University at Buffalo
State University of New York

Department of Computer Science and Engineering

January 13, 2013

Emil Sutovsky, President
Bartłomiej Macieja, General Secretary
Association of Chess Professionals

Dear Emil, Bartek, and ACP Board Members:

I am writing out of general concern at the increased spectre of cheating at chess, and from particular concern about issues of due process. It is prompted by current reports [1,2,3] and public discussion of the allegations against Mr. Borislav Ivanov from the 2012 Zadar Open, which was held last Dec. 16–22. In these reports, and in posts by other bloggers and commenters [4,5,6,7], the main evidence comprises tests of moves with computer chess programs, and assertions about the frequency of the played move 'matching' the program's first choice (or first few choices), alongside ideas about unlikelihood of rating-performance fluctuations.

None of these tests so far presented meets scientific needs of regular methodology, reproducibility, bases for comparison, or confidence intervals for conclusions. However, the need for such tests is recognized by many, no less the esteemed Leonard Barden, in his comments to the ChessVibes story [1] which appear immediately afterward at the top. Barden called for a test of Mr. Ivanov's nine games at the 2012 Zadar Open with "Houdini or another top program." I have worked on meeting these needs since October 2006, in co-operation with others including Guy Haworth of Reading University (UK) [8,9,10]. My appended report shows the workings and significance of my procedure. There has been wide concern about due process [11].

I pose two questions, of which at least the first should be an immediate concern of ACP in conjunction with FIDE and national organizations. The second is a deeper issue that I believe needs consultation with experts in statistics and computer sciences, and with representatives of bodies in other fields that have established protocols for using evidentiary statistics in fraud detection and arbitration.

1. What procedures should be instituted for carrying out statistical tests for cheating with computers at chess and for disseminating their results? Under whose jurisdiction should they be maintained?
2. How should the results of such tests be valued? Under what conditions can they be regarded as primary evidence? What standards should there be for informing different stages of both investigative and judicial processes?

Regarding the second question, my policy to date [12] has been that statistical evidence is secondary to physical or observational evidence of cheating when judging a case. This leaves open the question of using statistical evidence for recommending whether to pursue an investigation, in which ACP could play a role analogous to a "grand jury" in US jurisprudence.

In proper contexts, statistical results give important information. For one, especially when supplemented by scientific-control data showing natural incidence of deviations, they can inform calculations of potential costs, benefits, and damages. Speaking more generally, many walks of life are informed by projections from models, sometimes differing ones from many models, and the receivers need to have their own rules for evaluating them. The point of approaching ACP is to determine how the contexts and rules should be set for chess. The goals, shared by Haworth and others I have discussed this with, include:

(a) To deter prospective cheaters by reducing expectations of being able to get away with it.

(b) To test accurately such cases as arise, whether in supporting or primary role, as part of uniform procedures recognized by all as fair.

(c) To educate the playing public about the incidence of deviations that arise by chance, and their dependence on factors such as the forcing or non-forcing quality of their games.

(d) To achieve transparency and reduce the frequency of improper accusations.

(e) Finally, hopefully to avert the need for measures, more extreme than commonly recognized ones, that would tangibly detract from the enjoyment of our game by players and sponsors and fans alike.

I believe my question 1. should be a near-term action item, on which ACP can lead. I have already been asked for action in private; in two cases where my tests showed no statistical support, my results and opinion were implemented as a negative verdict determining the awarding of a delayed prize. This makes it a *de-facto* professional matter, also regarding the propriety of my acting as arbiter of recourse without at least a formal designation (including sufficient indemnification). For question 2. I wish to see a conference or panel of experts from chess and relevant other fields, and empower it to recommend procedures for evidence and arbitration in these cases.

I note that ACP has already begun action in response to the present controversy, including your communications released via Facebook. I trust that my information and requests are in harmony with the goals of your organization.

