University at Buffalo
State University of New York

Department of Computer Science and Engineering

June 8, 2013

Herrn Bernd Rosen, Chair
Prof. Dr. Bruno Müller-Clostermann, Vice-Chair
Herrn Ulrich Geilmann, Team Captain
Sportfreunde Katernberg SK
SGZ Zeche Helene
Twentmannstr. 125
D-45326 Essen, Germany

Dear SK Katernberg Officials:

Enclosed is a full report on two sets of accused games played by Mr. Jens Kotainy of your club, using the statistical model of my own sole development. It is important to note that the report is incomplete in two respects: three relevant games from the 2013 Neckar Open were unavailable for analysis, and the case forces consideration of individual games while my extensive comparison data has been processed only for whole-event performances.

Nevertheless the procedure used in prior cases has been followed to completion with the available data, and some definite conclusions are given for your consideration. It remains my understanding that there is no physical or observational evidence in this case, while the tests and conclusions presented herein cover-and-subsume all assertions about move-matching evidence that have come to my attention. As with the case of Mr. Borislav Ivanov from Bulgaria, this case therefore represents new ground, and I have devoted much space to conditions on how to evaluate such statistical-only results. I do not envy you, and as I have written privately to Prof. Dr. Müller-Clostermann, I encourage you to consult with outside parties, or otherwise widen the scope.

Report of Tests on the 2012-13 SBL and 2013 Neckar Open

My methodology and procedures were the same as detailed in my open letter and report from 13.i.2013 in the case of Mr. Borislav Ivanov at the Zadar Open, and unchanged since ii.2011 with my involvement in the case of Mr. Sebastien Feller. I refer to them also for terminology. I omitted a small update to make the handling of proto-repetitions in Rybka consistent with Houdini's evaluation policy, so the numbers are in fact the same as what I reported first on Arthur Kogan's Facebook group on 12th April.

More important, I have not implemented any part of the in-progress conversion of my platform from Rybka 3 to Houdini 3 for these tests. Time has not permitted this, but there is also a point of methodology that takes precedence over the fact that my own tests give higher move-match rates with Houdini 3 than Rybka 3. My model generates projections and error bars on which the statistical conclusions are based. The projections are generated by analysis with Rybka 3. In principle they apply to any sequence of moves, including bad moves, and in particular to the sequences of moves favored by various versions of Houdini. However, the unlikelihood assertions from those projections have been tested empirically only with respect to the sequence preferred by Rybka 3 itself. My tests stay with this, so that in particular the "actual" figure used for the seven accused Bundesliga games is the 68.9% Multi-PV figure reported for Rybka, not the 76.4% obtained in the main Single-PV test with Houdini 3. Nor is it correct to infer that the deviation from the projection of 59.7% would be enlarged by a factor of (76.4 - 59.7)/(68.9 - 59.7) = 1.82, because Houdini 3 has several peculiar properties which have not been fully tested.

The model has never had the forensic intent to discover "which engine?", for one thing because it lacks timing and comparison information that is run and recorded in real time on chess servers for online play. Instead it takes the experimental-phenomenology approach, with Rybka 3 as the apparatus. The idea is that if a systematic disturbance is caused with any engine, it will have an impact on measurements made with Rybka 3, for which I have compiled large-scale trial results and scientific-control data. Thus although Houdini versions have been exclusively cited in discussion of this case, the tests' sole use of Rybka 3 should not be considered strange. Nor is ignoring the current Rybka 4.1 version; my engine-match tests at fixed-depth rather than fixed-time have all favored Rybka 3 as stronger.

The regressions are run with a target precision of 0.0001, and for various other reasons, the values are significant only to three figures. This applies also to generated figures for "Intrinsic Performance Ratings" (IPRs), which should really be rounded to the nearest 10, or at least the nearest 05 as was done with professional FIDE ratings until recently. However, for accuracy in reporting I give all four figures for the IPR.

Settings For This Case

Mr. Jens Kotainy's FIDE rating stayed at 2413 on the monthly rating lists from August 2012 to April 2013, then increased to 2459 on the May 2013 and current June 2013 lists. With 18 games shown for the May change and 0 since September, I infer that the change is based entirely on the 9+9 games from the two events addressed in this report, the Bundesliga weeks since October 2012 having been treated as one event. My policy is to use the post-event rating, and take parameters based on the "s_{fit}" and "c_{fit}" columns in the table on page 9 of my AAAI 2011 paper with Guy Haworth.

neighborhood of Elo 2460. Owing to the 2400-level and 2500-level training sets having been above the fitting line, this now corresponds to Elo 2512 in the numerical formula defined and tested in later papers. I also used and tested the values $s = 0.100$, $c = 0.492$ given then for 2400 which now correspond to Elo 2462, and which I used and reported in GM Arthur Kogan's Facebook group in April while feeling they better reflected the span of what I knew about his rating. The two sets of parameter values also convey the extent to which deviation conclusions vary with regard to the player-strength setting.

