# Evaluating the Aesthetics of Endgame Studies

## A Computational Model of Human Aesthetic Perception

Azlan Iqbal[1], Harold van der Heijden[2], Matej Guid[3] and Ali Makhmali[4]

[1,4]College of Information Technology, Universiti Tenaga Nasional, Selangor, Malaysia,
azlan@uniten.edu.my, ali_makhmali@uniten.edu.my
[2]Deventer, the Netherlands, heijdenh@concepts.nl
[3]AI Laboratory, Faculty of Computer and Information Science, University of Ljubljana, Slovenia, matej.guid@fri.uni-lj.si

*Abstract*—**In this article we explain how an existing computational aesthetics model for three-move mate problems was improved and adapted to suit the domain of chess endgame studies. Studies are typically longer and more 'sophisticated' in terms of their perceived aesthetics or beauty. They are therefore likely a better test of the capability of machines to evaluate beauty in the game. Based on current validation methods for an aesthetics model such as this, the experimental results confirm that the adaptation was successful. In the first experiment, the new model enabled a computer program to distinguish correctly between composed studies and positions with sequences resembling studies taken from real games. In the second, the computational aesthetic evaluations were shown to correlate positively and well with human expert aesthetic assessment. The new model encompasses the previous three-mover one and can be used to evaluate beauty as perceived by humans in both domains. This technology pushes the boundaries of computational chess and can be of benefit to human players, composers and judges. To some extent, it may also contribute to our understanding of the psychology of human aesthetic perception and the 'mechanics' of human creativity in composing problems and studies.**

***Keywords: -aesthetics; chess; endgame; perception; creativty***

## I. INTRODUCTION

Computational recognition and evaluation of aesthetics or beauty in chess is a relatively new frontier of investigation in the game. Now that computers are able to defeat the best human players even on commercially available hardware [1], the concepts of beauty and creativity in the game have become of interest to some in the artificial intelligence (AI) community. It can be seen as an alternative direction of exploration compared to investigating games of greater complexity such as Go [2]. It may even be a more viable alternative in the interests of 'true' AI since the recognition and evaluation of beauty represents 'softer' aspects of human intelligence that are not as well-defined, and likely less susceptible to 'tricks' of programming, as is the case with many zero-sum perfect information board game *playing* technologies. Also, beauty usually stems from creativity which is difficult to describe in terms of a typical process [3-4].

The importance of aesthetics in the game of chess lies not only in the appreciation that human players and composers have for it but also in the larger implications to the boundaries of computation. It was once widely believed that a computer capable of playing chess on the level of a human expert would be proof of the ability to mechanize intelligence such that it could be applied to other domains, or that a computer could indeed 'think' [5].

However, efficient algorithms, clever programming techniques and brute force computation on powerful hardware proved to be the most effective solution tailored to this particular task. This is not to say that chess-playing programs are not an achievement or milestone in AI [6]. Indeed they are, and it is common for master players and composers today to use such programs to improve their games and compositions [7-8]. Lessons learned from computer chess are relevant to many problems in computer science – especially those that depend on search, e.g., automated reasoning, molecular synthesis, scheduling problems and even designing computers [9]. There have been investigations into computational aesthetics or a similar concept in other domains such as visual art [10-12] and music [13-15] but chess is perhaps the most amenable to computation given its finite and mathematical nature yet very high complexity that facilitates seemingly limitless beauty and creativity [16].

Section II describes briefly the existing three-move mate problem aesthetics model and its recent improvements; section III explains the new endgame studies model with an example evaluation of a study and a classic study evaluated in more detail; section IV presents the experimental results; and section V covers some of the issues related to human expert evaluation of aesthetics in the game. Section VI concludes with a summary of the main points and suggests some directions for future work. Appendix A includes examples that show the contrast between beautiful and less beautiful problems, and also the contrast between 'visual appeal' and 'depth appeal' in compositions. Appendix B provides some empirical support for the proposed changes in the new aesthetics model.

## II. THE THREE-MOVE MATE MODEL

The existing three-move mate aesthetics model was developed for the purpose of investigating if a computer could be made to detect and evaluate beauty in the game of chess within that scope (see Appendix A for examples). Experimental results showed that this was indeed the case [17-

This research is sponsored by the Ministry of Science, Technology and Innovation (MOSTI) in Malaysia under their eScienceFund research grant (01-02-03-SF0188).

19]. Essentially, seventeen aesthetic features comprising seven carefully-selected aesthetic principles (i.e., *violate heuristics successfully*, *use the weakest piece possible to checkmate*, *use all of the piece's power*, *win with less material*, *checkmate economically*, *sacrifice material, spread out the pieces*) and ten themes (i.e., *fork*, *pin*, *skewer*, *x-ray*, *discovered/double attack*, *zugzwang*, *smothered mate*, *cross-check*, *promotion*, *switchback*) were formalized as evaluation functions. These features are evaluated, as applicable, at different 'points of evaluation' (POE); i.e., in the initial position (POE = 0), after each move by the winning side (POE = 1, 2, 3) and the final position (POE = 4). More details on this classification are available in section 3.8 of [17].

A three-move mate sequence could then be computationally analyzed for the presence of these features and evaluated based on their corresponding evaluation functions for a (tentative) cumulative aesthetic score. Each feature that is evaluated at more than one point in the move sequence would have a cumulative score or 'total' of its own. This total is almost always between 0 and 1. Two detailed examples of such evaluations, move by move, are provided in section 6.4 of [17]. They are not included here due to space constraints and to minimize redundancy. Table I shows the aesthetic feature scores of a sample mate-in-3 combination or sequence. Next to the 'POE' are the seven *aesthetic principles* in the same order mentioned at the start of this section. There were no detected *themes* in this sample.

TABLE I. FEATURE SCORES OF A SAMPLE MATE-IN-3 SEQUENCE

| POE | VH | WPP | APP | WLM | ECO | SAC | SPA |
|---|---|---|---|---|---|---|---|
| 0 | | | | 0 | | | 0.417 |
| 1 | 0.083 | | 0.250 | | | | |
| 2 | | | 0.357 | | | | |
| 3 | | 0.400 | 0.357 | | | | |
| 4 | | | | | 0.750 | 0.357 | |
| **Total** | **0.083** | **0.400** | **0.964** | **0** | **0.750** | **0.357** | **0.417** |

The *use all of the piece's power* principle (APP) is evaluated at three points in the combination and has a cumulative score or total of 0.964. The *win with less material* principle (WLM) is evaluated only in the initial position (POE = 0) but is not applicable in this particular combination so it has no score. The combination included some material sacrifice – assessed at the end of the sequence or the final position – and has a score of 0.357 for that principle. The overall aesthetic evaluation for the combination continues like this.

Looking at the feature totals (the bottom row in Table I), the average sum of the *top-scoring* 5 or 6 features in the sequence, over 20 'iterations' [19], is used. The determination of whether to use 5 features or 6 – at each iteration – is random and based on a 20-80 probability split, i.e., a 20% chance of using 5 features and an 80% chance of using 6. The average sum of the remaining features that score > 0 [19] is subtracted for an 'adjusted' summative score. This is done for the main line, and also each of the 5 or 6 randomly-selected alternate variations, if any. If there are only 5 or fewer features or variations, all are taken into account and no randomness applies.

Comparing the average of the adjusted scores given all the lines (i.e., the main line *and* alternate variations), against the adjusted score of just the main line – the higher one is taken as the final aesthetic score for the three-mover. Table II shows the 20 iterations, the number of features randomly selected to use at each one (i.e., 5 or 6), the sum of those top 5 or 6 features, the sum of the remaining features (that scored > 0), the adjusted summative score (after subtracting the remaining features) and the cumulative total (an incrementing tally of the adjusted scores).