Yours sincerely,

Kenneth W. Regan

Reference Links

[1] www.chessvibes.com/reports/bulgarian-chess-player-strip-searched-after-suspection-of-cheating
[2] www.chessbase.com/newsdetail.asp?newsid=8751
[3] www.chess.com/news/suspected-cheater-strip-searched-4830 (no comment permalinks)
[4] www.youtube.com/watch?feature=player_embedded&v=JrOJ8SPENjM
[5] www.chessvibes.com/comment/84267#comment-84267
[6] User "doktor_nee" on comment page 3 of [3].
[7] schaken-brabo.blogspot.be/2013/01/vals-spelen.html
[8] www.cse.buffalo.edu/~regan/chess/fidelity/
[9] www.cse.buffalo.edu/~regan/papers/pdf/ReHa11c.pdf
[10] www.cse.buffalo.edu/~regan/papers/pdf/RMH11b.pdf
[11] www.chessbase.com/newsdetail.asp?newsid=8760
[12] www.cse.buffalo.edu/~regan/chess/fidelity/Golfers.html

Appendix: Report of Tests of the 2012 Zadar Open

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, and fraud detection quite 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 thousands of games by players rated at those points. Besides the peer-reviewed scientific papers [9,10] presented at the major international conferences of the AAAI and ICGA, my website has a two-page prose description and a six-page overview [13,14] with some technical details. Particular points relevant to this matter are:

1. As basis for comparison and scientific control, I have run almost every top-level event in the entire history of chess (except for national leagues) through Rybka 3 to reported depth 13. This stands at over 34,000 player-performances, including over 5,000 from recent Open tournaments comparable to the Zadar Open.
2. My statistical-analyzing program allows inputting parameters corresponding to any Elo rating, and generates confidence intervals for its hypothesis tests, which have been verified on practical data.
3. The process of taking and analyzing data is automated, with no ad-hoc judgment about when/whether a particular move is a "match," using chess engine settings that reproduce.
4. The "Intrinsic Performance Rating" (IPR) feature of [10] is independent of the move-matching tests, but quantifies the amount of benefit gained in the event of cheating.

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 Arena chess GUI [15] controls these engines and outputs their analysis and search statistics to text files; it is distinguished by the quality of its output.
2. Run scripts on the text files to extract the 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) and *average error* (AE). 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. Dependencies between games are reduced by the exclusion of book moves, and ones between moves in a game can be understood as meeting conditions called "nearest-neighbor" and "sparse" dependence, which mitigate their effect. The main effect is to reduce the effective sample size below the raw number of moves, which correspondingly widens the error bars when expressed as percentages. My empirical tests on tens of thousands of 9-game performances simulated from the games used to train the model indicate that a widening of 15% for the MM test is appropriate (and generally conservative), and indicate 40% for the AE test (which is subject to more sources of systematic 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.

Supplementary technical remarks—can be skipped on first reading

No warrant is made of parity between conditions for the Multi-PV and Single-PV tests; indeed the programs themselves follow different algorithms between the modes. Differences of 2–3% in results for individual players are typical, and owing to a selection effect, high Single-PV matching figures noticed while screening tournaments usually give rise to lower Multi-PV figures. However, for aggregates of at least a few dozen player-performances the distributions of results are similar enough to take observations in the large Single-PV game set about actual frequencies of outliers, and use them to interpret Multi-PV results in a bulk manner. (At about 6 processor-core hours per game in Multi-PV mode, it is not yet practical to have so much Multi-PV data for comparison, and the Single-PV test also replicates the competition mode of the programs.)

It should be emphasized that the model's equations already distinguish positions that have "obvious" or "forced" moves from ones with many reasonable possibilities. Forced moves are "automatic matches" to computers but are projected with close to 100% choice probability already when generating expectations and z-scores. The model and program provide for possibly down-weighting such moves within the sample, and up-weighting moves according to measures of "complexity"—and further experimentation may discover a uniquely natural and advantageous way to do so. My procedure in cheating testing has thus far used equal weights on all moves.

