University at Buffalo
State University of New York

Department of Computer Science and Engineering

March 7, 2017

Dr. Francois Strydom
FIDE Ethics Commission

Dear Dr. Strydom:

This is a cover letter for two expanded reports on the cases of Mr. Ivan Tetimov and Mr. Angelo Ricciardi. I have prefaced each with the same general statement of my methodology. I have footnoted some references to previous interactions with you and Dr. Mark Watkins, but my testimony is complete without them. I have avoided footnoting links to information whose lookup is unambiguous and readily available, such as tournament records at chess-results.com and FIDE rating profiles.

Yours sincerely

Kenneth W. Regan

Report on the Performance of Mr. Ivan Tetimov at Benidorm A, Dec. 2014

General Procedure

The methodology is the same as that outlined in my January 2013 example report in the case of Mr. Borislav Ivanov, except for the addition of material on the "screening" test. I have updated and simplified the general description. Some very technical information from the 2013 report remains in force but is elided here—it can be referenced in that report. Any technical information that has been updated is referenced here.

The statistical model employed falls into the category of *predictive analytics*, a kind of modeling used for risk assessment, insurance/assurance, clinical care decisions, credit scoring, quality and performance assessment, personnel evaluation, and fraud detection in general. It does not try to predict individual chess moves, but rather aggregates over large enough samples of games of the numbers of moves satisfying certain properties, in terms of parameters representing a (non-cheating) player's skill. The parameters have been trained to reference points on the Elo rating scale by analyzing many thousands of games by players rated at those points in an intensive "Multi-PV" mode. This is supplemented by analysis to the same depths in the engines' quicker "Single-PV" playing mode of approaching a million games including those from every "GM"-level event in the history of chess. My papers on this work have been presented at one AAAI conference, a workshop associated to another AAAI conference, several conferences of the International Computer Games Association and the IEEE Computational Intelligence in Games, and several other data-mining and AI-related conferences.

Almost all tests in personnel evaluation and many in fraud detection produce *statistical z-scores.* They are used in preference to generic "p-values" when the quantities being estimated are known to conform to the so-called (Gaussian) *normal distribution.* This is broadly the case with quantities that are themselves averages or quantiles of observations. Personnel evaluation often makes heavy reliance on results of multiple-choice tests or psychological score questions. The z-score is a measure of the deviation of actual performance from expectations based on prior means of rating the human subject (such as when he/she was hired). Insurance companies use them to judge particular patterns of claims and hazards against prior experiences for possible signs of fraud or changing environmental circumstances, respectively. Thresholds for statistical z-scores were recently used in judging *when* to proclaim that the Higgs Boson and gravitational waves had been *discovered.*

My model employed in these tests is organized around one simple and natural principle:

> The likelihood of a move being played is a function of its value in relation to the values of other possible moves, and of parameters expressing the skill profile of the player.

The essence is that if a certain move has clear standout value over other moves, then it is yea-more likely that a good human player will find such a *singular* move, and hence agree with strong computer programs which certainly find it. Whereas, when several moves in a position have nearly-equal and optimal value, then if many players—of any skill—are given the position, those moves will have nearly equal frequency. Together these principles (and a modification to the latter that distinguishes whether some of the tied-equal moves had lower values at earlier stages in the search) have great explanatory power. For example, the cheating accusations by Veselin Topalov at the 2006 world championship match were stoked by observing his opponent Vladimir Kramnik's frequent agreement with the computer program Fritz 9 in particular—indeed my own tests show over 90% agreement over the second half of the second game which was a gut-wrenching loss for Topalov. However, most of the 32 moves concerned were completely forced and several other positions had multiple moves of equal-top value. My mature model gives any player of Kramnik's class a *baseline projection* of over 80% agreement on those positions. Its average projection for top players is about 58% agreement over the broad normal run of chess positions, but the positions Kramnik faced had far more singular moves than this norm.

The model is also minimalist and "vanilla": there are no elements specific to any aspect of chess apart from the move values provided by strong chess programs which are used as a jury. All other chess content and all human relevance are derived from how the model is *trained* on large sets of games played by humans of all skill levels as measured by Elo ratings. No parameters are tailored to any one particular player based on "profiling" or other considerations. This avoids bias and implements an aspect of "judgment by your peers"—if your rating is R then your baseline projection comes from observed behavior by other players rated near R. (As discussed briefly at the end, it also enables a defense wherein a player might argue that his/her style is singular enough to mandate departure from the generic parameter settings.)