TABLE II. SCORING A PARTICULAR VARIATION OVER 20 ITERATIONS

| Iteration | Top Features to Use | Sum of Top Features | Sum of Remaining Features | Adj. Score | Cum. Total |
|---|---|---|---|---|---|
| 1 | 5 | 2.888 | 0.083 | 2.805 | 2.805 |
| 2 | 6 | 2.971 | 0 | 2.971 | 5.776 |
| 3 | 6 | 2.971 | 0 | 2.971 | 8.747 |
| 4 | 6 | 2.971 | 0 | 2.971 | 11.718 |
| 5 | 6 | 2.971 | 0 | 2.971 | 14.689 |
| 6 | 6 | 2.971 | 0 | 2.971 | 17.66 |
| 7 | 6 | 2.971 | 0 | 2.971 | 20.631 |
| 8 | 6 | 2.971 | 0 | 2.971 | 23.602 |
| 9 | 6 | 2.971 | 0 | 2.971 | 26.573 |
| 10 | 6 | 2.971 | 0 | 2.971 | 29.544 |
| 11 | 5 | 2.888 | 0.083 | 2.805 | 32.349 |
| 12 | 5 | 2.888 | 0.083 | 2.805 | 35.154 |
| 13 | 6 | 2.971 | 0 | 2.971 | 38.125 |
| 14 | 6 | 2.971 | 0 | 2.971 | 41.096 |
| 15 | 6 | 2.971 | 0 | 2.971 | 44.067 |
| 16 | 6 | 2.971 | 0 | 2.971 | 47.038 |
| 17 | 6 | 2.971 | 0 | 2.971 | 50.009 |
| 18 | 6 | 2.971 | 0 | 2.971 | 52.98 |
| 19 | 6 | 2.971 | 0 | 2.971 | 55.951 |
| 20 | 6 | 2.971 | 0 | 2.971 | **58.922** |

The average adjusted score is equal to 58.922 divided by 20, or 2.946. This will be the aesthetic score *for the main line or particular alternative variation that was analyzed in the sequence* (but not the final aesthetic score). The average adjusted score of *all* the lines (main line and variations) must be compared against the adjusted score for just the main line; the higher of the two is then chosen as the final aesthetic score for the combination. If the evaluation for the same combination was performed all over again (i.e., it is subjected to another 'cycle' or 'look'), the final aesthetic score could vary slightly. This is because for the main line or any of the variations there may be, say, five instances where 5 features are selected (instead of three as shown in Table II); affecting the average of the adjusted scores for that line and also the average score for all the lines that will be compared against that of the main line in order to determine the final aesthetic score.

The "5 or 6" features/variations and "20-80" probability split numbers mentioned earlier in this section, compared to lower and higher alternatives, proved to be the best compromise and most viable. Determination was based on the

normality of the results (using the Anderson-Darling Test Statistic) and significant positive Spearman rank correlation with mean human aesthetic assessment, as explained in [19]. Using the aforementioned numbers, the normality test statistic was 0.550 and the correlation was 0.648 (see Appendix B for a more detailed explanation). The capability to distinguish *between* compositions and real game sequences (but not *within* either) was also confirmed. The main change here compared to [19] is that variations are taken into account (previously, only the main line was assessed). Since including variations improves the normality and maintains the correlation, doing so is now part of the three-mover model.

## III. THE ENDGAME STUDIES MODEL

### A. New Challenges Compared to Three-Movers

The three-mover model explained in section II was adapted in order to function in the more complex domain of endgame studies. Studies are essentially compositions that have longer move sequences and where the initial position resembles some point in the endgame (when there are fewer pieces on the board). The solution is typically more 'open-ended' or less decisive than a three-mover. While three-movers have a stipulation such as "*White to play and mate in three moves*", studies have a stipulation such as "*White to play and win*" or "*White to play and draw*". This poses many challenges to the existing three-mover model. Specifically, the following had to be addressed.

(a) The longer sequence of moves and its effect on the aesthetic score.

(b) The detection of the *zugzwang* theme.

(c) The evaluation of the *use the weakest piece possible to checkmate* principle, the *checkmate economically* principle and the *smothered mate* theme.

(d) The selection of positions with sequences resembling studies taken from real games to compare against composed studies for experimental purposes.

The issue of studies having a longer sequence of moves is significant, especially in terms of an ever-inflating aesthetic score. A sequence should not be considered more beautiful simply because it is longer and has more 'content'. Aesthetic perception is unlikely a 'linear' affair. The opposite is probably true because 'difficulty' or 'complexity' (that tends to increase with the number of moves) is seldom seen as beautiful in itself [20]. In order to address this issue, cut-off points were introduced for all aesthetic features that are evaluated at 3 points of evaluation (POE) in the three-mover model because these have the potential of continuously increasing in magnitude with additional moves. The change affects seven features: *use all of the piece's power*, *fork*, *pin*, *skewer*, *x-ray*, *discovered/double attack* and *promotion*. Any of these features scoring more than 1.0 in total is scaled down to 1.0 as its 'feature total' (see Table I). The average was not used because it would dampen the effect of a particularly beautiful move within the sequence.

In studies, these features are assessed after every move in the sequence so it continues to the point at which the total for such a feature reaches or exceeds 1.0 where it is then 'capped' at 1.0 and ceases to increase any further. While it may seem intuitive that studies are more likely to achieve this for the seven features and score higher overall primarily because they are longer, no significant difference between the scores of three-move problems and studies could be detected, i.e., sequences from both domains had scores that ranged between 0-6.

In the three-mover model, zugzwangs are assessed after White's second move but in studies – with no decisive ending such as mate – it is theoretically impossible to determine with consistency if a zugzwang exists at any point in the sequence (i.e., it would require testing after every move to an indefinite depth). Therefore zugzwangs are tested only in cases of three-movers and three-movers presented as studies.

A similar situation existed for the features *use the weakest piece possible to checkmate*, *checkmate economically*, and *smothered mate*. In the three-mover model they were evaluated in the final (checkmate) position only. It was difficult to develop formulas for these that could apply consistently over a sequence of virtually any move length so they were tested only in cases where the study ended in mate, though not necessarily limited to three-movers.

In testing the three-mover model, it was relatively easy to find sequences of forced three-movers from real games that ended in mate to contrast against compositions of the same length; the latter being, on average, accepted as more beautiful. For studies it was more difficult because the end is often *not* mate. An FIDE endgame study judge[1] was therefore enlisted to select positions from real games with sequences that resembled study sequences. These were sourced from the ChessBase Fritz 10 database (1,128,478 games) with Chess Query Language (CQL)[2] and used to compare against composed studies to test the hypothesis that the compositions would score aesthetically, on average, higher than the real game sequences.

The selection criteria included considerations of material balances that are often used in studies (e.g., 'pawnless' endings, or queen and bishop against queen and any number of pawns, but at least 5 plies on the board) and study-like thematic moves. For example, the 'unguarded guard' theme, where in response to a check, an unprotected piece is interposed. In all cases only won games (by White) were considered. These game fragments were checked manually to remove non-thematic ones (e.g., pawnless endings with a large material advantage where Black just refused to resign). The *compositions* were sourced, at random, from an endgame study database with 76,132 studies (the largest collection in the world).[3]

### B. Changes to the Three-Mover Model

A new computer program, CHESTHETICA EG, was developed using the Visual Basic 6 programming language (over the course of about 18 months) to test the new model designed with endgame studies in mind (see Fig. 1).

[1] Harold van der Heijden.
[2] http://rbnn.com/cql/contents.html
[3] http://www.hhdbiv.nl


Figure 1. CHESTHETICA EG v0.03 main interface.

Some notable changes to the three-mover model in terms of the evaluation functions used are as follows. The *switchback* theme was, originally [17], calculated as the number of switchbacks detected divided by 2 (the constant or maximum number of switchbacks possible in a three-mover). For studies, the constant was replaced by *the number of plies in the sequence* divided by 2; the decimals truncated. For example, if 3 switchbacks were detected in a sequence with 5 moves (or 9 plies) the score would be: 3 / INT (9 / 2) = 3 / 4 = 0.75. Therefore, this is still applicable to three-movers (with 5 plies).

The *use the weakest piece possible to checkmate* formula was changed to: 1 / 5 × *piece value hierarchy*. The piece value hierarchy values for queen, rook, bishop, knight and pawn are 1 through 5, respectively. So using the queen to mate scores the lowest (i.e., 0.2) whereas using a pawn scores the highest (i.e., 1). This change was not due to the adaptation to endgame studies per se but rather because it provided a more proportional measure of using the weakest piece possible to checkmate.

For the principle *use all of the piece's power*, the evaluation at POE 4 (see Table I) was removed [17]. This was to minimize redundancy with the *use the weakest piece possible to checkmate* principle just explained. All the changes described in this section apply automatically to the new endgame studies model.

### C. Example Evaluation of a Study

Complete details pertaining to the individual aesthetic feature evaluation functions would be too lengthy for inclusion here, but they are provided in [17] and supplemented in [18-19] for the interested reader. Fig. 2 shows the moves of the example study as they appear in the popular chess database management program ChessBase. The initial or key move (*1.Qh4*) is highlighted by default, and the primary or main line in bold font.[4] The starting position is shown in Fig. 3 with the main line just below it. In addition to the main line, CHESTHETICA EG here selects 5 variations instead of 6 out of the 10 listed in the PGN file.[5] The selection is random and based on the 20-80 probability split (see section II).

