A Comparative Review of Skill Assessment: Performance, Prediction and Profiling

Guy Haworth[1], Tamal Biswas and Ken Regan[2]

[1] The University of Reading, Reading, UK
[2] The University at Buffalo (SUNY), Amherst, NY USA
guy.haworth@bnc.oxon.org
{tamaltan,regan}@buffalo.edu

Abstract. The assessment of chess players is both an increasingly attractive opportunity and an unfortunate necessity. The chess community needs to limit potential reputational damage by inhibiting cheating and unjustified accusations of cheating: there has been a recent rise in both. A number of counter-intuitive discoveries have been made by benchmarking the intrinsic merit of players' moves: these call for further investigation. Is Capablanca actually, objectively the most accurate World Champion? Has ELO rating inflation not taken place? Stimulated by FIDE/ACP, we revisit the fundamentals of the subject to advance a framework suitable for improved standards of computational experiment and more precise results. Other games and domains look to chess as demonstrator of good practice, including the rating of professionals making high-value decisions under pressure, personnel evaluation by Multichoice Assessment and the organization of crowd-sourcing in citizen science projects. The '3P' themes of performance, prediction and profiling pervade all these domains.

Keywords: average difference · Bayesian inference · cheating · chess · equal-value matching · fallibility · false positive · move matching · skill assessment · Skilloscopy · standards· statistics

1 Introduction

This position paper is motivated by the recent proliferation of studies and analyses of chess players' skill. A more serious requirement is the need for both scientific rigour and clarity of explanation stemming from the rise in the number of alleged and proven accusations of computer-assisted cheating over the board. FIDE and the Association of Chess Professionals (ACP) are taking urgent action, involving the third author, to defend the reputation and commercial dimensions of the game [1].

The aim here is to introduce a framework for the discussion and analysis of assessment methods that evaluate the 'intrinsic merit' of the subject's decisions. It clearly separates the goals of metrics for *Performance* and *Prediction*, and quantifies the use of *Profiling*. *Performance* refers to quantitative measures that can be correlated with ELO ratings. *Prediction* means anticipating the distribution of some 100-300 move choices that a player might make in a tournament. Both recognise that strong

adfa, p. 1, 2011.

players will find better moves but anti-cheating tests have to do better than merely identify someone who is 'playing too well'. This is required if analyses of excellent play are to have more statistical significance and more reliably trigger further investigation.

Profiling refers to the use of information about a player's behavior and achievements prior to the period of time relevant to that player's later assessment. Bayesian inference as used by Haworth et al [2-9] is the natural vehicle for this as it combines an expression of *prior belief* with the modification of that belief in the light of subsequent evidence. Players may also be profiled from a 'cold start' if 'know nothing' priors are adopted. The new FIDE Online Arena with its AceGuard cheating-detection system is notable for its user-profiling which is not done by other statistical methods [10]. Arguably it should have no role in measuring performance but its ability to predict is something that this paper's recommendations will help to study.

1.1 The Chess Engine as Benchmarking Player

All assessment research approaches now use computer-generated move analysis as the benchmark. The reality is that computers are now better and quicker decision makers than humans in the vast majority of chess positions. The top-ranked chess engines are now rated some 300-400 ELO points better than the top human players and no prominent equal-terms match has been played since December 2006. Chess engines deliver their verdicts on the available moves at each nominal ply-depth of their forward-search process. The current top two engines, KOMODO 8, STOCKFISH 6 and others do so via the standard UCI communication protocol [11]. Moves may 'swing up' or 'swing down' as they gain or lose apparent merit at increased depths. This illustrates both the engines' fallibility induced by the finiteness of their vision and the theoretically proven trend to greater precision at higher depths.

However, it is clear that chess engines' centipawn evaluations of positions are not definitive but merely best estimates: engines are fallible agents, even if the best are less fallible than human players. They tend to depart from 0.00 as a decisive result becomes more obvious with increased depth of search, and they vary from engine to engine on the same position [12, 13]. Only in the endgame zone where endgame tables (EGTs) have been computed does an infallible benchmark exist [3-6].