(With the upgraded regression procedure of my paper with Haworth and Macieja later in 2011, the "2400" training set has an aggregate IPR of 2435 while the "2500" set has 2528; the other sets are all within 11 Elo of their designated century markers and mostly below them, so these are the only two data points that poke above the regression line. Whether this has anything to do with an alleged problem in the FIDE rating system around 2400 is a subject for future work; in any event this report is taking precautions against it.)

The time controls of 40/100 + G/50 with 30-second increment for the SBL and 40/120 + G/30 (no increment) for the Neckar Open are close enough to the average standard of the training sets to need no adjustment for their difference. The opponents' Elo ratings play no role in my tests, except possibly for the selection of games.

My involvement began in the week of April 8th with a private communication including a PGN file with 13 games represented to me as accused, 7 from the SBL from 8.xii.2012 to 7.iv.2013, and 6 from the Neckar Open which was played over the long Easter weekend, 28.iii–1.iv, 2013. I discovered two other Bundesliga games played by Mr. Kotainy in November 2012, but erroneously believed the Neckar Open set to be entire, rather than rounds 3–8 of a 9-round event played over 5 days. This raises an issue of *selection* from both events. Willfully excluding 2 or 3 games from a 9-game sample can invalidate the statistical conclusions, but they can remain if there are grounds for selection that are (close-to-)uniquely determined by relevant factors wholly apart from the correspondence of moves to machines. The six Neckar Open games are the ones played on high enough boards to be recorded automatically and transmitted live, are the only ones of Kotainy given by TWIC and the tournament website, and as of this writing are still the only ones available to me. It helps in avoiding appearance of selection bias that they include two losses by Kotainy, and are the six played against titled players (all GMs). I am considering to adopt a standard of allowing to exclude games against opponents rated 200 or more Elo points lower; this would exclude the games in Rounds 1–2 and would almost exclude the final-round game against Jinshi Bai then rated 2236 (but currently 2400). For the Bundesliga I present my results three ways: *including* the two early non-accused games to make 9, staying with the accused set of 7, and including only the 5 games played since late February.

I determined the novelties using a standard of "book by 2300+ players" drawing on the ChessBase Big Database 2013, which extends through mid-November 2012, augmented by TWIC since Issue 941. For the available (and accused) games from rounds 3–8 of the Neckar Open, I determined the novelties to be

- Round 3, White versus L. Nisipeanu, 16...Nd7;
- Round 4, Black versus K. Piorun, 17...g5 by Kotainy (included in the sample);
- Round 5, White versus A. Istratescu, 11.Qd1 by Kotainy;
- Round 6, Black versus E. Bacrot, 9...a6 by Kotainy;

- Round 7, White versus A. Wirig, 13...Nd5 by Wirig;
- Round 8, Black versus Y. Solodovnichenko, 13...Qc7 by Kotainy.

For the Bundesliga, numbering the games 1–9 rather than by SBL round 1..15, it turned out not to be necessary to consult TWIC or check the dates:

1. White versus T. Heinemann, 12...Rb8 by Heinemann;
2. Black versus A. Jochens, 10...Bg4 by Kotainy;
3. White versus A. Cioara, 9...Qc7 by Cioara;
4. White versus L. Nisipeanu, 8...Rc8 by Nisipeanu;
5. Black versus R. Janssen, 10...a4 by Kotainy;
6. White versus G. Fish, 17...Rc8 by Fish;
7. White versus M. Feygin, 11.Nce2 by Kotainy;
8. Black versus F. Holzke, 11...Ncxe5 by Kotainy;
9. White versus J. Smeets, 11...e6 by Smeets.

In the training sets drawn from the years 2006–2009, players with Elo rating between 2390 and 2410 matched Rybka 3's top move 51.8%, and those with ratings between 2490 and 2510 matched 53.1%. Higher projections reported below in Mr. Kotainy'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 projections mean that it is consistent with the actual play of similarly-rated players that they would have the projected figures in the positions that Mr. Kotainy faced.