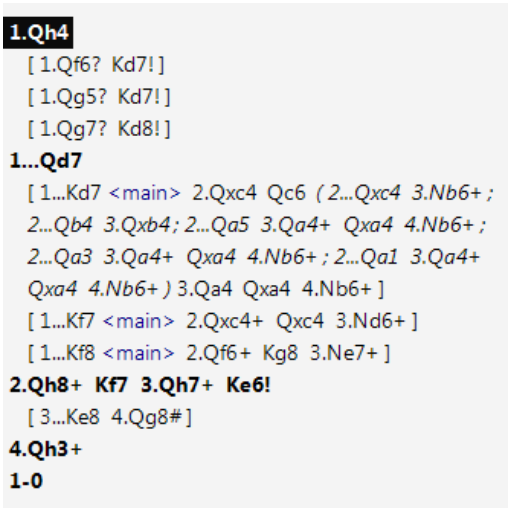
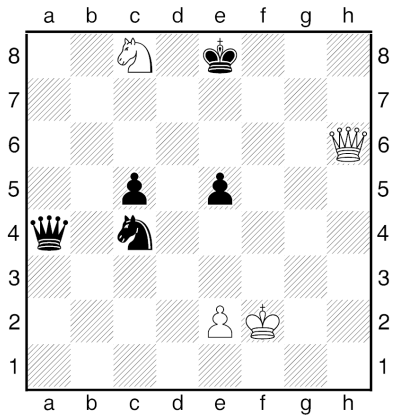



Figure 2. The main line and variations of the study.



H. Rinck .c Rigaer Tageblatt, 1905

1. Qh4 Qd7 2. Qh8+ Kf7 3. Qh7+ Ke6! 4. Qh3+

Figure 3. Initial position of the study and its main line.

For sequences from real games that resemble studies (that typically do not have the variations listed in the PGN file), a chess engine ('Houdini' in our case [21]) is used to generate the best alternate variations – limited in number to the top 16 or so for speed. From these the selection of 5 or 6 variations is made using the same probability split as before. In relatively rare cases, composed studies may not come with variations included and are treated as having none (just the main line).

[4] The "<main>" in the other variations (not in bold) are sub-lines intended as alternative main lines. In the model they are treated as alternative variations.

[5] PGN (Portable Game Notation) is a standard plain-text format for recording chess games and compositions. It can be read and processed by many computer chess programs.

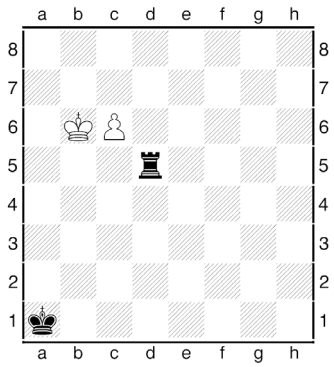
For each variation, including the main line, the top 5 or 6 aesthetic features is selected and assessed as explained in section II. In this example, and based on a single cycle of evaluation (i.e., one 'look'), the main line's aesthetic score is 2.266. The other 5 lines scored 2.560, 2.405, 2.399, 2.390 and 2.396. The average score of these 6 lines (i.e., including the main line) is calculated as 2.403. Since the overall average score (2.403) is higher than the main line's score (2.266), it is used as the final aesthetic score for this particular study. Alternatively, the process can be described using the following pseudo-code.

```
IF composed study THEN
   Read main line from PGN file
   Read variations (if any) from PGN file
ELSE (if real game sequence)
   Read main line from PGN file
   Read best variations generated by chess engine
End IF

IF variations > 5 THEN
   Choose random number between 1 and 100
   IF random number ≤ 20 THEN
      Select 5 variations at random
   ELSE
      Select 6 variations at random
   End IF
ELSE
   Select all variations
End IF

FOR each variation and the main line
   FOR 20 iterations
      Choose random number between 1 and 100
      IF random number ≤ 20 THEN
         Select the 5 top-scoring features
      ELSE
         Select the 6 top-scoring features
      End IF
      Sum the top-scoring feature scores
      Sum the remaining feature scores which have a
      positive value
   End FOR
   Line score = (average score of top-scoring
   features) – (average score of remaining features)
End FOR

Final aesthetic score = MAX (main line score, average
of all lines score)
```

### *D. A Classic Study Evaluated in More Detail*

Fig. 4 shows a famous study from the late 19th century known as the "Saavedra position" (see Fig. 4). It is named after Fernando Saavedra (1847-1922), a Spanish monk who spotted the win in a position that was previously thought to have been drawn [22-23]. A famous, 'good' or educational study such as this is not necessarily a prime candidate for high *aesthetic content*, but it is something readers are more likely to be familiar with and therefore a worthy specimen to use in illustrating the finer aspects of the proposed computational aesthetics model.

Considering just the main line (in bold in Fig. 4), the computational aesthetic evaluation is summarized as shown in Table III. The first three features (SPA, APP, WLM) are the same as in Table I. The other two represent the themes *promotion* and *switchback* (see section II).



F. Saavedra (correcting G. E. Barbier),
*Glasgow Weekly Citizen*, 1895

**1. c7 Rd6+ 2. Kb5** (2. Kc5 Rd1 3. ... Rc1) **Rd5+ 3. Kb4 Rd4+ 4. Kb3 Rd3+ 5. Kc2 Rd4 6. c8R** (threatening 7. Ra8+; if 6. c8Q Rc4+ 7. Qxc4 stalemate) **Ra4 7. Kb3 Kb1 8. Kxa4**

Figure 4. The Saavedra position and its solution.

TABLE III. FEATURE SCORES OF THE SAAVEDRA POSITION (MAIN LINE)

| Feat. | | SPA | APP | WLM | PRO | SB |
|---|---|---|---|---|---|---|
| POE | 0 | 0.500 | | 0.105 | | |
| | 1 | | 0.250 | | | |
| | 2 | | 0.125 | | | |
| | 3 | | 0.125 | | | |
| | 4 | | 0.125 | | | |
| | 5 | | 0.125 | | | |
| | 6 | | 0.250 | | 0.600 | |
| | 7 | | 0.125 | | | |
| | 8 | | 0.125 | | | |
| | 9 | | | | | 0.143 |
| **Total** | | 0.500 | 1.000* | 0.105 | 0.600 | 0.143 |

* Capped value

In the initial position (POE = 0), the 7th aesthetic principle (see section II) or 'sparsity' (i.e., the extent that the position is spread out) is evaluated as shown in (1).

$$P_7 = \left[ \left( n^{-1} . \sum_1^n s\left(p_n\right) \right) + 1 \right]^{-1} \quad (1)$$

$s(p_n)$ = the number of pieces in the field of a particular piece

The full details and logic of this evaluation function are provided in section 4.7 of [17] and also [24]. In short, every piece on the board is examined for pieces that reside immediately around it for an 'inverse density' score of the position. In Fig. 4 the white king has only a pawn in its immediate surroundings and so does the rook. The pawn has two pieces around it. The black king has no pieces around it so the sparsity feature score amounts to: $1 / [(1 / 4 \times 4) + 1] = 0.5$.

At POE = 0 we also have the principle of *win with less material* and it is evaluated as shown in (2); *v()* denotes the Shannon value of the piece. The principle only applies if Black

has more material than White at the start. More information about it is provided in section 4.4 of [17]. The score for this principle is therefore: 1 / 38 × (5 – 1) = 0.105.

$$P_4 = wlc^{-1}.\left[v(p_b) - v(p_w)\right], v(p_b) > v(p_w) \quad (2)$$

$wlc$ = principle constant (i.e., 38), $p_{b/w}$ = all Black/White pieces in the position

The APP (*use all of the piece's power*) principle is applicable after every move of the winning side (POE = 1-8) and relates to the nature and movement of the piece that just moved. Collectively (i.e., for the whole move sequence), it is evaluated as shown in (3). Further details pertaining to this evaluation function are provided in section 4.3 of [17], except for the minor change mentioned at the end of section III(B). So in this classic study, there is no evaluation for this principle at POE = 9.

$$P_3 = \sum_1^n \left(d(mp)_n\right).\left(r(mp)_n\right)^{-1} \quad (3)$$

$mp$ = the moving piece, $n$ = evaluation point
$b_k$ = the black king, $d$ = distance traveled, $r$ = piece power

After the first move (*c7*), the APP score is calculated as: 1 × 1 / 4 = 0.25, given that the pawn moved a distance of only one square and that its 'piece power' is 4 (see section 3.5.2(b) of [17] for more on this measure of a piece's inherent mobility). After the second move (*Kb5*), the APP score is: 1 × 1 / 8 = 0.125. This continues until after move 8, where the total score for the feature adds up to 1.25. However, to keep the value from increasing indefinitely in studies, there is a maximum cap of 1.0 (see section III(A)).

A promotion occurs on the 6th move (*c8R*) and its score is calculated as the reciprocal of the Shannon value of the promotion piece choice multiplied by the theme constant, i.e., 3 (further details and justifications for this function are available in section 5.9 of [17]). The score here is therefore: 1 / 5 × 3 = 0.6. The *switchback* theme (which for our purposes also includes the *rundlauf* or 'round-trip' theme) also occurs in the sequence and is assessed in the final position as explained in section II (in this study, POE = 9); the king 'originates' at *b3* on move 4, moves off it and then returns to *b3* on move 7. Given the formula mentioned in section III(B), the score for the switchback is: 1 / INT (15 / 2) = 1 / 7 = 0.143.

Since only five features (out of a total of seventeen, see section II) were detected in the main line, there is no need to select 5 or 6 randomly using a 20-80 probability split. There is also nothing to subtract in terms of 'remaining features' that scored > 0. The 'adjusted summative score' is therefore the same, i.e. 2.348. Ordinarily, all of the above would represent just one out of the 20 iterations for the main line. Assuming more than 5 features had been detected, in the second or third iteration, the score could be slightly different than 2.348.