Chess engines operate in one of two modes, *Single-PV* and *Multi-PV*. They focus on what they deem to be the best move in *Single-PV* mode, testing its value rather than also testing 'inferior moves' for comparison. *Multi-PV* guarantees full evaluation to search-depth *sd* of up to *k* 'best' moves as determined by the last round of search. Setting *k* to 50 essentially gives an *sd*-evaluation of every reasonable legal move and many others. *Multi-PV* working requires more time than *Single-PV* working but is required if the full move-context of a move is to be considered.

Alternative-move evaluations are the only input to the benchmarking processes below. No information about the clock-regime, move number, material, relative ELO difference or clock-times is considered, although there is evidence of the error-inducing *zeitnot* effect as players approach move 40 under classic conditions [14].

1.2 A Framework of Requirements for Assessment Methods

The approach here is to re-address the fundamental question 'What are the objectives of player assessment?', to identify the requirements in more detail, and then consider a portfolio of assessment methods in the context of those requirements. All this may be done under our headings of *performance* and *prediction*. As will be seen, all the methods reviewed below measure past performance, some less informed than others. Only a subset of the methods are suitable for reliable prediction.

A more detailed list of requirements is as follows:

1. identifying the current or overall performance of players on some scale,
2. identifying their 'intrinsic skill', i.e., on the basis of their moves, not results,
3. doing so in terms of a 'most likely' scale-point and with *c%* confidence limits,
4. identifying performance across the years, including the pre-ELO years,
5. ranking players relative to each other on the basis of their intrinsic skill,
6. understanding the stability of methods when subject to small input changes,
7. comparing methods as to the uncertainty budgets associated with their verdicts,
8. using 'robust' methods which are least sensitive to small input changes,
9. improving assessment methods where the opportunity to do so arises,
10. identifying suspected cheating with a suitably high degree of confidence,
11. identifying suspected cheating in real-time in order to trigger further action [15],
12. quashing 'false positive' accusations of over the board cheating,
13. discouraging players from cheating with evidence of good anti-cheating methods,
14. discouraging unfounded accusations of cheating in a similar way, and
15. estimating the probability that player *P* will play move *m* in a given position.

The following classification of information illustrates the range of algorithmic sophistication available to a notional punter betting on future moves:

A) Played move and engine-optimal move(s) as evaluated at greatest search-depth,
B) Values of all (reasonable) legal moves as evaluated at greatest search-depth,
C) Values of all (reasonable) move at all available depths of search,
D) Information about the chess position other than move values, and
E) Information as to a player's tendencies prior to the time of the moves assessed.

Category 'C' highlights the fact that this information has been available but has only recently been recognised as valuable [16]. We argue that 'C' is where the separation of performance and prediction should be focused. The demerit of a superficially attractive move which 'traps' the opponent only becomes visible at the greater depths of search. Heading 'D' includes considerations of pawn structure, attack formations and whether a move is advancing or retreating. Observations on how such factors influence move-choice are made by kibitzers but have not yet been captured by computer algorithm. Time management might be taken into account. Heading 'E' involves considering players' past game and how they might help predict future moves.

2 Useful notation

The following notation is used in subsequent sections:

- *AP*, the assessed player
- *BP*, the benchmark player against which *AP* is assessed
- *CP*, the cheating player, not only cheating but deceiving an assessment method
- *HP*, the honest player who is not cheating
- *RP*, a *Reference Player*, i.e., a stochastic agent with defined choice-behaviour
- $p_i \equiv$ position i, often implicitly understood to be one of a sequence of positions
- $\{m_j, v_{j,d}\} \equiv$ moves from a position, resulting in values (at depth d) $v_{j,d}$: $v_{j,1} \geq v_{j,2}$ etc.
- $ac_i \equiv$ the *apparent competence* of player *AP* after moving from position p_i

3 Survey of Assessment Methods

This section gives names to each of the methods known, and lists the methods' absolute and relative advantages ('+'), caveats ('±') and disadvantages ('–'). The first list applies to all methods.