In the present case there has been much talk of "top-3" tests, which means counting a match whenever the played move is among the engine's top three moves. This seems to entail running in Multi-PV mode, and has the problem that baseline projections are higher, leaving less room for deviations. The motive seems to be that a cheater might try to cross up the MM test by playing second- or third-best moves. I regard the AE test as better motivated here, with promise also of less dependence on which particular strong engine such a player uses and/or is tested with.

My work has used Rybka 3 as standard engine since January 2009, with both Multi-PV and Single-PV analysis run to reported depth 13. My own tests in engine-to-engine matches in both modes show rough parity to depths 15–19 on other major engines (generally 19 for Stockfish and 17 for Houdini versions), and there is also rough parity in the time required to analyze a game to the respective depths. Rybka 4.1 is somewhat speedier to depth 13, but in head-to-head fixed-depth games has also had weaker results, so I have stayed with Rybka 3 despite a "stalling" problem (which I reported a week after its release, and can now reproduce from a cleared-hash start) that sometimes forces manual intervention. A second issue of unexpected preservation of hash between games in analysis runs can extend the disruption of reproducibility from intervention. Absent that problem I find that reproduction of analysis output is assured by sameness of settings in three windows provided by the Arena GUI: (1) engine parameters—including the necessity of selecting 1-thread (plus "deterministic" if provided), (2) "Options for all UCI engines" under "Engines–Manage" where hash-table size and tablebase access are set, and (3) the "Automatic Analysis" dialog. In particular, Houdini 3 Pro x64 has given identical output to the Standard w32 version, even reported node-counts though of course not time-dependent data, with 1 thread and 512MB hash. The Multi-PV tests have all used the same 64-bit PC with 512MB hash and no tablebases; the Single-PV Rybka data has been compiled on various machines, mostly with 256MB hash for each of four simultaneous runs and 3-4 piece tablebases only. Both tests run forward preserving hash between moves of a game; with Houdini 3 I am testing the idea of running the Multi-PV test clearing hash between moves, using an EPD file of positions.

I use fixed depth rather than time limits not only for reproducibility, but also to assure as much equality of conditions as possible at different points of a run, without worrying about processor load variation. The fixed depth may shortchange endgame positions. In the same breath it must be noted that time information for individual moves in standard tournaments is often unavailable, and the model makes no expectation of having it. This is a major difference from cheating detection at on-line sites where such information is recorded and employed in a more "forensic" than "statistical" manner. The cheating process and timing described in Mr. Cyril Marzolo's published confession [17] in the Sébastien Feller case exemplifies the practicality of achieving (much) higher depths, and I have seen blog comments recommending to try "depths 24–25" with Houdini, while [6] used "depths 12–22/45 sec." However, the depths used suffice to preserve most moves chosen by the programs, and their strength estimated in the 2700s is authoritative enough for a 2300 player certainly. Recently published tests by Matej Guid and Ivan Bratko [18] used Rybka 3 to depth only 10 and certain other engines to depth 12. Single-PV runs to depths 19 and 21 for Houdini and Stockfish did not show appreciable difference in actual move-match percentage.

Settings For This Case

Borislav Ivanov's FIDE rating was 2227 entering the Zadar Open, and became 2342 after his event-specific performance rating of 2697. My tests used the "s_{fit}" and "c_{fit}" parameters published in [9] for 2200, 2300, and 2700, which now correspond to 2233, 2339, and 2709 after the revised regression procedure of [10] and other minor changes over two years. I used them for continuity with my tests in other cases and because they closely track Ivanov's three rating measures. I note that while most of the game data used to train the model come from round-robin and small-Swiss tournaments with time controls approximating 40/120 + 20/60 + G/30 or 40/90 + G/30 + 30-seconds-per-move increments, the Zadar Open was played at the appreciably faster pace of G/90 + the same increments [2]. This difference makes all the parameter settings "generous" to the player, hence less likely to yield false positives.