Numerical Results

As detailed in the referenced January report, the confidence intervals projected internally by my statistical analyzer are widened by a factor of 1.15 for the move-matching (MM) test, and by 1.4 for the average-error (AE) test. Thus the original z-scores are divided by 1.15 and by 1.4 in the respective tests to make *adjusted z-scores*, which are then compared with the standard 2.00 criterion for civil significance. A statement of odds based on the actual value of the adjusted z-score follows, using standard calculations for normal distribution (e.g., http://www.fourmilab.ch/rpkp/experiments/analysis/zCalc.html). These are commonly regarded as the odds against a *null hypothesis* which here would stand for "no cheating," but actually denote the expected frequency of such deviations from a large population that observes this hypothesis.

Neckar Open

The Multi-PV test observed move-matching of 59.9%, 169 of 282 moves after excluding book, repetitions, and positions with an advantage over 3.00. This is unusually many moves for 6 games, reflecting that the games were long. With the settings for old-2457, current-2512, the "baseline" projection 50.5%, which is lower than the norm from either the 2400 or 2500 training set. The lower baseline reflects the presence of more positions with many reasonable choices, for instance typical of endgames that have not yet reached a critical point. The "normal" interval for those moves is 45.1%–55.9%, giving the actual 59.9% a z-score of 3.47 which adjusts to 3.02. The AE test observes total scaled error of 10.52 when 18.70 was expected, "normal" two-sigma range 14.4–23.0, z-score of 3.84 adjusting down to 2.74.

With the settings for old-2400, current-2462, the MM baseline is revised downward to 49.9%, 2-sigma interval 44.4%–55.3%, z-score of 3.70 adjusted down to 3.22. The AE test expects total error 19.60, range 15.20–24.00, actual makes $z = 4.13$ which adjusts down to 2.95.

The Intrinsic Performance Rating (IPR) for these games is 2983, with range 2841–3125 widened to 2785–3181. The IPR does not have any statistical significance for unlikelihood assertions; the ranges represent error-of-measurement not hypothesis testing. The IPR includes book moves after the first eight turns, so 309 total moves in this case, and does not depend on the parameter settings for the player.

Schachbundesliga

For the 7 accused games, which comprises 212 moves after book and the other exclusions, the MM baseline for current-2512 is 60.4%. This is markedly higher than for the Neckar Open games—indeed it is higher than Mr. Kotainy's actual MM% for those games—reflecting that these games were more tactical. The observed matching is 68.9%. The normal 2-sigma range is 54.3%–66.5%, giving a z-score of 2.77 adjusting down to 2.41. The AE test sees total error of 5.95 when 15.33 is expected. This looks like a huge difference in proportion, but the 2-sigma range is 10.58–20.08, so the z-score is 3.95 which adjusts down to a less-huge-looking 2.82. (A "human" way of understanding this is that we *can* save a big block of error just by avoiding blunders for many moves; the model understands the "blocky" nature of error when it sets a wide normal interval. The difficulty of assigning values to blunders also factors into the larger 1.4 adjustment.)

With settings for current-2462, the MM baseline is revised down to 59.7%, normal range 53.6%–65.9%, so the z-score is 2.98 adjusting down to 2.59. The AE test expects 16.17 total error, range

onward, is 3018, with widened range 2817–3218.

When all 9 SBL games are included, 56 non-book moves are added to make a sample of size 268. The two extra November 2012 games by themselves have a low actual MM of $24/56 = 42.9\%$, and interestingly also an extremely low baseline of exactly the same for current-2512. The overall baseline for the 268 moves falls to 56.7% with normal range 51.3%–62.2%, with the actual of 63.4% lying "just" outside. The z-score is 2.45, adjusted to 2.13. The AE test expects 18.11 total error with range 13.19–23.03, and the actual of 9.89 gives $z = 3.34$ adjusted to 2.39. Note that these 56 moves have total error 3.94, almost 4.00, compared to 5.95 for the other 212 non-book moves. For current-2462 the projection goes down (only) to 56.1% for MM and up to 19.09 for error, giving adjusted z of 2.33 for MM and 2.58 for error. The IPR restoring book goes down to 2894 with 1.4-widened range 2668–3119.

