Using Chess Ratings as Data in Psychological Research

Mark E. Glickman[1]
Boston University

Christopher F. Chabris
Harvard University

Running head: *Chess Ratings as Data in Psychological Research*

[1]All correspondence concerning this document should be addressed to Professor Mark E. Glickman, Department of Mathematics, Boston University, 111 Cummington Street, Boston, MA, 02215. E-mail: *mg@math.bu.edu*

Since Binet (1893), psychologists have studied the game of chess to learn about fundamental properties of the human mind. Charness (1992) chronicled the rising influence on cognitive science of the seminal work of De Groot (e.g., 1965) and Chase and Simon (1973a, 1973b). The popularity of chess as a task domain for research on expertise stems from several factors, including its well-defined rules, susceptibility to symbolic representation and computer simulation, historical reputation as an unparalleled arena of pure thought, and its remarkable "cognitive fit" to the capacities of the human mind ("an hour to learn, a lifetime to master").

However, chess stands out also because of its rating system (Elo, 1986), which assigns to each player in official competitions a numerical value representing his "average strength." The larger the rating difference between two players, the better the higher-rated player is expected to score in a match between them. This relatively objective system allows researchers to select and group subjects with considerable knowledge of their abilities and the relation between their abilities and those of the larger population. There are numerous examples of this technique; to cite one, Chabris and Hamilton (1992), in a study of hemispheric differences in chess perception, restricted their sample to 16 players who were current or former holders of the "National Master" title, then defined by a rating of 2200 on the U.S. Chess Federation scale (the top 2.5% of competitive players in the country).

Charness and Gerchak (1996; hereafter CG) and Gobet and Simon (1996; GS) use chess rating data in new and intriguing ways to address questions about group differences and the tradeoff between search and knowledge in chess ability. They rely on ratings not merely to group subjects but to provide the actual data for analysis. Their approaches are valuable, but they each omit the crucial step of estimating the variability of their results [Footnote 1]. In this article we discuss how to perform this

analysis and the significant consequences for the CG and GS conclusions of accounting for this variability.

CG focus on the relationship between the size of a sample (e.g., of chess players) and its expected maximum value. They correctly argue that if two disjoint samples, one much larger than the other, are drawn from the same population, the larger sample will likely produce a larger maximum. As an example of this observation, CG examine the maximum playing strengths of different samples of chess players (men versus women, the United States versus the Soviet Union) and show that their "MILL7" approximation (i.e., for normal populations, the expected sample maximum is approximately linear in the logarithm [base 10] of the sample size with slope 0.7 standard deviations) seems to describe the observed differences. They conclude by suggesting that the maxima of samples be compared to a null hypothesis of the MILL7 expected values before generating new hypotheses (e.g., sex differences in cognitive ability, cultural differences in chess training) to explain why samples are different.

However, CG do not consider how to account for the variability of the sample maximum. As we describe below, the variability of a sample maximum is large enough that if a true difference exists between, say, the population mean ratings of men and women, this would not likely be detected by comparing the women's maximum rating and the men's maximum rating adjusted by MILL7 (and appropriately incorporating the variability of the sample maxima). Thus, while differences in the maximum chess rating between two different samples may be explainable by differential rates of participation, they will tell us very little (with any certainty) about the *average* differences between men and women, which may be a question of greater scientific interest.

Suppose in a sample of n observations from a continuous probability distribution

F (with density f), we wish to find the approximate distribution of the maximum. Instead of using the MILL7 approximation for the expected maximum of a sample, we can use an asymptotic normal approximation to the distribution of the t-th fractile, X_t, where we set $t = (n-1)/n$ for the maximum of a sample. The approximate distribution of X_t is given by

$$X_t \sim \mathrm{N}(F^{-1}(t), \frac{t(1-t)}{n[f(F^{-1}(t))]^2}), \tag{1}$$

where $F^{-1}(t)$ is the value corresponding to the t-th fractile of the distribution. This result can be found in advanced textbooks in statistics, such as Lehmann (1983) and Bickel and Doksum (1977). For most common distributions (e.g., Normal, Student's t, χ^2, Gamma), the value of $F^{-1}(t)$ can be calculated numerically using standard statistical software packages, for example, S-Plus (Statistical Sciences, Inc., 1995).

