decision options, and are framed via mathematical tools that work across application areas. For the first three, we have shown a strong response effect on performance, though we have not distinguished the measures from each other. The effect shows across skill levels and persists when restricting to controlled cases such as moves of equal highest-depth value.

# REFERENCES

Andersen, E. (1973). Conditional inference for multiple-choice questionnaires. *Brit. J. Math. Stat. Psych.*, 26:31–44.

Anderson, L. and Krathwol, D. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A revision of Blooms taxonomy of educational objectives: complete edition*. Longman, New York.

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika*, 43:561–573.

Andrich, D. (1988). *Rasch Models for Measurement*. Sage Publications, Beverly Hills, California.

Baker, F. B. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation.

Bloom, B. (1956). *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain*. David McKay Co., New York.

Bransford, J. D., Brown, A., and Cocking, R., editors (2000). *How People Learn: expanded edition*. The National Academies Press, Washington, D.C.

Busemeyer, J. R. and Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, 100(3):432.

Chabris, C. and Hearst, E. (2003). Visualization, pattern recognition, and forward search: Effects of playing speed

Table 3: Number of times best move played vs. number of total moves at positions of various complexity

| | $0 \leq$ Cpx. $< 0.2$ | | $0.2 \leq$ Cpx. $< 0.4$ | | $0.4 \leq$ Cpx. $< 0.6$ | | $0.6 \leq$ Cpx. $< 0.8$ | | $0.8 \leq$Cpx. $\leq 1$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| Level | #EMP | TM | #EMP | TM | #EMP | TM | #EMP | TM | #EMP | TM |
| WB | 14851 | 20983 | 7804 | 19270 | 5216 | 17155 | 1395 | 5104 | 106 | 416 |
| 2200 | 3083 | 4098 | 1863 | 4051 | 1216 | 3894 | 340 | 1244 | 32 | 125 |
| 2300 | 5227 | 6779 | 2933 | 6308 | 1984 | 5790 | 519 | 1849 | 59 | 174 |
| 2400 | 5672 | 7236 | 3421 | 7315 | 2282 | 6574 | 619 | 2036 | 46 | 148 |
| 2500 | 4590 | 5731 | 2916 | 5923 | 1946 | 5369 | 474 | 1625 | 57 | 160 |
| 2600 | 2061 | 2547 | 1384 | 2706 | 954 | 2496 | 232 | 761 | 23 | 46 |
| 2700 | 1840 | 2197 | 1146 | 2182 | 830 | 2084 | 200 | 605 | 16 | 49 |

Table 4: Best move and number of total moves for positions with various difficulty

| | Diff. $< 1$ | | $1 \leq$ Diff. $< 2$ | | $2 \leq$ Diff. $< 3$ | | $3 \leq$ Diff. $< 4$ | | $4 \leq$ Diff. $\leq 5$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| Level | #EMP | TM | #EMP | TM | #EMP | TM | #EMP | TM | #EMP | TM |
| WB | 10438 | 13038 | 3790 | 6979 | 3462 | 7497 | 3189 | 8049 | 8493 | 27365 |
| 2200 | 2178 | 2569 | 754 | 1269 | 853 | 1585 | 741 | 1680 | 2008 | 6309 |
| 2300 | 3666 | 4318 | 1333 | 2089 | 1278 | 2450 | 1211 | 2665 | 3234 | 9378 |
| 2400 | 3925 | 4492 | 1481 | 2396 | 1463 | 2743 | 1452 | 3133 | 3719 | 10545 |
| 2500 | 3140 | 3512 | 1205 | 1848 | 1262 | 2304 | 1189 | 2459 | 3187 | 8685 |
| 2600 | 1392 | 1558 | 572 | 843 | 558 | 996 | 605 | 1198 | 1527 | 3961 |
| 2700 | 1183 | 1300 | 538 | 761 | 534 | 855 | 460 | 888 | 1317 | 3313 |


and sight of the position on grandmaster chess errors. *Cognitive Science*, 27:637–648.

Donovan, M. S. and Bransford, J. D. (2005). *How Students Learn*. The National Academies Press, Washington, D.C.

Elo, A. (1978). *The Rating of Chessplayers, Past and Present*. Arco Pub., New York.

Ferreira, D. (2013). The impact of search depth on chess playing strength. *ICGA Journal*, 36(2):67–80.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, 48:377–394.

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

Guid, M. and Bratko, I. (2011). Using heuristic-search based engines for estimating human skill at chess. *ICGA Journal*, 34(2):71–81.

Haworth, G. (2003). Reference fallible endgame play. *ICGA Journal*, 26:81–91.

Hotiu, A. (2006). The relationship between item difficulty and discrimination indices in multiple-choice tests in a physical science course. M.Sc. thesis.

Krathwohl, D., Bloom, B., and Bertram, B. (1973). *Taxonomy of Educational Objectives, the Classification of Educational Goals. Handbook II: Affective Domain*. David McKay Co., New York.

Linacre, J. M. (2006). Rasch analysis of rank-ordered data. *Journal of Applied Measurement*, 7(1).

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47:149–174.

Morris, G. A., Branum-Martin, L., Harshman, N., Baker, S. D., Mazur, E., Dutta, S., Mzoughi, T., and McCauley, V. (2006). Testing the test: Item response curves and test quality. *American Journal of Physics*, 74(5):449–453.

Moxley, J. H., Ericsson, K. A., Charness, N., and Krampe, R. T. (2012). The role of intuition and deliberative thinking in experts' superior tactical decision-making. *Cognition*, 124(1):72 – 78.

Ostini, R. and Nering, M. (2006). *Polytomous Item Response Theory Models*. Sage Publications, Thousand Oaks, California.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings, Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 321–334. University of California Press.

Regan, K. and Haworth, G. (2011). Intrinsic chess ratings. In *Proceedings of AAAI 2011, San Francisco*.

Thorpe, G. L. and Favia, A. (2012). Data analysis using item response theory methodology: An introduction to selected programs and applications. *Psychology Faculty Scholarship*, page 20.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323.

Webb, N. (1997). *Criteria for Alignment of Expectations and Assessments on Mathematics and Science Education*. Monograph No. 6. CCSSO, Washington, DC.