+ chess engines are the only benchmarks which perform consistently across time,
+ the best chess engines are now thought to be better than humans at all tempi,
+ increasing engine ELO decreases move-choice suboptimality by the engine,
+ increasing search-depth increases engine ELO and decreases suboptimality,
+ 'cold start', defined-environment, single-thread running ensures reproducibility,
+ skill-scales may be calibrated in ELO terms using 'Reference ELO *e* players'
+ skill assessments on such calibrated scales lead to inferred ELO ratings,
± results from different benchmarking engines BP_i may be combined with care,
– *AP's* actual competence varies within games, tournaments and over the years,
– move-choices stem from a plan but are modelled as independent events,
– chess engines are not fully independent, tending to make the same mistakes,
– multithread processing, though attractive, introduces lack of reproducibility,
– there is a probability $p_m > 0$ that cheating player CP will not be detected,
– there is a probability $p_{fp} > 0$ that honest player HP will be accused of cheating.

The eight methods reviewed below are classified under three headings:

- 'Agreement': the observance of agreement between *AP* and *BP*,
- 'Average Difference': the recording of centipawn value 'lost' by *AP*, and
- 'Whole Context': the appreciation of *AP's* move in the full context of options.

3.1 Agreement between *AP* and *BP*

The methods here are 'MM: Move Matching' and its enhancement 'EV: Equal-value Matching' requiring a little more data as from Multi-PV mode.

MM: Move Matching. Observers of games commonly note whether the human player's choice move matches that of some 'kibitzer-engine' and compute a %-match *MM*. Specific merits (+) and demerits (-) of this method:

+ Engine- and human-moves are easily generated, communicated and compared,
+ the method applies to all moves, even those where engines see 'mate in *m*',
+ '*MM(AP)*=1.00' is a clear 'best possible performance' calibration point,
+ there is no need to scale centipawn values provided by engine *BP*,
– *MM(AP)* changes on forced moves when nothing is learned about *AP's* skill,
– different but equivalent/equi-optimal moves are not regarded as 'matches',
– some engines may randomly choose different equi-optimal moves to top their list,
– cheater *CH* can easily lower their *MM*(*CH*) at minimal cost,
– 'Canals on Mars syndrome': observers are attracted to high-*MM*(*AP*) coincidences,
– this method uses the least information of any method.

EV: Equal-value Matching. Disadvantages 2-3 of MM are addressed [17]. Equi-optimal moves are regarded as 'matches', requiring the engines to identify them all:

+ *EV* is not susceptible to random ordering by chess engine *BP* whereas *MM* is,
+ *EV* results are reproducible whereas *MM* results from a 'randomising *BP*' are not,
– *EV*, unlike *MM*, requires all equi-optimal moves to be communicated.

3.2 'Average Difference' methods

AD: Average Difference. Note that we chose *AD* not *AE* (Average Error) as the error may come from the benchmarking computer *BP* rather than from *AP* [18-20]:

+ The requisite information is again easily generated, communicated and used,
+ '*AD*(*AP*) = 0.00' is a clear 'best possible performance' calibration point,
+ *AD*, using more information than *MM* or *EV*, should give more robust ratings,
– *AD*(*AP*) changes on forced moves when nothing is learned about *AP's* skill,
– *AD*(*AP*) changes when AP has only equi-optimal moves to choose from,
– *AD* only uses information about the best apparent move and *AP's* choice,
– BP_1 and BP_2 may return different *AD*(*AP*), even when choosing the same moves,
– *AD* cannot be used where the engine gives 'mate in m' rather than an evaluation,
– *AD* does not scale 'differences' in consideration of the position's absolute value.

AG: Accumulated Gain. This method [21] varies from *AD*. *AP's* move is credited with the difference between *BP*'s evaluation of the position before and after the move.

+ *AG* uses only *BP*'s position evaluation at depth *d* before and after the played move,
± *AG* guarantees that the winner will be higher rated than the loser,
– *AG* conflates the 'horizon effect' with *AP's* performance,
– *AG* can give a positive score for a suboptimal move if *BP* sees a win more clearly,
– *AG* can penalize an optimal move by the loser as *BP* sees the win more clearly,
– *AG*, unlike *AD*, does not produce a clear mark (0.00) of perfect performance.