As an example of its application, consider CG's calculation (p. 49) of the expected rating for the best player in the Soviet Union in 1982. The authors assume the distribution of players' ratings follows $\mathrm{N}(1550, 250^2)$, and calculate an estimated maximum for a sample of 4,000,000 to be 2640. Numerically approximating (1) with $t = 3,999,999/4,000,000$ and F corresponding to $\mathrm{N}(1550, 250^2)$, we calculate that $F^{-1}(t) = 2807$, and that the sample maximum is approximately distributed as $\mathrm{N}(2807, 48^2)$.

Several points are worthy of mention. First, our estimated expected maximum (2807) differs substantially from CG's value of 2640. Regardless of the variability of the sample maximum, this difference points out the large potential for error in the approximations used (ours as well). Secondly, the (asymptotic) standard deviation of the sample maximum with $n = 4,000,000$ is about 48. Compared to the standard deviation of the sample mean, which in this example would be 0.125, the standard

deviation of the sample maximum is 384 times as large. That is, for repeated large samples, the maximum will be expected to vary by as much as 300–400 times as much as the sample mean will be expected to vary.

This large variability has a profound impact on statistical inference. Suppose, in the spirit of the preceding example, that we observe a maximum rating of 2720 in a sample of 4,000,000 with ratings that follow $N(1550, 250^2)$. In an independent second sample of 100,000 players drawn from $N(\mu, 250^2)$, with μ unknown, suppose we observe a maximum rating of 2730. At first glance, it would appear obvious that the second sample must have been drawn from a normal population with a higher mean μ than 1550 because not only was the maximum in the second sample higher, but that the sample maximum was based on a 40 times smaller sample (in smaller samples, according to the MILL7 approximation, one would not expect the maximum to be as large). But computing approximate 95% prediction intervals for the maximum from each sample demonstrates that this conclusion is not warranted. For the first population, the approximate 95% prediction interval is $2807 \pm 1.96(48) = (2713,\ 2901)$. Assuming $\mu = 1550$ for the second population, the sample maximum can be computed to come from $N(2616, 63^2)$ according to (1). So an approximate 95% prediction interval for the sample maximum is $2616 \pm 1.96(63) = (2493,\ 2739)$. In each case, the sample maxima fall in the prediction intervals, so they are consistent with the population distributions being identical. This demonstrates that using the sample maximum is not a powerful test statistic for detecting group differences.

If one's goal is to detect average differences among groups, one should choose procedures that are based on less variable statistics than the sample maximum. An obvious candidate is the sample mean, which is considerably less variable than the sample maximum. Even using lower order statistics, such as the top percentiles of

the sample, would reduce the variability appreciably relative to the sample maximum: with $n = 4,000,000$ values drawn from $\mathrm{N}(1550, 250^2)$, the standard deviation of the top one-tenth percentile (according to (1)) can be computed to be 1.17. The standard deviation of the top percentile is 0.467. The standard deviation of the top tenth percentile is 0.214. While obtaining such data is often more difficult than finding the sample maximum, doing so greatly reduces the variability. Another reason to avoid the sample maximum for carrying out inference is that the maximum is very sensitive to parametric assumptions, whereas a statistic like the sample mean is much less so (particularly for symmetric distributions). Making an inference about the center of a distribution using the sample maximum will result in a different conclusion depending on whether the data is assumed to be, say, normal versus t-distributed on small degrees of freedom, whereas using the sample mean will provide identical conclusions. This is an important issue to consider for measuring chess playing strength, because the distribution of chess ratings is not necessarily normal (Glickman, 1995).

GS carry out a study to determine whether recognition processes in chess are more important than planning processes in determining playing strength. To do so, they examine a sample of recent game results of the current world chess champion, Garry Kasparov, under conditions where he had less time to think than his opponents. From the 56 game outcomes, GS estimate that Kasparov's strength was not notably lower than his playing strength under normal tournament time limits. They therefore conclude that recognition processes, which occur nearly instantaneously, probably dominate planning and look-ahead processes, because otherwise Kasparov would have performed much worse than he did.

Like CG, GS do not account for the variability in their estimate of Kasparov's strength under restricted time limits, and therefore cannot describe the uncertainty

of their conclusions. Furthermore, the Elo (1986) linear approximation formula that GS use to compute an estimate of Kasparov's strength has poor statistical properties (e.g., it is not a consistent estimator), and is only intended as a simple approximation to more principled estimators, so it should not be relied upon. We analyzed their data using maximum likelihood estimation to obtain a more principled estimate of Kasparov's average playing strength, and calculate its standard error.