Exactly what the model does is take the move values and player-parameters and generate projected probabilities for every legal move in every position. Simply summing the projections $p_{1,t}$ for the engine's first move m_1 at every game turn t gives the baseline projection for agreement with these moves (MM for "move-match"). Summing $p_{i,t}$ times the loss in value of move m_i compared with m_1 gives the projected error at turn t, and this can be summed or averaged over t (ASD, for average scaled difference). A key simplifying assumption is that the distributions over choices m_i of $p_{i,t}$ are *independent* between different game turns t. Using this assumption maps chess into the domain of rolling dice: if m_1 is the move Nf4 (knight to f4) and $p_{i,t} = 0.375$ then it is as if we paint "Nf4" on 375 of the faces of a 1,000-sided die. Rolling many-sided dice is a form of "Bernoulli trials" whose statistical theory is so simple as to be often taught in high school. The theory provides not only baseline projections but also their standard deviations which combine to define the z-scores of actual observations.

Strictly speaking, the key assumption is false—decisions on consecutive game turns are often not independent. A witness is that Anand's half of the famous "double blunder" in the 2014 match against Carlsen may owe to his fixation on counterplay with a pawn that he pushed three times in four turns over this phase. However, the dependence is evidently of a "nearest-neighbor" kind that is *sparse* amid all pairs of moves, so that the departure from independence is felt mainly as a reduction in the sample size—that is, as if there were slightly fewer separate decisions represented by those moves. This in turn necessitates *adjustment* of z-scores to reflect the lower effective sample size and other possible sources of modeling error. A multiplicative adjustment a is determined empirically by running 10,000s of trials on sets of (say) 9 games each by (presumably) non-cheating players and seeing that the distribution of z/a values over each trial stays within the confines of normal distribution on the positive side.

There don't exist 90,000 games by players of a given Elo rating level R (within $R \pm 10$ say) in any one year or sequence of a few recent years, so the trials are done by *resampling* the training data for each milepost Elo level. The point is to ensure that the rate of "false positives" does not exceed a predictable *normal* level. These runs were done first in summer 2011, and were repeated on larger data in summer 2014 with no change in the previous adjustment factors of $a = 1.15$ for the MM tests and $a = 1.4$ for the ASD test, plus 1.2 for a test of playing moves of equal value (EV) to the engine's first move. These adjustment factors are snug for ratings in the range 2300–2500 and become progressively more conservative away from there. (Further rationale for this kind of adjustment has been discussed previously in this company.)

It is also important that all use of statistical regression in the generation of z-scores has been done on the *large data* of the training sets, not from the *small data* of one player's tested games. Regression on the small data is done only for a separate task of skill assessment, which produces what I call an "Intrinsic Performance Rating" (IPR) matched to the Elo rating scale used by FIDE. The IPR is not part of the statistical judgment but can serve to explain it in terms more familiar to players and the public at large.

To summarize, the model is of a generically recognized kind and is trained by standard data-gathering and regression procedures. The statistical ingredients are the simplest possible: more sophisticated notions such as maximum-likelihood estimation, distributional distances, and analysis of variance are absent.

Specific Procedure

Aspects of my procedure that depend on the current state of my work use first-person pronouns, while others are general. Upon receiving a set of games to test, and players' ratings plus any other information, the procedure is as follows:

1. Run a chess program to obtain sufficiently deep evaluations of all available moves in every relevant position. I use the program's so-called Multi-PV mode for this purpose, setting it to evaluate up to 50 legal moves with Rybka 3 and 32 (the limit) with Houdini 3. The analysis data is output by scripts confirming to the UCI chess program protocol either through the Arena chess GUI (for Windows-only engines) or directly in UNIX/Linux environments.
2. Run further scripts on the analysis data to extract the particular information needed for the statistical tests.