I determined the novelties in the respective rounds [added 11 April, 2014: this was using the standard of "book by 2300+ players"] as 1–15.Bh4, 2–9...Be7, 3–10...b6, 4–15.h4, 5–10.Bc4, 6–9.Qc1, 7–10.g3, 8–8.Nf3, and 9–11.h4, discarding all moves before them. For round 4, using ChessBase Big Database 2013 augmented by TWIC [19] corrected 14...h6 given by [20], which had missed a prominent game owing to a transposition of earlier moves. Despite conflicting reports about extra security measures taken in Rounds 8 and/or 9, my primary tests included both games. The sample size is 258 moves. All analysis was scripted as described above using the logging feature of the Arena chess GUI, running Rybka 3 to reported depth 13 and Houdini 3 to depth 17, using 512MB hash, no tablebases, and neither "own book" nor "common book."

In the training set drawn from the years 2006–2009, players with Elo rating between 2190 and 2210 matched Rybka 3's top move 48.3%, those with ratings between 2290 and 2310 matched 50.4%, and those between 2690 and 2710 matched 56.3%. The higher corresponding projections reported below in Mr. Ivanov's games mean that they had somewhat higher differences in value between the engine's first and second moves—one could say they were more "tactical." The projection for, say, 2300 means it is consistent with the actual play of 2300 players in the many training games that they would have matched 53.2% in the positions Ivanov faced. In this manner, the procedure respects the principles of judgment-by-one's-peers, consideration of particulars, and choices that give benefit-of-doubt.

Results

The Multi-PV test observed move-matching of 69.6% overall, 69.0% after excluding book moves. With the settings for 2300 (2339), on the included positions with Ivanov to move, my analyzer projects 53.2% matching, with margin of error 47.4%–59.0% before adjustment, 46.5%–59.9% after. The value $2 * (69.0 - 53.2)/(59.9 - 53.2) = 4.72$ is the adjusted z-score for the test. For 2200 (2233) the projection is 51.9%, giving a z-score of 5.86 which adjusts to 5.09.

With the 2700-player settings, the projection moves up to 58.5%, with error bars 52.9%–64.2% that widen to 52.0%–65.1%. The actual figure is still outside, and the adjusted z-score is 3.21.

The total error over 9 games judged by Rybka 3 is 795 centipawns, call it 7.95 pawns, while the projected error for a 2300 player on the positions faced by Mr. Ivanov is 21.48 pawns. After 40% widening, the adjusted z-score for the AE test is 3.73. The IPR for the nine games is 3061 on non-book moves, 3089 on all moves. It is subject to the same widening as the AE test. The resulting 95% confidence intervals for the IPR have 2815 and 2868 as their respective floors. Ivanov's opponents played to an aggregate IPR of 2476 for the nine games, with error bars 2260–2692 before the widening, and 2173–2778 after it. Their average rating of 2572 is within both ranges.

When the Round 8 and/or Round 9 games are excluded, and/or a Move-70 cutoff is used for long games, the adjusted z-scores are higher and go beyond 5 on some tests under the 2300 (2339) settings. The corresponding IPRs for Ivanov range between 3126 and 3258, several with confidence-interval floors above 3000.

When the tests were re-run using Multi-PV analysis by Houdini 3 to fixed depth 17, the Multi-PV test observed (only) 64.7% matching, 64.1% on non-book moves. For the same 2300-player setting the Houdini-based projection is 52.2%, giving an adjusted z-score of 3.51. For matches to the top 3 moves of Houdini 3 the test shows 91.6% against a projection of 80.7% for 2300 players, 85.9% for 2700 players. The actual figure is lower than the '98%' claimed for Houdini 2 by FM Valeri Lilov [4]. It should be noted that the model has not yet been calibrated for Houdini 3, so the Houdini MM results are currently provisional at best. These results are offered only for comparison to the Rybka results and to other reported tests with Houdini, and to honor the letter of Leonard Barden's request.