Had *AG* evaluated the position after *AP's* move at search-depth *d-1*, it would be close to *AD.* However, it moves *BP's* horizon on by one ply and therefore credits *AP* with *BP's* change of perception one ply later. It does not compare *AP's* and *BP*'s decisions at the same moment. The concept seems flawed and is not considered further here.

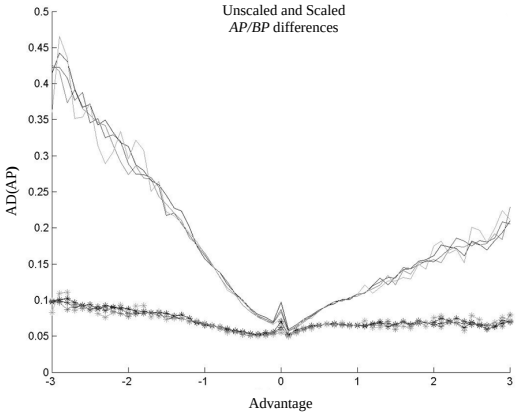

Fig. 1. Unscaled *AP*/*BP* differences, before and after scaling [14].

ASD: Average Scaled Difference. The last caveat on the *AD* method anticipates a key finding by Regan [14, 22] that average-difference correlates with the absolute value of the position, see Figure 1. This may be because (a) humans are only sensitive to the relative values of moves, (b) humans with an advantage tend to avoid the risk associated with the sharpest tactical plans, and/or (c) engines see the win more clearly when the position is relatively decisive already. The case for scaling seems clear.

If *pv* = |position value|, AP's 'difference *ad* relative to *BP's* choice is scaled to be *ad*/ln(1 + *pv*). Regan reports that he now prescales all differences in his industry-scale computations. The recommendation here is that all results produced by the *AD* method should be revisited and scaled in this way.

A detailed study of results from the EV and AD methods [17] also notes the danger of 'false positive' conclusions about suspected cheating over the board. It points to extreme ratings, which any corpus of results will have, which would at first sight be suspicious had they not been achieved before the availability of chess engines at grandmaster level. Table 1 highlights some games analysed with STOCKFISH 3.0.

Table 1. Achievements over the board which would be or are ' false positives' [17]

#	Year	White	Black	Res.	Book depth	Search Depth	Moves Anal.	CV	Comment
1	1857	Kennicott	Morphy	0-1	29	18	10	—/1.00	Morphy moves 15-24
2	1857	Schulten	Morphy	0-1	8	16	13	—/1.00	Morphy moves 5, 17
3	1866	Morphy	Maurian	1-0	12	18	12	1.00/—	Morphy moves 7, 18
4	1889	Weiss	Burille	1-0	13	20	26	1.00/—	Weiss moves 8-33
5	1965	Carames	Fedorovsky	½-½	18	18		0.85/0.82	Dead drawn, positions 62b-101w
6	1980	Browne	Timman	1-0	33	8	23	1.00/—	Browne moves 18-40
7	2009	Mamedyarov	Kurnosov	0-1	31	*var.*	6	—/—	too few moves; CV insignificant

3.3 'Whole context' analysis: deepest evaluations only

These methods potentially draw on the full context of a move-choice to assess the choice made by *AP*. They deploy a set $SBP \equiv \{BP(c_i)\}$ of stochastic benchmark players of defined competence c_i. As c_i increases, the expected value of BP_i*'s* chosen move increases if this is possible. For these methods:

+ a much fuller use of the move-context is being made,
+ 'apparent competence' does not change if nothing is learned from the move-choice,
+ these methods can easily calculate MM/EV and AD/ASD as byproducts,
– the method potentially requires all moves to be evaluated,
– the method uses the evaluation of the moves at the greatest depth only,
– the number *MultiPV* of 'best moves' considered is a computation parameter,
– the definition of $q_{j,i} \equiv \{\Pr[m=m_j \mid BP(c_i)]\}$ requires some domain-specific insight,
– the task of communicating statistical significance is greater than for other methods,
– the results of two *SR* computations cannot easily be combined.