Not only is the main line evaluated as above, but 5 or 6 of the *variations* as well – all in the same way as the main line (i.e., each also with 20 iterations). In the case of the Saavedra position, it was treated as a composed study (since the variations were provided in the PGN file we had). In principle, it could also be treated as a 'real game sequence' with the variations generated by a chess engine but these could be missing some lines the composer or publisher intended for us to see.

There were 7 variations in addition to the main line, and based on the 20-80 probability split, 6 were chosen at random as shown in Table IV (numbered in no particular order). The score for each of the variations shown in Table IV is actually the average for that line after 20 iterations. The main line scored 2.348 and therefore the average for *all* these variations including the main line is: (2.105 + 3.064 + 3.050 + 1.730 + 1.605 + 1.563 + 2.348) / 7 = 2.209.

TABLE IV. AVAILABLE VARIATIONS IN ADDITION TO THE MAIN LINE

| | Variation | Score |
|---|---|---|
| 1 | *1. c7 Rd6+ 2. Kb5 Rd5+ 3. Kb4 Rd4+ 4. Kb3 Rd3+ 5. Kc2 Rd4 6. c8Q Rc4+ 7. Qxc4* | 2.105 |
| 2 | *1. c7 Rd6+ 2. Kb5 Rd5+ 3. Kb4 Rd4+ 4. Kb3 Rd3+ 5. Kc2 Rd4 6. c8R Rh4 7. Ra8+ Ra4 8. Rxa4#* | 3.064 |
| 3 | *1. c7 Rd6+ 2. Kb5 Rd5+ 3. Kb4 Rd4+ 4. Kb3 Rd3+ 5. Kc2 Rd4 6. c8R Ra4 7. Kb3 Rh4 8. Rc1#* | 3.050 |
| 4 | *1. Kb7 Rc5 2. c7 Rxc7+ 3. Kxc7* | 1.730 |
| 5 | *1. c7 Rd6+ 2. Kc5 Rd1 3. Kb6 Rc1 4. Kb7 Rxc7+ 5. Kxc7* | 1.605 |
| 6 | *1. c7 Rd6+ 2. Ka7 Rc6 3. Kb7 Rxc7+ 4. Kxc7* | 1.563 |
| 7 | *1. c7 Rd6+ 2. Kb7 Rd7 3. Kb8 Rxc7 4. Kxc7* | *Not chosen* |

Since the average (2.209) is a lower value than the main line's score (2.348), the latter is used as the final aesthetic score for this classic study. This is neither a poor score nor an outstanding one, aesthetically (we return to this point toward the end of section IV(A)). Incidentally, had variation no. 2 or 3 (see Table IV) – that allow White to mate more quickly – been presented as the main line in the study, it would have scored higher. Note, however, that all this represents one cycle or 'look' and while it is sufficient in most cases, a 'crisp' or 'final' aesthetic assessment for a study may sometimes be desired (e.g., in composition tournaments or 'tourneys'). This can be obtained by using instead the average of three cycles or more. While the aesthetic evaluation of a single study is not likely to change in magnitude by much, its ranking in a collection of studies might.

Also noteworthy is that while the feature evaluation functions play a critical role in the aesthetics model, the final aesthetic score is distinct from their mere 'summation' and does not necessarily suffer due to a lack of features that some human experts may insist upon. Just as we typically treat the mind of a human expert like a 'black box' (without, say, scrutinizing the logic behind the firing of specific neurons in his brain or their 'senselessness' in relation to beauty as we

know it), the aesthetics model should also be afforded the same 'courtesy' in practical use so long as it has been experimentally-validated and its evaluations are reasonable.

## IV. EXPERIMENTAL RESULTS

### A. Comparisons between and within Domains

The validation method or 'qualifying standard' for the endgame studies model was based on the guidelines as described in [18]. In short, there should be *no* statistically significant difference between the mean aesthetic scores of two sub-samples (of equal size) *within* each domain, i.e., compositions and real game sequences. However, in total as a domain, the compositions must score aesthetically, on average, higher than the real game sequences; and the difference in means between the two domains should be statistically significant.

The first experiment consisted of 310 composed studies and 310 sequences taken from real games that resembled studies (see section III(A), the last two paragraphs, for source information and other details). This was the number of viable sequences that we had from real games; a matching number of composed studies was used for statistical consistency. Each was split evenly into two sub-samples of 155. Since the aesthetics model can generate a slightly different evaluation for a study or sequence if given another 'look', each was subjected to three cycles of evaluation and the average used. CHESTHETICA EG operated at a rate of approximate 9 endgame study compositions a minute and 6 real game sequences a minute on an AMD Athlon™ 64 X2 Dual Core 4600+@2.4 GHz desktop computer with 2 GB of DDR3 SDRAM running Windows XP Pro SP3.

The mean aesthetic score for each of the four sub-samples was calculated. Comparisons *between* and *within* the domains (composed studies, real game sequences) were done using a two-sample t-test assuming equal or unequal variances (two-tailed, significance level of 5%). The variances were first analyzed using an F-test to determine the proper t-test. The results are as shown in Table V. The rows labeled 't-stat' contain the statistical results for the comparisons made between and within domains. SD denotes 'standard deviation' for the item in the row immediately above it.

TABLE V. COMPARISONS OF MEAN AESTHETIC SCORES

| **Type** | Composed Studies | | Real Game Sequences | |
|---|---|---|---|---|
| **Database** | Sub-sample 1 | Sub-sample 2 | Sub-sample 1 | Sub-sample 2 |
| **Mean** | 2.902 | 3.126 | 1.267 | 1.255 |
| **SD** | 0.733 | 0.653 | 0.427 | 0.390 |
| **t-stat** | $t(308) = -2.841$, $P < 0.05$ | | $t(308) = 0.247$, $P > 0.05$ | |
| **Mean** | 3.014 | | 1.261 | |
| **SD** | 0.702 | | 0.408 | |
| **t-stat** | $t(496) = 37.997$, $P < 0.05$ | | | |

The difference in means *between* the domains of composed studies and real game sequences (310 each) was statistically significant with the former scoring significantly higher than the latter; more than double the magnitude, in fact.[6] However, *within* domains, the difference was statistically significant for only the composed studies, not the real game sequences. The distributions of both compositions and real game sequences were analyzed for normality using the Anderson-Darling Test. The compositions, in total, had a test statistic value of 0.520 (normal) but the real game sequences, in total, did not (4.255).

The comparison within the domain of real game sequences was therefore performed again but this time using the Mann-Whitney U test which is non-parametric, i.e., it does not make any assumptions about the probability distribution of the samples. Using this test, no significant difference in means was found between sub-sample 1 and sub-sample 2 for the real game sequences; $U_A = 11894.5$ ($z = 0.15$), $n_1 = n_2 = 155$, $P > 0.05$, two-tailed. Comparing both domains in total, compositions and real game sequences, the difference in means was still significant; $U_A = 2328$ ($z = 20.5$), $n_1 = n_2 = 310$, $P < 0.01$, two-tailed. Even though it was expected that *within* domains there should be no difference of statistical significance – i.e., no aesthetic differences, on average, between samples of the same type – the difference within composed studies was relatively small (0.224) and the difference within real game sequences smaller still (0.012).

In cases of a relatively small discrepancy but where statistical analysis shows it to be significant, other factors should be considered. For instance, the sub-sample size of 155 each within domains may have been insufficient for these comparisons or the compositions and real game sequences used may not have been of the best 'quality'. More likely is that 'statistical significance' in these cases should be seen in the proper context. Since the differences in means between the sub-samples were small (relative to their actual sub-sample means) it was decided that their statistical significance can be justifiably ignored. Similarly, should a relatively small difference have existed *between* domains, any claims of significance should also be reconsidered.

Perhaps a general rule to follow in computational aesthetics with regard to the 'significance' of a difference in means between two samples is that the difference should be more than 10% of the amount of the smaller of the two samples. So if one sample has a mean of 2.5 and the other a mean of 3.0, the difference should be both statistically significant *and* more than 0.25 in magnitude in order to be taken seriously. It is unclear at this time whether humans are able to perceive an aesthetic increment of ≤ 10%, but it seems safe to assume they are not. Given that most competent players and composers are probably able to tell a real game sequence from a composition that shares similar characteristics (see footnote 6), humans are most likely able to discriminate aesthetically given an increment in beauty of approximately 30%. A computer, however, is expected to be more 'sensitive'.

[6] On the same scale as 5,000 three-move mate problems that scored, on average, 2.303 and 5,000 three-movers taken from tournament games that scored, on average, 1.728, the results suggest that composed endgame studies (3.014) are generally aesthetically superior to three-move mate problems. A separate analysis of 4,394 composed endgame studies from the same database had a mean aesthetic score of 3.198. So overall, on average, three-move mate problems score about 30% higher than their real game counterparts, and composed studies about 30-40% higher than three-move mate problems.