When just the 5 SBL games played since late February are included, the sample size falls to 103 moves. The actual matching is $77/103 = 74.8\%$, but also the baseline rises to 61.9% for current-2512, 61.2% for current-2462. Because of the smaller sample size the normal ranges are much wider: 53.3%–70.5% for 2512, 52.6%–69.9% for 2462. The adjusted z-scores for MM are 2.60 for current-2512 and 2.72 for current-2462. For the AE test, even though the actual total error of 2.74 is only about a third of the projections of 7.70 for current-2462 and 7.29 for current-2512, the adjusted z-scores are a flat 2.00 for the former and only 1.91 for the latter, which by my policy is deemed *insignificant*. The IPR is 3036, again with wider expanded range 2705–3366. (I have not heard any reason alleged to include only these games, nothing saying 'cheating only in 2013'—I have included this paragraph in order to illustrate the effect of sample size on my error bars, and hint that perhaps the higher actual match rate was noticed by onlookers especially during the final SBL weekend in April. My report makes no further reference to this grouping.)

Interpretation

I currently hold it correct to take the maximum of the MM and AE tests in any one set of moves. The two tests are heavily correlated, but 'error' represents a fairly independent criterion from matching just the first-listed move, so one could argue to combine them into a figure slightly higher than the maximum. However, this is offset somewhat by the selection bias of choosing from two tests, though this bias decreases with their correlation. The tests supplement each other in that AE lessens the dependence on a particular engine (and also provides against the often-expressed albeit-tendentious "top-3 match" idea that a live accomplice seeing several close moves might communicate the 2nd-best or 3rd-best etc.), while the MM test gets at the major point. Using the settings for old-2460/current-2512 and the sets of games actually accused, the results are:

- Neckar Open: $z = 3.02$, giving frequency 1-in-791.
- Bundesliga: $z = 2.82$, giving frequency 1-in-416.

With the settings for old-2400/current-2462, the results are:

- Neckar Open: $z = 3.22$, giving frequency 1-in-1,560.
- Bundesliga: $z = 2.98$, giving frequency 1-in-694.

This is the first formal report I have given on two events with finishing dates within a week

of each other, and where there is no physical or observational evidence from either event. Some aspects of this need to be kept in mind:

1. First, if there were just *one* such event, even with 1,500-1 odds, and again with no independent evidence, *the result should be ignored*. In any given week of TWIC one can find games from about 1,000 players, so by "Littlewood's Law" among them one should expect to find about a 1,000-1 deviation. One might argue that there are well fewer than 1,000 *winners* of tournaments, when restricting attention to winners, but note here that Mr. Kotainy finished in the pack of the Neckar Open, and that the SBL is a team event. (On the other-other hand, one could restrict to makers of GM norms.)
2. Second, if there were independent evidence, then that removes the selection factor of the first point, and having even 400-1 odds becomes strongly significant. Indeed the civil significance threshold is only 40-1 odds (one-sided), even if the outside evidence does nothing more than pin down the selection, let alone be separate evidence of guilt or responsibility.
3. Then again, the damage to reputation lasting even beyond the timespan of recently-given sanctions argues for an odds threshold approaching those used in penal courts. This is represented by many commenters in many chess-cheating threads comparing to "DNA results" which colloquially involve odds of "a million to one" or higher.
4. If the two events are regarded as independent so that the odds are multiplied, they are over 300,000-1 for current-2512, and over that 'million-to-one' for 2462. Even if we insist on using 2512 and the *lesser*-significant test for both events, giving 2.74 (AE for NO) and 2.41 (MM for SBL-7), this is still over 40,000-1, which would take a year of trawling TWIC to expect by chance.
5. When the 13 accused games are tested as one pile, the MM and AE tests give adjusted z-scores that are nearly equal, 3.86 and 3.94 respectively. The 3.86 represents odds of 1-in-17,369. For current-2462 they are again nearly equal, 4.13 and 4.19, and the 4.13 represents 55,000-1 odds. The combined odds should be expected to be slightly less in either case—by analogy, the odds against getting 8 heads in 10 coinflips are lower than demanding to get 4 heads in each of two events of 5 flips each.

At this point I am not able to give an authoritative opinion on how these five considerations should be weighed in general. The overriding purpose of my letter and freelance-report in January on the Ivanov case was that the chess world needs to judge these factors at large and reach standards that can be commonly agreed, and a commission charged with this and more-immediate issues of prevention and policing has begun work only this week. However, in this case I think even the least of the test assessments is cause for *concern*, and also sheds the following light:

(a) First, the results are strongly distinct from cases such as the voiced accusations against GMs Shengelia and Kreisl at the recently-concluded EICC in Legnica, Poland, which similar tests show to be baseless. Likewise also from results of much-talked-about "mercurial" results by similarly-rated players, of which I have collected a few at

`http://www.cse.buffalo.edu/~regan/chess/fidelity/data/Mercurial.txt`

(b) The assertions published in the magazine *Schach* and in BBS comments that certain moves are "computerlike" can be criticized individually, doubting some assertions even in their own chess-specific terms, but the results of my tests—which are purposely entirely chess-neutral

(c) The reaction by players and online viewers during the Neckar Open was spontaneous and strong, even leading to actions taken by the tournament directors, long before I came on the scene. Again the test results are enough to say that the players' actions have grounds, and that this is a problem with greater dimension than the opinions of some players.