SR: Statistical Regression. This method, deployed by Regan [14, 22, 23] identifies the BP_i which best fits the observed play: it is essentially frequentist. The probability of $BP(c_i)$ playing moves m_1-m_k is $p(c_i) \equiv \Pi q_{j,i}$ and c_i is found to maximize $p(c_i)$. The model also generates variances and hence provides *z-scores* for statistical tests employing the MM, EV, and AD/ASD measures.

– the results of two *SR* computations cannot easily be combined.

We report here that SR, carried out to FIDE/ACP guidelines, comes to a negative rather than a positive conclusion on all the games of Table 1, and on the aggregate of Morphy's moves. Given a distribution of MM/EV figures for players of Morphy's standard, the MM/EV figures' z-scores are less than the minimum mark of 2.75 stated [1] as needed to register statistical support for any 'positive' conclusion about cheating likelihood. The Browne-Timman and Mamedyarov-Kurnosov results are less than 0.50. The reason is that the whole-context analysis finds these and Weiss's and Morphy's games to be unusually forcing, so SR gives higher projections than the simpler MM/EV or AD/ASD analyses as [17] would expect. Thus, our category 'B' outclasses 'A' here for the purpose of prediction. This distinction is legislated in [1].

SK: Skilloscopy, the Bayesian approach. Classical probability asks how probable a future event is given a defined scenario. Bayesian analysis asks instead 'What is the probability of each of a set of scenarios given (a) a prior belief in the likelihood of those scenarios and (b) a set of observed events?' An important advantage is that his simple formula can be used iteratively as each new observation arrives.

Skilloscopy is the name given to the assessment of skill by Bayesian Inference [2-9]. It proceeds from initial inherited or presumed probabilities p_i that *AP* 'is' $BP(c_i)$: *AP's* initial presumed apparent competence *ac* is therefore $\Sigma_i p_i$. Given a move m_j and the probability $q_{j,i} \equiv \{\Pr[m=m_j \mid BP(c_i)]\}$, the $\{p_i\}$ are adjusted by the Bayesian formula

$$p_i' \propto q_{j,i} \times p_i$$

The $\{p_i'\}$ continue to represent how specifically *AP's* apparent competence on the c_i-scale is known: *AP's* apparent competence $ac = \Sigma_i p_i'$.

Skilloscopy was first conceived [3-5] in the context of that part of chess for which endgame tables (EGTs) hold perfect information. These EGTs provided infallible benchmark players BP_c so the above caveats about the fallibility of *BP* do not apply.

\+ *SK* can combine the results of two independent, compatible computations,
\+ *SK* may evaluate the moves in any order, chronologically or not,
– the choice of $\{BP(c_i)\}$ affects AP_j*'s* rating ac_j after a defined input of evidence,
– AP_j*'s* rating ac_j is meaningful only relative to other ratings ac_j.

3.4 'Whole context' analysis: evaluations at all depths

The most recent addition to the spectrum of assessment methods [16] is labelled '*SRA*' here, being *SR* but taking move-valuations from all depths of *BP's* search. It is clear that if a move takes and retains top ranking early in the search, it is more likely to be selected by *AP* than a move that emerges in the last iteration of the search. Therefore, to ignore shallower-depth evaluations is to ignore valuable information.

Similarly, one can study the way in which such indices as MM/EV and AD/ASD plateau out as search-depth increases. It appears that greater depths are required to get stable MM/EV/ASD ratings for better players. Figure 2 generated from the STOCKFISH v2.31 and v3 data [16] shows this for ASD and also corroborates the contention of Guid and Bratko [18-20] that even CRAFTY'S relatively shallow analysis of world champions suffices to rank them accurately if not to rate them accurately. The sixty players in the 2013 World Blitz championship (WB) had average rating 2611 but showed a competence lower than 2200 at classical time controls.