The Saavedra position shown in section III(D) scored 2.348 aesthetically. Given its classic nature, we might have expected that it scored higher; perhaps significantly exceeding 3.0. A dispassionate look at the Saavedra position shows that, while it is undoubtedly a 'good' or educational study, its *beauty* may not be quite as stellar. Especially since, thanks to relatively recent endgame tablebases, it is now known that Black can resist longer by playing *3. ... Kb2*; this also forces White to promote to a queen in order to win rather than the more attractive and unexpected *under*promotion to a rook to which the Saavedra study arguably owes its fame.[7]

The real issue, however, may have more to do with the classic's emphasis on 'depth' as opposed to 'visual' appeal (see Fig. 6, Appendix A). With reference to the experimental results in Table V and the information in footnote 6, computationally, at 2.348, the Saavedra position is aesthetically well above something that typically occurs in a real game (1.261, 1.728) and leaning toward or on the side of a composition (3.014, 2.303). As mentioned toward the end of section III(D), the choice of the main line by the composer or publisher – as transcribed in the PGN file – is also a significant factor. Had no. 2 been chosen (see Table IV), the final aesthetic score would have been 3.064, since it is higher than the average of 2.209 (see same section).

Even so, such alternatives do not always represent the best defenses for Black. Unless a powerful chess engine is always used to determine the soundness or 'strength' of the main line presented relative to all alternative lines (not done in this research), aesthetic scores in these cases may become inflated. The Saavedra position was tested again, but this time it was treated as a 'real game sequence' which means that the main line was the same but the ('best') variations were generated by the Houdini chess engine (16 or so, from which 6 were selected at random). The final aesthetic score was still 2.348 given that the average was 2.213. Notably, the highest aesthetic score for an alternative variation – one of the 6 – was only 2.422.

### *B. Correlation with Human Expert Assessment*

The second experiment performed was to test the correlation strength of the computer's aesthetic evaluations against those of human experts. Since expertise in properly evaluating composed endgame studies is scarce, and the resources to obtain the opinions, under controlled conditions, of a sufficient number of players competent enough to evaluate beauty in move sequences also limited, we relied on the expertise of one FIDE judge of endgame study compositions and one FIDE master of chess.[8] They were asked to rate 30 randomly selected composed studies and 30 randomly selected real game sequences that resembled studies in terms of their beauty, on a scale of 0-10. These were sourced from the 76,132 sample database mentioned in section III(A) for studies and the '310' sample database used in the previous experiment for real game sequences.

The judge was asked to exclude considerations, like 'anticipation', that pertain specifically to composition *conventions* in that domain and rate them solely on beauty per se. The expert player was also asked to rate solely on beauty perceived. Their experiences in rating these studies are described in the following section. If the average assessments of competent human players (not necessarily experts) are preferred, they can be obtained with reasonable data integrity via an online survey as was done in [17]. Due to the relatively esoteric nature of endgame studies, this was considered unsuitable here.

The FIDE judge rated the studies and sequences as whole numbers whereas the FIDE master did the same for the studies but rated the sequences to one decimal place. The scores generated by the computer program – originally to three decimal places – were rounded to match their evaluations; whole numbers for the judge and to one decimal place for the master. As in the previous experiment the average of three cycles was used. Table VI shows the Spearman rank correlation coefficients (two-tailed, significance level of 1%). Since the evaluations pertain to beauty and it is better to have a larger sample size than 30, both the composed studies and sequences were combined when testing for correlation. The bottom row shows the correlation given the average score of both experts (taken to one decimal place). This is a fresh correlation coefficient calculated given the average score of both experts for each study and sequence evaluated.

TABLE VI. CORRELATION WITH HUMAN EXPERT ASSESSMENT

| **FIDE Endgame Study Judge** | **FIDE International Master of Chess** |
|---|---|
| 0.754 | 0.760 |
| 0.813 | |

The results show good correlations with the aesthetic assessments of both experts, including with their average aesthetic scores (which can be considered a *very* good correlation at > 0.8). Unexpectedly, correlation was about the same for both experts despite their vastly different domains of expertise. This suggests that they may have been able to isolate and focus on purely the *aesthetic* aspect of the studies and sequences they were asked to rate given that it exists in both their domains of expertise (compositions and real games). The even stronger correlation with their average scores suggests that the computer's evaluations mirror more closely a *collective* assessment by human experts (e.g., a panel of judges) than just that of an individual expert.

A perfect correlation of 1.0 with human expert assessment would be virtually impossible (and undesirable) given the many different tastes and biases almost certainly present in any kind of human assessment of aesthetics. In summary, the results here further validate the endgame studies model in terms of its capability to assess beauty in studies as perceived by humans competent in the game. It is important to note that, if comparing against aesthetic scores attributed by humans, the computer's scores should be taken only in terms of their utility in *ranking* compositions and sequences in a database or

[7] There is also the systematic maneuver of the white king and black rook and the double threat after *7. Kb3* (both of which are not explicitly accounted for in the model).

[8] Harold van der Heijden and Matej Guid, respectively.

collection. This is where the correlation with human assessment is likely to be found.

## V. THE EXPERTS HAVE THEIR SAY

### A. FIDE Endgame Study Judge

There are no standard criteria for judging of endgame studies. R. Pye in [25] lists: *novelty*, *thematic content* (motif), *naturalness* (credibility), *economy of means* (material, analyses), *activity*, *spaciousness*, *clarity*, *counter-play* (conflict), while J. Levitt and D. Friedgood in [26] use four elements to define chess beauty: *paradox* (surprise), *depth* (complexity), *geometry*, *flow*. A common problem in judging – especially when scores of different judges are compared – is personal taste. For instance, for me, surprise is a key element of an endgame study, and I consider complexity less important.

Despite that, in general there is consensus among judges which studies should *not* make it into an award (poor studies), and which studies are candidates for the top prizes. It proved very difficult to judge the game fragments (i.e., the positions with sequences that resembled studies) as if they were endgame studies. In the latter the solution is more compact, and almost every move has a particular point (motif), while in the game fragments, one or the other interesting move occurred, while the rest of the moves were just technical play.

### B. FIDE International Master of Chess

Most people seem to be convinced that evaluating aesthetics in chess games is a task which should be relatively easy for chess players but nearly impossible for a computer. But is it indeed so? Chess players are usually able to perceive beauty in particular sequences of moves in chess games; this ability can be seen in their comments in numerous chess books (e.g., GM Kasparov gives several comments on beautiful moves, combinations, sacrifices etc. in his well-known series of books *My Great Predecessors* [27]). However, they are very rarely (or practically never) faced with a task to determine *how* beautiful a certain sequence of moves is. In other words, chess players usually think in *qualitative* terms (e.g., a particular combination *is* beautiful in their opinion, or a certain sequence of moves is *more beautiful* than another sequence) rather than in *quantitative* terms, that is, to assign a specific aesthetic score to a certain sequence of moves.

The latter task also proved to be difficult for our expert player, which was reflected in a rather long time spent to assign scores to the sequences of moves in the two data sets that were subjects of evaluation – the expert spent more than 20 hours so that his numeric scores met all his qualitative comparisons of the particular sequences. Moreover, it was difficult for him to put seemingly 'plain' sequences of moves (some of which in his opinion still contained subtle points that can be perceived as somewhat beautiful) on the same scale as several combinations in the composed studies, which were much more beautiful in his opinion.

## VI. CONCLUSIONS

In this article we have shown how an aesthetics model for endgame studies was developed based on an existing model for three-move mates. As compared to the previous model [18-19], the new one incorporates variations. Based on experiments, it is able to assess automatically beauty in both three-movers and studies in such a way that resembles the assessments of humans competent in the domain. From a psychological standpoint, given the (apparently) necessary emphasis on just 5 or 6 of the top-scoring aesthetic features in each variation of a sequence and similar numbers of variations in addition to the main line (both based on a random 20-80 probability split), it would seem that humans are less objective and look at less than what they might think they do when it comes to beauty in the game.

In previous work [19], it was thought that the 5 or 6 feature selection phenomenon represented some kind of 'strictness rule' that suggested humans appreciate only the top 30% of aesthetic features associated with an object while 'penalizing' it for the rest. This is because there were seventeen features being evaluated in total and 5 or 6 represented approximately 30% of that. However, *this* research has shown that when it comes to factoring in variations for aesthetic analysis – which can number anywhere between 1 and 20,000 or more for a single combination or study – the 5 or 6 (in addition to the main line) applies still. The 'strictness rule' therefore likely does not apply when it comes to human aesthetic perception or at least does not pertain to variations. Humans may, in fact, be rather whimsical and quite limited in terms of the amount they can or care to see when it comes to evaluating beauty.