3. I exclude moves 1–8 of any game before analyzing it. After the engine analysis is obtained, my programs exclude moves in repeating sequences and positions where the engine judges one side ahead by more than 3.00 (colloquially, three Pawns). I also exclude "book moves" after move 8, using the novelty-finding feature of [16], unless superseded by published statements, private communications, or my own databases. Other move exclusions may be made based on supplementary information.
4. Run a statistical program to generate projections, measure deviations, and compute confidence intervals for several aggregate statistics, following the theory and equations of [9,10]. The program runs tests with parameter settings representing a player's rating. I use the rating before the tournament, the rating after, and the tournament performance rating (TPR), giving most weight to the rating after. Although the parameter space is multi-dimensional and infinite, for uniformity and continuity I select the closest setting published in [9], or average two settings, to reach a value close to and preferably above the rating.
5. The main tests are *move-matching percentage* (MM), *equal-top value* moves (EV or ETV), and *scaled difference* (SD, or ASD when averaged over all analyzed positions). The program computes projections and **z-scores** according to the well-known theory of independent Bernoulli trials (technically, multinomial trials) and Gaussian normal distribution.
6. An adjustment is made to allow for move decisions not being truly independent, and for other possible sources of systematic modeling error. The resulting **adjusted z-scores** are final outputs used to indicate statistical (un-)likelihood.
7. I then run the last tests "in reverse"—actually, in the original training mode—to generate the parameters that would give them zero deviation. These yield the player's IPR for the given set of games. This has no official standing but serves to explain interpretations to the chess community.
8. As a separate test, I run the same games in the so-called Single-PV mode used for maximum performance by engines in timed competitions. This is for purpose of scientific control: I have been able to run about 200,000 games from almost every important competition in the history of chess (excepting national leagues) in this mode, including many from recent major Open tournaments. The large data set, currently giving over 34,000 player-performances of at least 120 analyzed moves each (including book moves after move 8), provides context and comparison for outlier results.
9. Informed by both tests, render a report giving interpretation of the results.

Settings For This Case

Ivan Tetimov's FIDE rating is listed as 2158 in the report of the 2014 Benidorm-A tournament. The *chess-results* entry for him shows a rating gain of +77.8 points over the games that were played. Under the "null hypothesis" of fair play, those points are credited to him as the official best estimate of his skill after the tournament. Hence I entered 2236 as the rating in my tests. With built-in slack, my analyzer used settings (sensitivity $s = 0.12176$, consistency $c = 0.4837$) corresponding to 2260 in the central fit of my settings-to-Elo regression on large data. The Benidorm A time control of 90 min. + 30s/move (lacking a 30 minute increment at move 41) is among the fastest of standard controls used in my model's training data, which is a slight conservative factor (that is, tending away from false positives).

Using the Chessbase Big 2017 data disc, I have double-checked the following determination of novelties under a "book by 2300+ players" standard which was used in my original tests in early 2015:

1. Tetimov—Cucarella Montell: 11.d4xc5
2. Garcia Martin—Tetimov: 9.Nf3
3. Tetimov—Ibanez Aullana: 7.d5 (turns 7 and 8 skipped anyway)
4. Jimenez Martinez—Tetimov: 15.Rfd1 (in other game, GM Shulman played 15.Ne5)
5. Tetimov—Cano Gimenez: 9...Qc8 (so move 10 is the first by Tetimov to be tested)
6. Platel—Tetimov: 10.a4
7. Olhovik—Tetimov: 14...Qb6
8. Tetimov—Kultiyasov: 14.fxe5
9. Tejedor Fuente—Tetimov: 11.Qd2

Mr. Tetimov was disqualified in his round-10 game before enough moves had been played to test. Turns before the above-listed moves were eliminated, as well as turns in repeating sequences and turns where one side was ahead more than 3.00. The game with Tejedor Fuente went to turn 90. Subsequent to the 2016 World Cadets in Batumi, I have adopted a standard policy of also eliminating turns 71 and later. Here the difference does not matter because Rybka 3 and the other engines report advantage beyond -9.00 to Black over all of these moves. Rybka 3 gives a sample size of 288 moves by Mr. Tetimov; the other engines differ slightly owing to some positions nipping under or over the 3.00 cutoff with different engines.

Official Results With Rybka 3

With settings entered for 2236 (giving 2260), the model projects:

1. Agreement with Rybka 3's first line (move-match, MM) on 146.5 turns (50.86%) with a pre-adjustment 2-sigma range 130.8–162.1 (45.42%–56.29%).
2. Equal-top-value (EV) moves (in positions where two or more moves are tied for optimal) on 169.1 turns (58.73%) with 2-sigma range 153.8–184.5 (53.41%–64.05%).
3. Scaled difference (SD, i.e., total error) of 24.92 pawns-in-equal-positions with 2-sigma range 19.57–30.27, giving an average scaled difference (ASD) of 0.0865 with 2-sigma range: 0.0679–0.1051.

The observed test results are:

1. MM: agreement on 181 turns (62.85%), giving an adjusted z-score of $4 \cdot (181 - 146.5)/1.15 \cdot (162.1 - 130.8) = 3.833$. (My program gives 3.836 using non-rounded values of the range endpoints.)
2. EV: 208 turns (72.22%) giving $z = 4.23$ after adjustment by 1.2 in the denominator.
3. SD: 7.27 (ASD: 0.0253), giving $z = 4.71$.