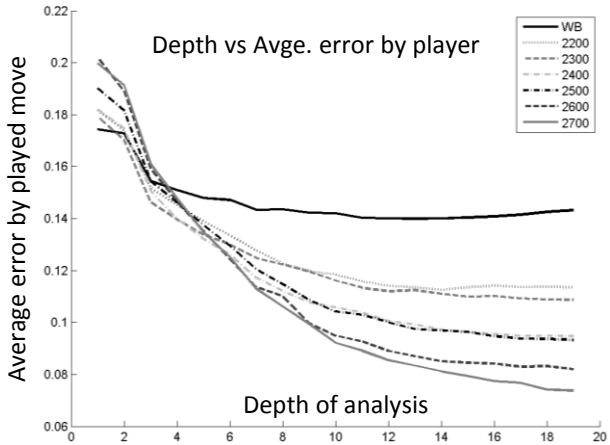

Fig. 2. 'Average Difference' statistics reaching a plateau as *BP's* search depth increases.

4 The Reference ELO Player

RP_e, a Reference Player with ELO *e* may be defined by analyzing the moves of a set of players with ELO e±δ, e.g., [2690, 2710]. This was done [7-9], in fact restricting chosen games to those between two such players.[1] The players' ratings in MM/EV, AD/ASD and SR/SK terms may be used to calibrate their respective scales.

Following such calibration, any set of move-choices may be given an Inferred Performance Rating, IPR. That IPR may be placed in the distribution of IPRs by nominally similar players and may be admired or investigated as appropriate.

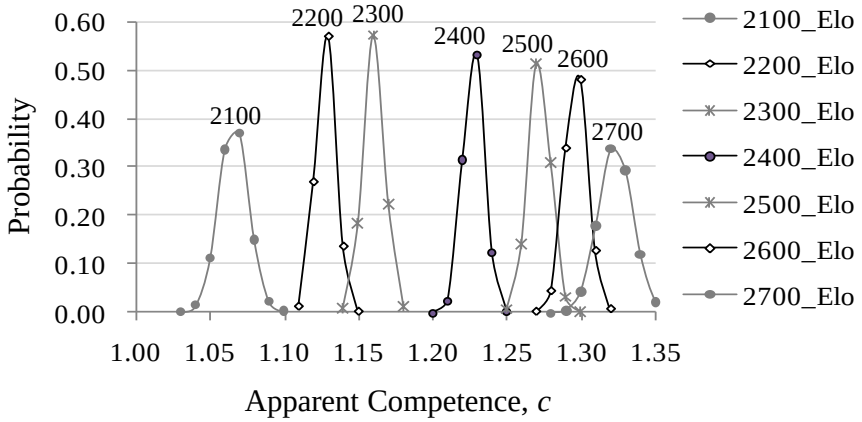

Fig. 3. The set of ELO *e* Reference Players used by Skilloscopy [7-9].

5 Standards for a Research Community

The statistical assessment of IPRs requires large amounts of relevant data. The large choice of chess engines, versions, search depths and other computational parameters does not help in combining results by different workers. There is a natural preference to use the best available engines to reduce position-evaluation inaccuracy, and the typical reign of the 'best engine' is usually short.[2]

However, greater interworking in the community may be assisted by:

- the 'separation' of move-analysis, skill-rating and inferred performance rating,
- computational experiments being done in a defined and reproducible way,
- a comprehensive data-model encompassing the computations' results, and
- a robust, accessible repository of results consonant with the data-model about:
  move analyses, skill-rating exercises and inferences of 'apparent ELO'.

[1] This probably increased the apparent competence *ac* of RP_e: draws exhibited higher *ac*.

[2] Over the last four years, the winners of the TCEC events [12] have been HOUDINI 1.5a, HOUDINI 3, KOMODO 1142, STOCKFISH 170514 and KOMODO 1333.

The reproducibility of computational experiments certainly requires single-thread mode [17], non-learning mode, and the full specification of UCI parameters.[3] Figure 4 is a proposed data-model which may be implemented in Relational or XML databases. The advent of the web-related XML family of standards[4] and the lighter weight 'JSON' Javascript Object Notation have greatly improved the communication and manipulation of hierarchical data.