There are also many difficulties in doing so, even for experts. These difficulties can be attributed to both the differences of opinion that usually arise between people (e.g., players, composers, judges), and in the process of 'objectively' ranking the beauty of one combination or study relative to another, or several others. It becomes even more difficult for humans to evaluate beauty between, say, three-move mate problems and studies; which, though they involve the same game rules and pieces, are quite different in their own rights. The problem is compounded when there hundreds, or even thousands of combinations and studies to evaluate and compare.

These issues do not necessarily imply a level of sophistication when it comes to aesthetic assessment that is beyond computers; rather, they perhaps expose the limitations of humans in this regard. Aesthetics is something that cannot, or is very difficult to 'learn' or get 'good at' (even for humans). Hence traditional machine learning approaches may not be as effective as the approach used in this research. Experts are used mainly because we can be sure of their domain competence, but it is unlikely that a grandmaster will be any 'better' (e.g., more efficient, more objective) at evaluating beauty in the game than a master, or a master much better than a good club player. *Sufficient* competence in the game is therefore sufficient. Where, exactly, this line should be drawn when it comes to humans is difficult to say for certain.

The new model as a form of computing technology may be useful in aiding judges at composition tournaments (as an attentive tool for identifying beauty in studies), composers compose and players identify and harvest beautiful sequences from large databases too difficult to be explored manually. The technology may also be useful in making chess-playing

programs play in a more human-like way, and in the automatic annotation of chess games. The latter would enable computers to automatically select interesting games to be annotated and key moments in games to supplement them with automatically generated annotations.

Future work might include similar experiments pertaining to single moves and sequences taken from any point in the game, and also efforts toward assessing the beauty of *whole* games, even though a beautiful *game* is usually considered such because of a particular outstanding sequence within it, and not due to the beauty of every single move played. Applications of the aesthetics model at the 'knowledge level' to compose chess problems comparable to those of human experts and to understand better the 'creative process' in this domain would also be viable directions for further work [28].

## ACKNOWLEDGMENTS

We would like to thank FIDE composition judge and solver, Michael McDowell, for the examples and explanations of the three-movers in Fig. 6, Appendix A. We would also like to thank the official and unofficial reviewers of the many drafts of this article for their valuable comments and feedback.

## REFERENCES

[1] K. Muller, "The Clash of the Titans: Kramnik-FRITZ Bahrain," ICGA Journal, vol. 25, no. 4, pp. 233–238, 2002.

[2] F. H. Hsu, "Cracking Go," IEEE Spectrum, October, pp. 44–49, 2007.

[3] M. Boden, "Computer Models of Creativity," AI Magazine, vol. 30, no. 3, pp. 23-34, 2009.

[4] S. Bushinsky, "Deus Ex Machina – A Higher Creative Species in the Game of Chess," AI Magazine, vol. 30, no. 3, pp. 63-70, 2009.

[5] C. E. Shannon, "Programming a Computer for Playing Chess," Philosophical Magazine, vol. 41, no. 314, pp. 256-275, 1950.

[6] M. Newborn, Deep Blue: An Artificial Intelligence Milestone, Springer-Verlag, New York, Inc., USA, 2003

[7] P. E. Ross, "The Expert Mind," Scientific American, August, pp. 64-71, 2006.

[8] I. Sukhin, Chess Gems: 1000 Combinations You Should Know, Mongoose Press, Massachusetts, USA, 2007.

[9] M. Newborn, Kasparov versus Deep Blue: Computer Chess Comes of Age, Springer-Verlag, New York, Inc., 1997.

[10] R. Datta, D. Joshi, J. Li and J. Z. Wang, "Studying Aesthetics in Photographic Images Using a Computational Approach," in Proc. of the European Conference on Computer Vision, Part III, 2006, pp. 288-301.

[11] C. D. Cerosaletti and A. C. Loui, "Measuring the Perceived Aesthetic Quality of Photographic Images," Proc. of the First International Workshop on Quality of Multimedia Experience, 2009, pp. 47-52.

[12] R. Datta, J. Li and J. Z. Wang, "Algorithmic Inferencing of Aesthetics and Emotion in Natural Images: An Exposition," in Proc. of the 15th IEEE International Conference on Image Processing, 2009, pp. 105-108.

[13] D. Cope, Virtual Music: Computer Synthesis of Musical Style, MIT Press, Cambridge, Ma., USA, 2001.

[14] F. Friedel, "Ludwig - A Synthesis of Chess and Music," ChessBase News, 4th December. http://www.chessbase.com/newsprint.asp?newsid =3522, 2006.

[15] B. Manaris, P. Roos, M. Penousal, D. Krehbiel, L. Pellicoro and J. Romero, "A Corpus-Based Hybrid Approach to Music Analysis and Composition," in Proc. of the 22nd Conference on Artificial Intelligence (AAAI-07), 2007, pp. 839-845.

[16] S. Chinchalkar, "An Upper Bound for the Number of Reachable Positions," ICCA Journal, vol. 19, no. 3, pp. 181-183, 1996.

[17] M. A. M. Iqbal, A Discrete Computational Aesthetics Model for a Zero-sum Perfect Information Game, Ph.D. Thesis, Universiti Malaya, Kuala Lumpur, Malaysia, 2008. http://metalab.uniten.edu.my/~azlan/Research/pdfs/phd_thesis_azlan.pdf

[18] A. Iqbal, "Aesthetics in Mate-in-3 Combinations, Part I: Combinatorics and Weights," ICGA Journal vol. 33, no. 3, pp 140–148, 2010.

[19] A. Iqbal, "Aesthetics in Mate-in-3 Combinations, Part II: Normality," ICGA Journal, vol. 33, no. 4, pp. 202–211, 2010.

[20] S. Margulies, "Principles of Beauty," Psychological Reports, vol. 41, pp. 3–11, 1977.

[21] R. Houdart, Houdini Chess Engine v1.5a , 2011. http://www.cruxis.com/chess/houdini.htm

[22] D. Hooper and K. Whyld, The Oxford Companion To Chess, Oxford University Press, Oxford, 1996.

[23] S. Giddins, "Studies – It's the Thought that Counts," ChessBase News, 2nd March. http://www.chessbase.com/newsdetail.asp? newsid=7960, 2012.

[24] A. Iqbal and M. Yaacob, "Computational Assessment of Sparsity in Board Games," Proc. of the 13th International Conference on Computer Games: AI, Animation, Mobile, Educational and Serious Games (CGames 2008), 2008, pp. 29-33

[25] R. Pye, "An Enquiry into Excellence in Study Composition," EG, vol. 7, no. 117, pp. 638–642, 1995.

[26] J. Levitt and D. Friedgood, Secrets of Spectacular Chess, 2nd edition, Everyman Chess, London, England, 2008.

[27] G. Kasparov, My Great Predecessors, Parts 1-5. Everyman Chess, London, 2003-2006.

[28] A. Iqbal, Increasing Efficiency and Quality in the Automatic Composition of Three-Move Mate Problems, in Entertainment Computing - ICEC 2011, Lecture Notes in Computer Science, vol. 6972, pp. 186–197. Anacleto, J.; Fels, S.; Graham, N.; Kapralos, B.; Saif El-Nasr, M.; Stanley, K. (Eds.). 1st Edition., 2011, XVI. Springer.

**Azlan Iqbal** received the B.Sc. and M.Sc. degrees in computer science from Universiti Putra Malaysia (2000 and 2001, respectively) and the Ph.D. degree in computer science (artificial intelligence) from the University of Malaya in 2009. He has been with the College of Information Technology, Universiti Tenaga Nasional since 2002, where he is senior lecturer. He is a member of the IEEE and AAAI, and chief editor of the electronic Journal of Computer Science and Information Technology (eJCSIT). His research interests include computational aesthetics and computational creativity in games.

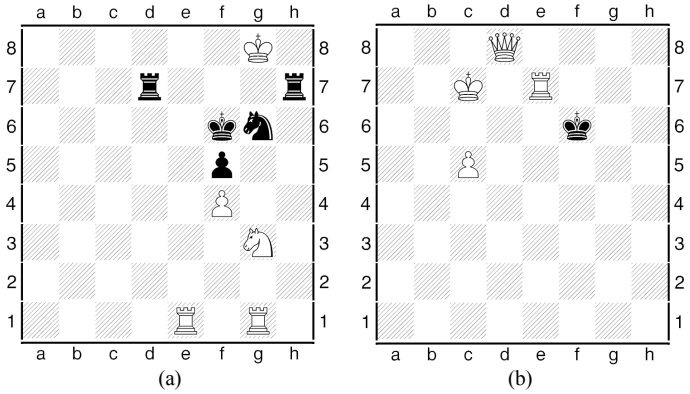
**Harold van der Heijden** (1960) finished his HBO-B (university of applied sciences) study in Biochemistry in 1981. He has been a research technician in a veterinary institute (GD, Animal Health Service) in the Netherlands since 1982. In 2009 he obtained a Ph.D. degree from the veterinary faculty of the Utrecht University. Since 2010 he is heading the research and development laboratory of GD. In the domain of endgame studies, he has been active as a collector (largest endgame study collection of the world), writer (three books), main editor of EBUR (1993-2006) the magazine of the Dutch/Flemish endgame circle ARVES, and main editor of the international magazine EG (since 2007). He obtained the title of international judge of endgame studies from FIDE in 2001, and organized and judged dozens of endgame study tourneys.