In my original report I used a policy of reporting the maximum of the three tests as the official result. This was considered a benefit of doubt from a combined-score policy that would often produce a higher score than any test, but in this case the rule adopted from a 2011 paper by Dmitri Zaykin (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3135688/) produces a lower score. It uses the conservative stipulations that MM and EV are completely correlated and ASD correlates 0.75 with the other tests. Then Zaykin's version of the Fisher-Stouffer rule says to add the three scores and divide by $\sqrt{8}$. This gives a final combined test score of **4.517**. The interpretations from my earlier report remain valid, starting from how far this exceeds the threshold of **2.75** for statistical support of allegations *in the presence of physical or observational evidence* given in the FIDE ACC regulations.

The interpretation and rationale in the present instance is as follows: Under standard normal distribution, a performance equaling or exceeding this deviation can be expected once every 318,884 entries (by one player in one tournament). To convert this into a statistical expression of unlikelihood of the null hypothesis of fair play, one must consider the prior likelihood of cheating. On general grounds, a rounded estimate of 1-in-10,000 for this "prior" can be justified (as conservative) by comparing the total entries in tournaments at comparable levels under FIDE auspices per year and the frequency of actions against players in a given year. This would translate into about 1-in-32 odds, clearly short of "comfortable satisfaction" and in fact below the common 97.5% one-sided confidence level associated with a two-sigma deviation.

However, the presence of other evidence cuts into the 9,999-in-10,000 presumption of innocence. Even if it were judged that only 1-of-10 persons giving the observed physical feedback would be using it to cheat, the odds adjustment would still be about 32,000-to-1 against the null hypothesis. This is over a 99.99% confidence standard. A more-balanced judgment might set the prior at 1-in-2, "as likely as not" given the physical evidence.

My "Intrinsic Performance Rating" (IPR) measure does not give statistical test scores. It does, however, indicate whether the measured quality of play is in ranges commonly associated with computer engines as opposed to human players. I measure an IPR on non-book moves (IPRNB) of 3105 +- 105 (2-sigma error bars adjusted by 1.20, everything rounded to the nearest 05). Apart from the context of cheating I include book moves after the first 8 turns—judging opening choice and home prep as skill factors from then on—in the IPR figures which I circulate after prominent matches and tournaments. Then out of 310 moves his IPR is 3110 +- 105.

Results With Other Engines

The ACC procedures call for a test with a second engine by a second party. It has not been clarified whether the second test must be given equal official standing to the first or is to be regarded as a sanity check and source of benefit-of-doubt to the accused. If the former, then it should be combined by another application of the Fisher-Stouffer rule taking into account the measured (high) correlation of the tests; at 100% correlation the scores would be averaged but otherwise a higher aggregate score might emerge. I have declared my understanding to be the latter in communications with ACC. Whereas Rybka 3 is a constant “standard candle” for the official test—and a conservative choice since current engines have grown apart from it—the other engines are tested in current versions. Short of recreating the entire training set and model for each version of each engine, a conversion of their evaluations via a standard logistic curve (the same curve used for Elo ratings but with different scale factors for each engine) is employed. The observed conversion scale factors vary between player Elo levels, so conservative catchall factors are employed for each engine used. The office of a second party has not been filled; this is tied to larger issues about the mounting of a server under FIDE auspices to house and manage my software and the departure of a person who had the required expertise. In compensation thereof, I have expanded my suite of backup tests from 2 to 4, using newer versions of the champion Komodo and Stockfish engines in addition to Houdini 4 from the time of my original reports. In several instances besides this one, the newer tests have provided significant benefit of doubt for the subjects of my reports.

I report the secondary tests in a shorter form, giving just the observed values and the adjusted z-scores, from which the model’s projections can be back-calculated. The rating and book settings are the same as above. The 2236 entry settings print out as 2270 with Houdini 4, 2272 with Komodo 10.2, and 2284 with Stockfish 7.

- Houdini 4: (MM $180/287 = 62.72\%$, EV 203, ASD 0.0290) $4.13, 4.15, 4.46 \rightarrow 4.50$; IPRNB 3070 +- 130
- Komodo 10.2: (MM $186/285 = 65.26\%$, EV 204, ASD 0.0319) $4.85, 4.33, 4.07 \rightarrow 4.69$; IPRNB 3030 +- 175
- Stockfish 7: (MM $162/292 = 55.48\%$, EV 190, ASD 0.0394) $2.33, 2.70, 3.46 \rightarrow 3.00$; IPRNB 2900 +- 155

The great divergence with Stockfish 7 is notable for its *absence* in the other report for the proceedings in Madrid and for quite the opposite phenomenon in another case. The 3.00 aggregate score is still a positive by the ACC guidelines in the presence of other evidence. It corresponds to a standard frequency of 1-in-741, which on division by 2 leaves 99.7% confidence. Results from screening tests show that the divergence carries forward to the current Stockfish 8 as well.