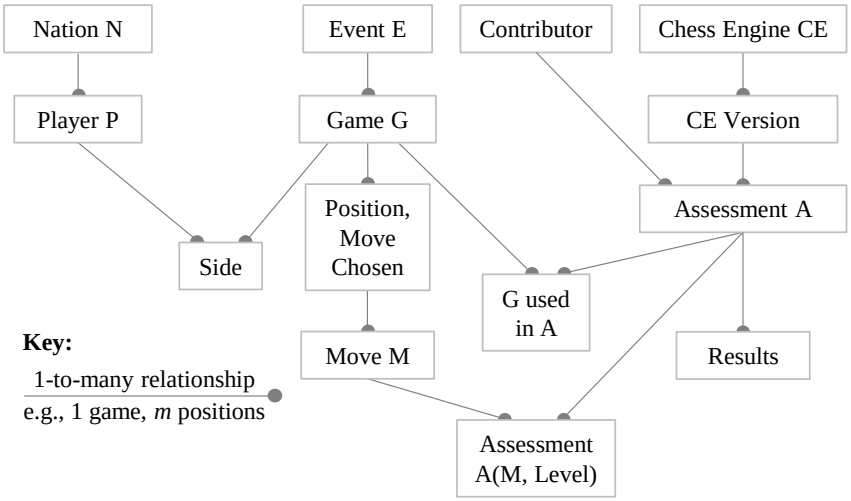

Fig. 4. Data Model: Computer-assessments of game-moves at various search-levels.

6 Summary and View Forward

A number of skill assessment methods have been compared. They vary in their conclusions but differences between workers' computations make definitive comparison difficult at this time.

Greater interworking within the community of those interested in skill assessment is required to quantify the intuitive, widely held but qualitative belief that:

"The more information is used by a method, the better the method is."

Specifically, here, it is believed that MM ⟨ EV ⟨⟨ AD ⟨⟨ ASD ⟨⟨ SR/SK ⟨ SRA.[5] The FIDE/ACP committee certainly regards MM/EV/AD as 'screening methods' but looks to more informed methods for definitive assessments [1].

Therefore the first requirement is to agree on a shared computational approach and on a set of computation subjects in order to quantify the belief above. Agreed tools and data-management interfaces will facilitate progress within the community.

[3] UCI = Universal Chess Interface [11]
[4] An example of chess-position in XML format is given in [17].
[5] The notation here is: ⟨ means 'worse than' and ⟨⟨ means 'much worse than'.

Finally, the authors have sought here not only to bring a new coherence to the community of those assessing chess skill but to explore better ways to communicate the subtleties of assessment to the non-specialist and the public. Data supporting this article is freely available and is being evolved [24].

Acknowledgements. In addition to thanking various parties for stimulating discussions on skill assessment, the authors thank David Barnes and Julio Hernandez Castro of the University of Kent for advance and subsequent discussion of their paper [17].