Some side observations: Non-matches to Houdini were observed in moves 10,12,14,15 of the round-5 game against GM Robert Zelcić, moves 10, all of 15–24, and 29,32 of the round-7 game against GM Andrey Sumets, and moves 9–15 and 18,23,25,29,30 of the last-round game against GM Ivan Sarić, with most of the difference from Rybka occurring on these moves. (This shows that it is possible to obtain significant results even when tests log ten consecutive non-matches in a game; that others are on early moves also argues against the defense of "home preparation" or that the above exclusion of book is conservative to the detriment of Mr. Ivanov.) In almost all tests with Rybka, the pre-adjustment z-score for the AE test was very close to that of the MM test, but for Houdini the differences were greater. This supports my belief that the AE test can be sharpened by upgrading the model and improving the determination of how much quantified error should result from a given 'blunder,' and work to do this is current.

Interpretation and Comparison With Single-PV Data and Historical Record

Under standard interpretations typified by [21], z-scores correspond to assertions of odds against the designated **null hypothesis**, which in this instance can be called "no cheating" for short. More properly they designate the odds of getting a reading of the same or higher deviation on the same side, given that the null hypothesis were actually true. The term **p-score** is used with general distributions, while z-score is specific to projections of expectations under normal distribution, as applicable here.

A z-score of 4.72 represents odds of 1-in-848,103. For 5.10 this is 1-in-5,582,934. For 3.21, from the test for 2700-level players, the odds are 1-in-1,507. The z-score of 3.73 for the AE test with the 2300s setting means 1-in-10,445 mathematically. There is a civil convention that a z-score of 2.00 is the minimum requirement to argue significance; any lesser deviations are "within the margin of error" and ascribed by default to chance.

Such odds figures, especially the lower ones, must however be interpreted in relation to the population they come from, including possible biases from their manner of selection. Most crudely put, suppose we were to test 1,507 performances by 2700-level players known not to be cheating, and press a red button upon seeing a z-score of 3.21 or higher. We can expect to press it once, and hence be wrong once. This does not violate the mathematical meaning of "confidence," because we were right not to press the button the other 1,506 times. It remains correct to say that the odds against a randomly-selected 2700-player hitting 3.20 are 1,506-to-1 against, and that the odds against a 2300 player (or any player) hitting 4.72 are almost a million-to-one against, but the word 'a' can be deceiving.

The interpretation becomes difficult when only a small number of items are tested. What caused us to select them for testing? If the cause comes from preliminary indications of the same kind—which in this case means the mere fact of beating strong players as well doing a quick check of games with an engine—then the bias in selection can upset the interpretation of probabilities. This is why my policy [12] has been to require an independent factor that determines the selection, such as behavioral or physical evidence of cheating. (It should be footnoted that the same policy can allow excluding at least the round-8 game, whose prior circumstances were different.)

The population-based caveats, however, have their own limits. Have there been 1,500 performances in comparable events by 2700-rated players in all of chess history? There certainly have not been 850,000 recorded 9-game performances by 2300+ players—databases representing the entire recorded history of chess number under ten million games. Determining how far the caveats apply requires looking at the historical record, and this is where the Single-PV data set comes in. For reasons of practicality and uniformity, "book" (whatever it was at the time) is not excluded from this set.

With Rybka 3, the Single-PV test showed 68.8% matching. Albeit lower than the Multi-PV score (inlcuding book), this still stands in 6th place among 5,352 performances in Opens, including large selections from every Gibraltar and World Open (since 2006), and the whole of every Aeroflot A, Lone Pine, and European Individual Championship (except 2009–2011 are selections). Significantly, no higher percentage has been found since last March when there were about 2,300 performances at press time for [22], and Feller remains the only one above 70%. On the global list of chess performances, 68.8% is tied with Viswanathan Anand and Vinay Bhat for 29th all-time, sandwiched between Lubomir Kavalek at the 1984 Olympiad and Garry Kasparov in his 1985 match with Ulf Andersson. The next highest figure by a non-GM over at least 120 moves, apart from IM Diwakar Prasad Singh who was acquitted after formal investigation in 2007 [23], is 67.7% (66th highest). When WGMs are excluded too, the highest is 66.7% (116th).