Historical Comparison and Particular Factors

In the Ivanov report I have described my database of over 35,000 whole-event performances analyzed with Rybka 3 in Single-PV mode. I have also run all tournaments of category 20 and higher, all major matches, and a few entire Swiss events, in the Multi-PV mode used by my full model to do tests and compute IPR's. The only human whole-event IPR's over 3000 that I have recorded apart from credible cheating accusations are three by Kramnik and one each by Anand, Aronian, Kasparov, and (Igor) Khenkin. It is possible I will find more as I have not yet itemized the IPR for every player-performance, but for instance the only other recent performance[1] with over 70% matching to Rybka or Houdini, by Shirov at Shanghai 2010, gives 2929. Only for sake of illustration, here is the short summary table from the just-completed category-21 Thessaloniki Grand Prix tournament, enough to show that the IPR is not simply a function of the AE figure:

ThessalonikiGP	AE	IPR	Opp.IPR	Diff	Score
Whole event:	0.048	2696			
Bacrot	0.043	2696	(2964)	-268	-3
Caruana	0.048	2788	(2522)	+266	+4
Dominguez Pe	0.047	2643	(2597)	+047	+5
Grischuk	0.040	2672	(2698)	-026	+1
Ivanchuk	0.064	2610	(2777)	-167	-4
Kamsky	0.048	2842	(2578)	+264	+4
Kasimdzhanov	0.030	2883	(2863)	+020	=0
Morozevich	0.069	2478	(2683)	-205	-3
Nakamura	0.040	2762	(2711)	+051	-1
Ponomariov	0.037	2794	(2632)	+162	+1
Svidler	0.060	2662	(2727)	-065	-2
Topalov	0.054	2604	(2605)	-001	-2

I have begun re-doing my Rybka 3 Single-PV data with Houdini 3, going back to 2010. For a selection including all round-robin tournaments of Category 10 and higher, numbering 2,113 whole-event performances in all, the MM% figures stay close to each other in all percentiles. Shirov has 72.5% with Rybka and 70.7% with Houdini, Khenkin 71.0% with Rybka and 67.7% (fifth) with Houdini, place-50 is 64.3% with Rybka and 68.3% with Houdini, place-100 is 63.0% with Rybka and 62.8% with Houdini, and so on. Hence I intend soon to combine the results with Houdini since January 1 into the same master list.

The 76.4% matching in the Houdini Single-PV test from the seven accused SBL games is higher (by just 0.1%) than the highest in that list, that is higher matching than any performance in virtually the entire history of chess. When the two November games are restored, the resulting 71.2% is tied for 8th with Kramnik at the 1992 EU team championship in Debrecen. I have assembled comparison data from the past seven SBL seasons (since 2006–07), viewable at

[1]Besides Khenkin, 3071 while winning the 2011 German championship. By contrast, Christoph Natsidis had an IPR

`http://www.cse.buffalo.edu/~regan/chess/fidelity/data/, files Bundesliga20xx--yyHou3d17.{r3,sc3}`

I have not yet taken time to fix the SBL 2012–13 files' separation of names between Oct.–Nov. from Big Database 2013 and the rest from TWIC. This happens to coincide with the selection of seven accused SBL games. I have fixed some such accidents in the past, and have never seen a higher MM% figure over 120 or more moves in a partial result either.

Within the SBL, the next-highest figure is 67.1% by IM Bart Michiels in 2009–10 when his rating stayed around 2460. The other 12 performances of at least 65% are by Pelletier, Parligras, Yusupov, Ivanchuk, Shirov, Berkes, Wojtaszek, Bacrot, (Karsten) Müller, Gashimov, (WGM Inna) Gaponenko, and Kempinski, none from the past two seasons. Mr. Kotainy has raised comparison with Mr. Leon Mons, with a similar rating and excellent performance this SBL season. Mr. Mons has 53.4% matching in 73 analyzed moves before the split, 55.4% matching in 323 moves after it, i.e., from December onward, with AE (that is, scaled error) 0.067, 0.076 since December. Mr. Kotainy's AE in the accused games since December is 0.032, and there are 26 SBL performances tied or better in AE.