**Matej Guid** received his B.Sc. (2005) and Ph.D. (2010) degrees in computer science from the Faculty of Computer and Information Science at the University of Ljubljana, Slovenia. He is a researcher at the Artificial Intelligence Laboratory, University of Ljubljana. His research interests include heuristic search, computer game-playing, automated explanation and tutoring systems, and argument-based machine learning. Chess has been one of his favorite hobbies since childhood. He was also a junior champion of Slovenia a couple of times, and holds the title of FIDE master.

**Ali Makhmali** completed his B.Sc. degree in computer science in Universiti Tenaga Nasional, Malaysia (2009). He is now completing his M. Sc. (Artificial Intelligence) in the same university and for 18 months served as research assistant to Azlan Iqbal under the eScienceFund research grant (01-02-03-SF0188). His main tasks under the project were to program and test CHESTHETICA EG. His research interests include computational aesthetics in computer games, Web development, and computer programming.

APPENDIX A

To illustrate the concept of beauty in three-move mate problems, consider the two sequences shown in Fig. 5. In (a), we have an ancient composition [9] that is renowned for its beauty whereas in (b), we have a forced three-mover taken from a tournament game between two master players that is not quite as attractive.



(a) Al-Adli, *'Book of Chess'*, 9th Century
1. Nh5+ Rxh5 2. Rxg6+ Kxg6 3. Re6#

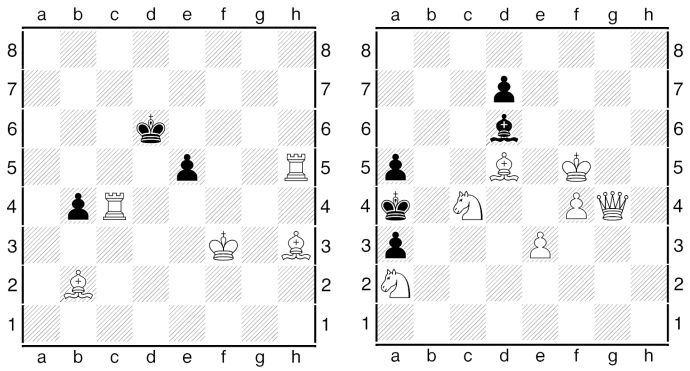
(b) Marcantoni vs. Gervais, FRA-chT2 North 0607, 2007
1. Qf8+ Kg5 2. Rg7+ Kh4 3. Qh8#

Figure 5. Beauty contrast in two three-movers.

The one in (a) is arguably more beautiful because, for example, the solution is not immediately visible or even expected. White seems to be one step away from being mated (Rh8#). However, he is able to secure the win by sacrificing a significant amount of material in a way that forces Black to entrap his own king; a highly satisfying win. The one in (b) is understandably less attractive because – having occurred in an actual tournament game – it does not have the benefit of creative forethought. The win is obvious and straightforward even to a borderline competent player.

The reasons we usually provide to justify our perceptions of beauty, however, need not necessarily all be mirrored computationally. For example, a computational model of aesthetics may arrive at generally the same aesthetic conclusions about a collection of three-movers or studies by having analyzed them in a way that may appear to us to have left certain 'essentials' out. This does not make the model inferior in any practical sense. It only demonstrates that there are alternative methods to ours in reaching what we consider an acceptable answer or solution.

Even so, there is a notable difference between the levels or types of aesthetic perception experienced by some humans with the patience and knowledge as compared to others. The two three-movers in Fig. 6 might illustrate this. In (a), the second move sets up the discovered check by sacrificing a rook. This move would probably not be a surprise to the experienced solver but might be quite impressive to players, especially if it occurred in a real game. The first or 'key' move does not follow the composition convention that there should not be any capture in that move. It also limits the king's movement, which is another violation of convention. In short, it is a strong move by White when the usual practice in composing is for the key move to be a 'quiet' or subtle one. Nevertheless, (a) may still appear as quite beautiful to the majority of sufficiently competent chess players and some composers.



(a) #3 by Earle Jenner, *The Guardian Chess Book*, 1967
1. Rxe5 b3 2. Rc3 Kxe5 3. Rd3#

(b) #3 by Erich Zepler, 1st Prize, Olympic Ty., 1936
1. Ke4! Bc5 2. Nd6 Bxd6 3. Qxd7#

Figure 6. Contrasting visual appeal and depth appeal.

In (b), ignoring for a moment White's key move, Black has the option of moving his king or bishop (the pawns cannot move). However, 1. … *Kb3* can be met by 2. *Qe2*, and Black cannot stop *Nb6#* or *Nb2#*. 1. … *Kb5* allows the king to escape to *a6*, unless in the key move White had moved his king to a white square (e.g., 1. *Kg6*) – to avoid 'check' by the black bishop – so as to allow 2. *Qxd7+ Ka6* and mate in the next move with 3. *Qb7#*. However, the black bishop's possible moves present a problem. White would like to guard *b3* and

[9] The rules of chess pertaining to the pieces involved in this composition have remained unchanged for the last millennium or so, hence the example and its solution are still applicable.

*b5* to allow for 3. *Qd1#* or 3. *Qxd7#*; after 1. *Kg6*, the bishop moves (e.g., 1. … *Bc5*), and 2. *Nd6* facilitates this plan.

Unfortunately, after 1. *Kg6*, instead of just moving the bishop away, Black has a subtle defense with 1. … *Bxf4!* after which the planned 2. *Nd6* gives stalemate (the black bishop is pinned against its king by the queen on *g4*). Now the composer's full idea emerges. The correct key move is 1. *Ke4!*, which neutralizes the effect of 1. … *Bxf4*. White prevents a stalemate potentially caused by his knight moving to *d6* after 1. … *Bxf4*. This may be considered by many composers to be a deep and clever idea, and a beautifully constructed problem, even though its main line and variations are not as 'visually' appealing as (a).

It is, however, not unlikely that *more* people with sufficient competence in the game would consider (a) more beautiful than (b) by virtue of its visual appeal (particularly the main line presented) and the speed by which it can be appreciated. It is therefore difficult to say, which, is 'truly' more beautiful because there are justifications for both. A computational aesthetics model should, however, seek to please the largest audience.[10] In terms of computer creativity in automatically composing problems, (b) would probably be more impressive and illustrative of machine intelligence – perhaps even 'awareness' – than (a) because there would be evidence of the capability of generating a deep and beautiful *idea* that transcends mere chance and (visual) aesthetic considerations.

## APPENDIX B

In comparing the performance of one aesthetics model in this domain to another like it, a reasonable determination of improvement could be made by analyzing the normality of the distribution of scores and correlation strength with human assessment. The details of this process are explained in [19].

The Spearman rank correlation coefficient (SRCC) was used to measure correlation with mean human-player aesthetic ratings that were derived from four online surveys comprising 80 randomly-selected three-move mate combinations. These included a total of 40 compositions and 40 sequences from tournament games between expert players that ended in mate. Further details can be found in [17-18]. The normality of the mean human ratings based on the Anderson-Darling test statistic (ADTS) was 0.351.

Table VII shows the performance of the original three-mover model that generated a fixed aesthetic score for a combination using a simple linear summation of all seventeen aesthetic features [17] as compared to an improvement that incorporated the use of some randomness or stochastic technology to sum the top 5 or 6 features in a line (while subtracting the scores of the remaining features that scored > 0), averaged over 20 iterations for a final aesthetic score [18-19]. The most recent implementation of the three-mover model that incorporates variations to the main line as explained in this article is also shown in the rightmost column. All of these tests were performed again after [18-19] given some improvements to the aesthetic feature detection algorithms and evaluation functions of the CHESTHETICA program used in [17-19]. Fig. 7 shows the main interface of a recent version of CHESTHETICA (incorporating the latest three-mover model) running on the Windows 7 operating system.

TABLE VII. PERFORMANCE OF THE THREE THREE-MOVER MODELS

| | **Original Three-Mover Model** | **Stochastic Three-Mover Model (No Variations)** | **Stochastic Three-Mover Model (Variations)** |
|---|---|---|---|
| **ADTS** | 1.951 | 0.746 | 0.550 |
| **SD** | - | 0.058 | 0.068 |
| **SRCC** | 0.663 | 0.660 | 0.648 |
| **SD** | - | 0.004 | 0.005 |

The ADTS values for the models other than the original one (that had fixed values) were averaged over 30 iterations because in those cases aesthetic evaluation for a combination may vary slightly from one cycle of evaluation to the next. The same is true for the SRCC. SD denotes 'standard deviation' for the item in the row immediately above it.