The IPR figure of 3089 is determined directly from the quality of the moves, rather than the results of games, and is completely independent of the opponents' ratings and performance. By comparison, since January 2010, Vladimir Kramnik has two IPRs above 3000, Levon Aronian one (3002 in the match against Kramnik), and Magnus Carlsen none [24]. IPRs at the recent London Classic ranged from 2455 for Anand to 2991 for Kramnik, with similar error bars. The interpretation of all this is that Mr. Ivanov's IPR figure is unusual even for a top player, and represents a quality of moves significantly beyond the level normally required for a 2697 performance. The IPRs above 3100 that result from certain reasonable exclusions are commensurate with the published ratings of Houdini 3 and several other top programs [25].

The comparisons also provide a sanity-check on the shorter odds, which is one function of a scientific control group. If one ascribes odds of 1,500–1 when using settings for a 2700-player at an Open tournament, it is important to be able to point to 1,500 performances by players in the range (say) 2600–2800 that didn't reach the level achieved by the one. This supports the overall *bona fides* of the setup that is making these projections. The second question posed in the cover letter revolves around whether this grounding in reality is sufficient to extend the significance accorded to the results that produce assertions of longer odds.

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

The real-world comparisons do give some additional ways to be skeptical of the model's conclusions, and to these I now turn.

Evaluation and Further Steps

One one three main skeptical points I consider is to note that several move-match percentages (to Rybka 3 in Single-PV mode) higher than Mr. Ivanov's have been achieved in Open tournaments by players not accused of cheating. One, GM Le Quang Liem at the 2012 Aeroflot Open, even had a *minus* score in the tournament. The model bases its projections, however, not only on a player's rating but also the nature of the particular positions he/she faced. These other instances have higher baseline projections, and yield smaller z-scores, ones that are fully in line with expectations when you have several thousand data points. While stated odds of a million to one fall under such considerations in cases like DNA testing over large populations [26], I believe their basis does not apply at this scale in the world of chess. The furthest I see this line being taken is to claim that the fact of Liem and the others means that it is prejudicial to assign settings of "2300" (2339), which produce those odds, to Mr. Ivanov. However, the standards of due process requested in the cover letter must settle on a policy commensurate with the player's actual rating. A study of intrinsic rating-lag in developing young players might determine how far it can reasonably be bent.

The second defense is to note that players rated under 2300 do have performances above 2697 a fair portion of the time. The IPR figures indicate, however, that the intrinsic quality of the moves is significantly beyond what is normally required to achieve a 2700-level performance. They are closer to the published ratings of many top programs, and above recent results by top players and improving young ones I have studied thus far.

The third critique is to argue that the model's own error bars are out-of-true by more than the 15% adjustment, that they are subject to systematic dependencies (besides those considered above) that obviate the amount of independence assumed in applying z-tables of normal distribution. Here it is noteworthy that I myself have co-authored a blog article [27] advancing exactly this critique of the projected error bars for discoveries at CERN, prompted by several noted instances of "three-sigma events disappearing," i.e., failing to reproduce. My response is that the chess application apparently has more ability to generate simulations from composite data than physicists have to simulate unseen particles. This has been done only to test z-scores up to 4.00, however, representing odds of about 30,000-to-1.