Screening Results and Comparison With Historical Record

The screening tests use the standard 1-PV playing mode of the engines rather than the Multi-PV mode needed to compile rigorous values for all reasonable moves. This enables processing a typical game in 5–10 minutes on one core thread as opposed to 4–8 hours in the highest Multi-PV mode. Beginning with 2015 I have employed my university supercluster to screen every game in tournaments covered by "The Week in Chess" plus others brought to my attention, using current versions of Komodo and Stockfish which are the only two top engines that are natively scriptable in a UNIX/Linux batch environment. In 2015 I recorded over 31,000 player-performances with Stockfish 6 and Komodo 9.2, and in 2016 over 33,900 with Stockfish 7 and Komodo 9.3. So far in 2017 I have screened over 5,540 performances with Stockfish 8 and Komodo 10.3.

The screening tests do not provide theoretical confidence intervals for z-scores and other statistical judgments. Instead they tally the MM and ASD figures without regard to the context of game situations which require Multi-PV mode to analyze with authority. (EV cannot be judged directly in Single-PV mode; a non-drop in eval at the next game turn is used as a stand-in.) These two figures are synthesized along with the player's rating into a "Raw Outlier Index" (ROI) on a 0–100 scale. The scale is designed so that 50 is the expectation for one's rating at standard time controls, 40–60 is the "completely normal" range (nominally 2-sigma radius but currently set "aggressively" to around 1.5-sigma), 60–70 the "yellow" range and above 70 is "orange." Conversion factors between engines and engine versions are used to center these scales. As "screening" this test is intended to have high sensitivity but low specificity—put in simple terms, the intent is to identify a moderate subset of the lambs so that if there is a wolf-in-sheep's-clothing then the wolf will almost certainly be among them. This is to help arbiters to allocate resources of watching some player(s) more carefully but discreetly, and also as a quick filter for whether any complains that may arise are well-founded or ill-founded.

I ran the entire Benidorm A and B tournaments with the current engines for 2017. The ROI for Tetimov was 73 with Stockfish 8 but 87 with Komodo 10.3. The 87 is higher than any ROI figure recorded since 2015. The other defendant in Madrid was tied for top in 2015 at 83 with Stockfish and alone at 82 with Komodo, these being the only scores over 79.

These figures provide immediate comparison and contrast to the mass of other players, from Carlsen down through children rated under 1500 in FIDE junior events—all equally graded in proportion to rating on the ROI scale. There are no statistical theories or manipulations here, just a simple mass tally of results. To be sure, there will always be *someone* at the top of such a table, and conditioned on this, the full test with Multi-PV data and z scores will plausibly find a large deviation (but, often for players in the low or mid 70s it does *not*). However, the presence of physical and behavioral evidence in the case changes the reckoning: The screening tests show the accused are not being unfairly "picked on" by the z-scores in relation to other players—rather, they already stand well apart from them.

The IPRs over Elo 3000 can also be compared with the historical record. I have logged only a few IPRs over 3000 by the world's top players. Bobby Fischer and Garry Kasparov both hit the mid/high 2900s on occasion, as did Capablanca at New York 1927. As with the screening test, the IPR results do not provide confidence tests—the error bars are only ones of measurement by regression over small data—but they refute any move to dismiss the full statistical results on *sui-generis* grounds. They also quantify the benefit obtained by the alleged use of engine(s) in comparison to they player's Elo rating.

I can provide upon secure request the game analysis files from which the above conclusions and comparisons are drawn; their locations are already visible to some co-workers and confidants.

Conclusions

The results provide extremely strong support of the other evidence in this case. They do fall short of thresholds that had been proposed for statistical judgment alone, and it must be repeated that no re-introduction of such a proposal is envisaged in the near term. Defenses can be raised by arguing that more moves than those indicated were home preparation and so should be struck, and/or that the player has such a distinctive tendency toward positional or tactical play compared to peers that a non-central pairing of the s, c parameters to Elo should be used. The effects of these are likely to be minor. There is also the anomalous result from Stockfish 7. This might lead to discussions about the recent history of chess engines that leave the grounds of solid science; staying on the ground we note the consistency of the results from the other engines.