References

1. FIDE/ACP. Anti-Cheating Guidelines approved by FIDE. http://www.fide.com/images/-stories/NEWS_2014/FIDE_news/4th_PB_Sochi_Agenda_Minutes/Annex_50.pdf (2014)
2. Bayes, T.: An Essay towards solving a Problem in the Doctrine of Chances. *Phil. Trans. Royal Soc.*, Vol. 53, pp. 370-418. doi:10.1098/rstl.1763.0053 (1763)
3. Haworth, G.M^{c}C.: Reference Fallible Endgame Play. ICGA J. 26-2, 81--91 (2003)
4. Haworth, G.M^{c}C. and Andrist, R. B.: Model endgame analysis. In: van den Herik, H. J., Iida, H. and Heinz, E. A. (eds.) *Advances in Computer Games: Many Games, Many Challenges*. Advances in Computer Games, 135 (10). Kluwer Academic Publishers, Norwell MA, pp. 65-79. ISBN 9781402077098 (2004)
5. Andrist, R. B. and Haworth, G.M^{c}C.: Deeper model endgame analysis. Theoretical Computer Science, 349 (2). pp. 158-167. ISSN 0304-3975 doi: 10.1016/j.tcs.2005.09.044 (2005)
6. Haworth, G.M^{c}C.: Gentlemen, Stop Your Engines! ICGA J. 30-3, 150-6 (2007)
7. Di Fatta, G., Haworth, G.M^{c}C., Regan, K.: Skill Rating by Bayesian Inference. In: IEEE CIDM Symposium on Computational Intelligence and Data Mining (2009)
8. Haworth, G.M^{c}C., Regan, K. and Di Fatta, G.: Performance and prediction: Bayesian modelling of fallible choice in chess. Lecture Notes in Computer Science, 6048. pp. 99-110. ISSN 0302-9743 doi: 10.1007/978-3-642-12993-3_10 (2010)
9. Di Fatta, G. and Haworth, G.M^{c}C.: Skilloscopy: Bayesian modeling of decision makers' skill. Systems, Man, and Cybernetics: Systems, IEEE Transactions on , 43 (6). 1290-1301 . ISSN 0018-9472. doi: 10.1109/TSMC.2013.2252893 (2013)
10. FIDE Online Arena with AceGuard: http://www.fide.com/fide/7318-fide-online-arena.html, http://arena.myfide.net/ (2015)
11. Huber, R., Meyer-Kahlen, S.: http://wbec-ridderkerk.nl/html/UCIProtocol.html. Universal Chess Interface specification (2000)
12. TCEC: http://tcec.chessdom.com/archive.php. Thoresen's Chess Engine Competition, Seasons 1-7 (2014)
13. Ferreira, D.R.: The impact of search depth on chess playing strength. ICGA J. Vol. 36, No. 2, pp 67-90 (2013)
14. Regan, K.W., Macieja, B. and Haworth, G.M^{c}C.: Understanding Distributions of Chess Performance. ACG13: Advances in Computer Games. Tilburg, the Netherlands. LNCS 7168, pp. 23-243. doi: 10.1007/978-3-642-31866-5_20 (2012)
15. Friedel, F.: Cheating in Chess. In: Advances in Computer Games 9, pp. 327-346. Institute for Knowledge and Agent Technology (IKAT), Maastricht, The Netherlands (2001)
16. Biswas, T., Regan, K.W.: Quantifying Depth and Complexity of Thinking and Knowledge. ICAART 2015, the 7th International Conference on Agents and Artificial Intelligence, January 2015, Lisbon, Portugal (2015)

17. Barnes, D.J., Hernandez-Castro, J. On the limits of engine analysis for cheating detection in chess. *Computers and Security*, Vol. 48, pp. 58-73 (2015)
18. Guid M., Bratko, I.: Computer Analysis of World Chess Champions. ICGA J. 29-2, 65-73 (2006)
19. Riis, S.: http://www.chessbase.com/newsdetail.asp?newsid=3465. Review of "Computer Analysis of World Champions" (2006)
20. Guid, M., Perez, A., Bratko, I.: How Trustworthy is CRAFTY's Analysis of Chess Champions? ICGA J. 31-3, 131-144 (2008)
21. Ferreira, D.R.: Determining the Strength of Chess Players Based on Actual Play. *ICGA J.*, Vol. 35, No. 1, pp. 3-19 (2012)
22. Regan, K.W. and Haworth, G.M^{c}C.: Intrinsic chess ratings. AAAI-11, the 25th AAAI Conference on Artificial Intelligence, 07-11 August 2011, San Francisco, USA, pp. 834-839. ISBN: 9781-5773-5507-6 (2011)
23. Regan, K.W., Biswas, T.: Psychometric Modeling of Decision Making Via Game Play. CIG-13, the 2013 IEEE Conference on Computational Intelligence in Games, August 2013, Niagara Falls, Canada (2013)
24. Haworth, G.M^{c}C., Biswas, T., Regan, K.W.: This article and related, evolving datasets including pgn and statistics files. http://centaur.reading.ac.uk/39431/ (2015)