These comparisons do not carry formal statistical unlikelihood assertions—rather they place the performances in the relevant large human/historical contexts, and act as a "sanity check." The overall message strikes me as similar to the statistical results above. In the Ivanov case, as I reiterated in a statement quoted in an article at ChessBase.com last Wednesday, I cannot imagine stronger statistical results (using his pre-event rating most tests are over 6-sigma where the cited odds-giving applet just says "Infinity") and the comparisons gave no daylight. Here the odds are less extreme and some comparisons could be more so—for instance, Mr. Kotainy's AE from the seven accused SBL games could be lower.[2] It seems that a search for greater certitude would have to turn to considering individual games, where it may be possible to find corroboration from finer details and player/eyewitness accounts.

Here I must say that I have not compiled anything near the above comparison data sets for single games, as opposed to whole-event performances. Automating my setup to do so will probably require several person-weeks of student help. I also feel that my position requires refraining from chess-specific commentary of the kind made in certain videos and online articles and bulletin-board posts and the May issue of *Schach*. I can, however, convey these facts connected to games from the Neckar Open, including the two lost games among the six accused:

1. The lost game to GM Bacrot has almost exactly the expected MM (48%) and AE (4.5 scaled pawns total) expected for the current-2460 settings, including the final blunder 70...Rc8?? Moves 11–22 match, then move 24, then 7 of the next 10 moves; from move 35 onward most moves are non-matches. Mr. Kotainy's IPR for the whole game comes out exactly as 2462, with 2980 to move 34, 2870 to move 40. The article says that Bacrot suspected cheating and called for measures which were implemented during the game, as I have heard also in private, but I have not heard at which move that occurred.
2. The lost game to GM Solodovnichenko gives the winner an IPR of only 2218, while Mr. Kotainy receives 3314, with two-sigma confidence above 3187. The only other game I have recorded where I have noted a similar discrepancy between winner and loser is a blindfold game I lost to my son as we walked through the center of Vienna after Christmas 2011—he

[2] I have previously excluded league events from my database and training sets because of non-contiguous dates, widely varying numbers of games per player, and team-over-individual considerations. For instance, the top AE figure of

played some nice moves to win after I recovered almost from losing a piece in the opening. Both Rybka to depth 13 and Houdini to depth 17 fall into the trap of winning a pawn on the Kingside, with Houdini 3 (started before Black's move 23 with 256MB hash in single-PV mode on one core thread) still thinks Black is ahead with the 23...Qxh4 grab until giving 0.00 at depth 21 and +0.08 to White at depth 22. Since Black is otherwise in difficulty, this evaluation determines earlier moves.

3. The won game over Nisipeanu shows similar high MM% to the other games, and gives an IPR of 3274, even though there was no live transmission on Good Friday in compliance with German law. So it was on a high enough board to be recorded automatically, but was not transmitted from the Neckar Open website.

I have not delved into individual games beyond this. I can do so upon request, and will do so upon being apprised of relevant outside information, but a thorough statistical treatment with confidence intervals for projections of individual games would require well more than the effort already taken for this report.

Conclusions

The statistical data show significant deviations from each accused event, in each case well above the 99% confidence standard and approaching 99.9%. In the absence of independent evidence, one such occurrence would have to be ascribed as an effect of selection, since there are up to 1,000 new player-performances in comparable events every week, but two such occurrences cannot be. Comparisons to extensive historical data give several senses (IPR, MM% by non-GM) in which one or the other performance is an unprecedented outlier, but show at least one sense (AE in one SBL season) by which it is not. There are reflections in the data of incidental aspects of the two lost games. At the very least the results are sharply distinct from those for other strong player performances that have been mentioned for comparison, including some accused publicly of cheating. They also witness that spontaneous actions of other players have had reasonable objective grounds, and unless and until superseded by an official decision such as you are making, those actions may be considered defended by these results.

Nothing known about this case or shown in this report yet escapes the essential dilemma of whether statistical data alone should be regarded as determinative for sanctions against players. Reflecting the grade between civil and penal standards in society on the whole, this may depend on the nature and interpretation of the sanction. This is where my special expertise ends and why I have publicly called for wider consultation on these issues.

Attachment signed,

Dr. Kenneth W. Regan, 8 June 2013.