Much more data is needed to create a bank of thousands of independent rather than "composite" player-performances of 9 (or more) games each, while the 6-hour average to analyze one game in Multi-PV mode to sufficient depth on a single CPU core limits even the ability to do a same-day test of one player during a tournament. Thus I am also asking for volunteers with spare CPU time (desktops better than laptops) to help with compilation, using different engines toward our goal of creating an "engine panel" to serve as ultimate authority for training the parameters. Near-term, further calibration work needs to be done with Houdini. Then a major upgrade will combine results from different search depths, so as to distinguish "trappy" moves and evaluate the amount of "challenge" a player creates for his/her opponent. The reward of all this will not only be to make statistical cheating-testing more reliable, but also to improve skill assessment, player training, and even the realism of strength-level settings in playing programs, as described in our papers and presentations.

Conclusions

The bottom line of the test is that the results are about as strong as one can reasonably expect a statistical move-matching test, done scientifically and neutrally and with respect for due process, to produce. My model projects that for a 2300 player to achieve the high computer correspondence shown in the nine tested games, the odds against are almost a million-to-one. The control data and bases for comparison, which are wholly factual, show several respects in which the performance is exceptional even for a 2700-player, and virtually unprecedented for an untitled player. The z-scores I am reporting are higher than in any other instance I have formally tested, which is what prompts me to raise the questions in the cover letter now.

Postscript: In response to my (accidentally anonymous) request as a comment to [4] for evidence besides move-matching stated in his video, FM Lilov replied, "The evidence shown in the video is pretty clear. First of all, we have 98% of engine moves (Houdini 2.0c to be exact) as the other 2% are the opening moves and second, we can see this player's rating chart. It's impossible to lose or/and drew 1900 players with a very poor play and suddenly destroy 2600 rated GMs in 20 moves with absolute engine match moves. In this case, I believe physical evidence is not required." Whether the policy expressed by his last sentence, apart from any matter of fact, should be adopted—and under what circumstances—is the "deeper question" in a nutshell.

Reference Links (some items cited only in the cover letter)

[1] www.chessvibes.com/reports/bulgarian-chess-player-strip-searched-after-suspection-of-cheating
[2] www.chessbase.com/newsdetail.asp?newsid=8751
[3] www.chess.com/news/suspected-cheater-strip-searched-4830 (no comment permalinks)
[4] www.youtube.com/watch?feature=player_embedded&v=Jr0J8SPENjM
[5] www.chessvibes.com/comment/84267#comment-84267
[6] User "doktor_nee" on comment page 3 of [3].
[7] schaken-brabo.blogspot.be/2013/01/vals-spelen.html
[8] www.cse.buffalo.edu/~regan/chess/fidelity/
[9] www.cse.buffalo.edu/~regan/papers/pdf/ReHa11c.pdf
[10] www.cse.buffalo.edu/~regan/papers/pdf/RMH11b.pdf
[11] www.chessbase.com/newsdetail.asp?newsid=8760
[12] www.cse.buffalo.edu/~regan/chess/fidelity/Golfers.html
[13] www.cse.buffalo.edu/~regan/chess/ChessResearchProspectus.pdf
[14] www.cse.buffalo.edu/~regan/chess/ChessResearchOverview.pdf
[15] www.playwitharena.com
[16] www.chessgames.com
[17] www.chessbase.com/newsprint.asp?newsid=8370
[18] www.chessbase.com/newsdetail.asp?newsid=7621
[19] www.theweekinchess.com
[20] www.chessgames.com/perl/chessplayer?pid=138575
[21] www.fourmilab.ch/rpkp/experiments/analysis/zCalc.html
[22] www.nytimes.com/interactive/2012/03/19/science/chess-players-whose-moves-most-matched-computers.html?ref=science
[23] www.chessbase.com/newsdetail.asp?newsid=3724
[24] www.cse.buffalo.edu/~regan/chess/fidelity/data/CarlsenAronianKramnikCompare.txt
[25] www.computerchess.org.uk/ccrl/4040/rating_list_all.html
[26] en.wikipedia.org/wiki/Prosecutor's_fallacy
[27] rjlipton.wordpress.com/2011/12/13/the-higgs-confidence-game/