The stochastic three-mover model that did not implement variations – as described in [19] – scored a borderline average normality of 0.746. Anything higher would typically be suggestive of non-normality for its aesthetic score distribution. The SRCC, however, was quite good; averaging 0.66. The stochastic three-mover model featuring variations (explained in *this* article) had a better or more stable average ADTS of 0.55 and a comparable average SRCC of approximately 0.65. This is why it was considered an improvement over the previous two three-mover models and used also in the aesthetic evaluation of endgame studies.

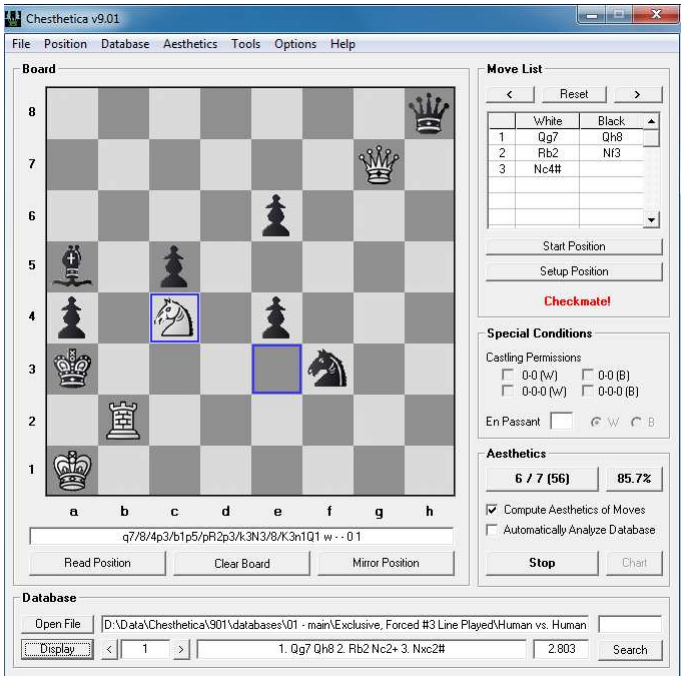



Figure 7. CHESTHETICA v9.01 main interface.

[10] The three-mover aesthetics model that incorporates the latest developments described in this article considers (a) to be of higher aesthetic value than (b).

In any case, the no variations model could be used to assess the aesthetics of three-movers that are *not* forced because the variations model assumes that the main line *is* forced (in order to generate alternate mating lines using an external engine). Alternatively, the variations model as incorporated into the new CHESTHETICA EG program could perhaps treat such a three-mover (that is not a forced sequence) as an endgame study. However, this requires further study and experimentation.

The 'top 5 or 6' feature determination was experimentally derived in [19] and confirmed again in this research. For instance, based on the stochastic three-mover model with no variations, the correlation when using just 5 features had a good average ADTS (0.384) but suffered in terms of the SRCC (0.582). An SRCC of ≥0.6 was considered a minimum for an aesthetics model such as this [18]. Using just 6 features, the opposite happened: the ADTS was 0.904 and the SRCC was 0.666. So it was reasonable to assume that a suitable compromise lay somewhere in between using a probability split (i.e., *n* percent chance of selecting the top 5 features and 100 - *n* percent chance of selecting the top 6 features).

Figs. 8 and 9 show the different probability splits tested based on the stochastic three-mover models not including and including variations, respectively. The vertical axis unit is shared by both the ADTS and SRCC. In Fig. 8, the *x* axis, the first number refers to the probability of the top 5 features being selected; and the second number, the top 6 features. Based on observations of the ADTS and SRCC values as plotted in charts like Fig. 8 (see also [19]), where the ADTS and SRCC values cross or come very close the ADTS usually still reflects normality and the SRCC is usually ≥0.6 or close to it; hence the model may be considered close to optimal performance, or at a point at which further considerations should be taken into account. The crossing point can be seen as a maximization of the correlation coefficient while retaining normality. Why, exactly, this happens at only one point or even happens at all given these variables is not yet known.

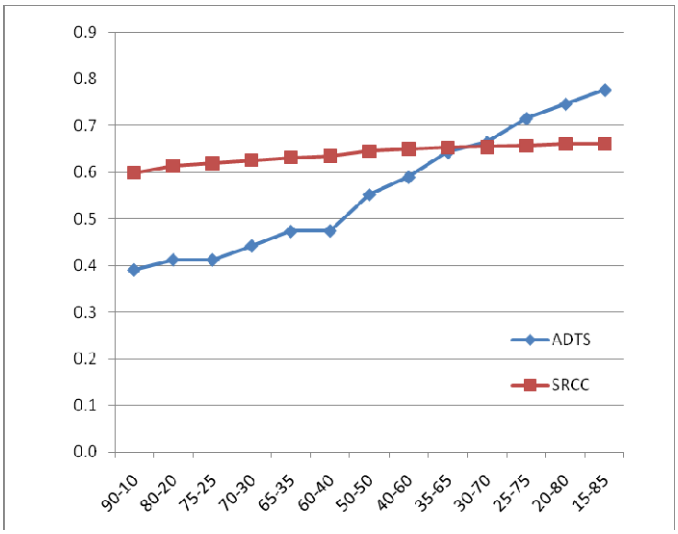



Figure 8. Performance analysis of different probability splits (top 5 or 6 features, respectively) using the stochastic three-mover model with no variations.

Here, it seems to occur at the 35-65 split; that is, a 35% chance of using the top 5 features and a 65% chance of using the top 6. Non-viability occurs where the ADTS exceeds 0.75 or so (15-85 split in Fig. 8) or the SRCC drops below 0.6 (90-10 split). Ideally, the ADTS should be as small as possible and the SRCC as high as possible even though they seem to correlate negatively with each other. The decision here was based on the highest correlation attained that still retained normality and this was using the 20-80 split (the actual ADTS and SRCC values are as in Table VII, center column). However, the average ADTS value was considered too high or 'unstable'.

The inclusion of randomly-selected variations was tested to see if the normality could be improved, i.e., pushed back from the borderline 0.746 normality test statistic without affecting the SRCC (0.660) too much. The variations were generated using the powerful 'ChestUCI' mate-solver engine[11] interfaced with CHESTHETICA. The results are shown in Fig. 9. The *x* axis represents the number of variations used (this includes the main line). Here, we can see that the SRCC is hardly affected given any number of variations used to calculate the aesthetic score. However, there were significant improvements to the average ADTS. The best improvement (that is, that minimizes it), is apparently somewhere between using 6 or 7 variations (5 or 6 in addition to the main line). Something similar can also be seen when using 9 or 10 variations but the ADTS incline (improvement) is less steep; in any case, using fewer variations is computationally more efficient and more likely to resemble what humans probably perceive.

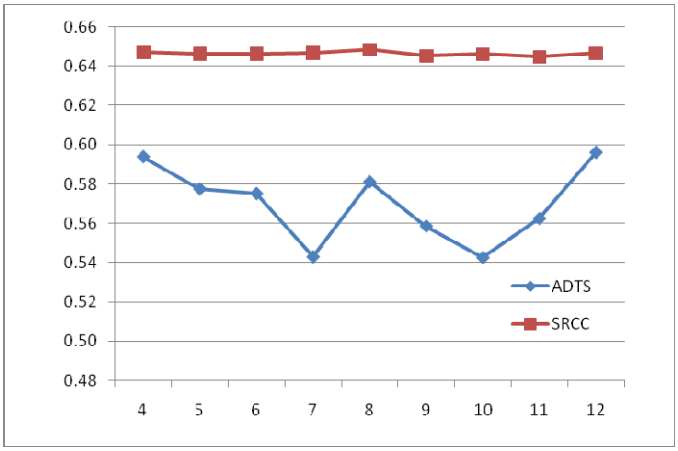



Figure 9. Performance analysis of including different numbers of randomly-selected variations (which includes the main line).

For symmetry with the feature selection process (i.e., top 5 or top 6 features), the same probability split of 20-80 was applied to variations in deciding whether or not 5 or 6 variations (in addition to the main line) should be randomly selected for evaluation. Based on the chart (Fig. 9) that distribution also appears to be more consistent with the incline or improvement in the ADTS than say, a 50-50 split might be. The higher probability should be associated with the '7' rather than the '6'. The results are as shown in Table VII, where the

[11] http://fhub.110mb.com/

ADTS is a stable 0.55 and the SRCC a good 0.65 (approximately).

Using exclusively 7 variations was also possible – no statistically significant differences between the ADTS and SRCC scores – but it would have been slightly more computationally intensive and provided less variability in the overall aesthetic assessment of a combination, i.e., the capability of the computer to 'change its mind' slightly from one cycle of evaluation to the next, as a human judge might if given a second look at a set of compositions to evaluate (especially if after a period of time).

Overall, incorporating variations into the aesthetic evaluations can be said to have essentially retained correlation strength while improving normality. This seemed the most reasonable and viable option for an improvement over the three-mover model that before this did not factor in variations. It would seem that the general advice of chess composers not to exclude them in the aesthetic assessment of combinations was indeed sound.