The probability model used in the Elo rating system is commonly known as the Bradley-Terry model (Bradley & Terry, 1952) for paired comparisons. The Elo system uses a reparametrization of the original Bradley-Terry model as follows. Suppose θ_i and θ_j are the true unknown ratings of players i and j. Then according to the Elo version, the probability that player i scores s_{ij}, where $s_{ij} = 1$ if player i wins and $s_{ij} = 0$ if player i loses, is given by

$$\Pr(s_{ij} = s) = \frac{(10^{(\theta_i - \theta_j)/400})^s}{1 + 10^{(\theta_i - \theta_j)/400}}, \tag{2}$$

The likelihood, $L(\theta|s_1, \ldots, s_n)$, over n games is therefore

$$L(\theta|s_1, \ldots, s_n) = \prod_{k=1}^{n} \frac{(10^{(\theta - \theta_k)/400})^{s_k}}{1 + 10^{(\theta - \theta_k)/400}} \tag{3}$$

where the k-th factor in the product is the result of the k-th opponent with rating θ_k. (When a draw occurs against opponent k, we can set $s_k = 0.5$ in the likelihood. This corresponds to information worth half a win and half a loss.) To perform maximum likelihood estimation from a set of collection of game outcomes, one can maximize the expression in (3) using the Newton-Raphson algorithm, a standard iterative numerical procedure, substituting the estimated opponents' Elo ratings for the θ_k. The algorithm also provides an asymptotic standard deviation of the rating estimate, which is the negative reciprocal of the second derivative of $L(\theta|s_1, \ldots, s_k)$ evaluated at the maximum.

After correcting a minor error in the GS data set (Kasparov's score in the United States junior team match was actually 0.5 worse than reported by GS), we carried out this analysis on the 56 game outcomes. The maximum likelihood estimate is 2594 with a standard error of 53.6. Thus, an approximate 95% confidence interval for Kasparov's true rating over the 56 games is $2594 \pm 1.96(53.6) = (2489,\ 2699)$. This interval estimate suggests a conclusion different from the one reached by GS, who obtain a point estimate of 2646. While this analysis shows that Kasparov played worse than his rating range of 2700–2790 during the same time period (July 1985 – July 1992) would indicate, it also shows that there is not enough information to conclusively determine how much worse. For example, the confidence interval suggests that Kasparov may be playing at a rating level of 200 or more points below his normal tournament strength [Footnote 2].

The apparent objectivity of Elo-type ratings when compared to other measures of relative ability (peer ratings, impact analyses, prize winnings, etc.) can mask the fact that they are still imperfect measures of underlying parameters, and the consequence that conclusions derived from them will be subject to variability. The articles reviewed here make innovative methodological contributions to the study of chess ability and expertise in general. We have tried to refine their approaches by estimating the variability of their conclusions – a step that all researchers using chess ratings as data should be careful to remember.

References

Bickel, P. J., & Doksum, K. A. (1977). *Mathematical statistics: Basic ideas and selected topics.* San Francisco: Holden-Day.

Binet, A. (1893). Les grandes mémoires: Résumé d'une enquête sur les joueurs d'échecs. [Mnemonic virtuosity: A study of chess players.] *Revue des Deux Mondes*, 117, 826–859.

Bradley, R., & Terry, M. (1952). The rank analysis of incomplete block designs. 1. The method of paired comparisons. *Biometrika*, 39, 324–345.

Chabris, C. F., & Hamilton, S. E. (1992). Hemispheric specialization for skilled perceptual organization by chessmasters. *Neuropsychologia*, 30, 47–57.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4, 55–81.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing.* (pp. 215–281). New York: Academic Press.

Charness, N. (1992). The impact of chess research on cognitive science. *Psychological Research*, 54, 4–9.

Charness, N., & Gerchak, Y. (1996). Participation rates and maximal performance: A log-linear explanation for group differences, such as Russian and male dominance in chess. *Psychological Science*, 7, 46–51.

de Groot, A. D. (1965). *Thought and choice in chess.* The Hague: Mouton.

Elo, A. E. (1986). *The rating of chessplayers, past and present* (2nd ed.). New York: Arco.

Glickman, M. E. (1995). Chess rating systems. *American Chess Journal*, 3, 59–102.

Glickman, M. E. (1996) *Parameter estimation in large dynamic paired comparison experiments.* Manuscript submitted for publication.

Gobet, F., & Simon, H. A. (1996). The roles of recognition processes and look-ahead search in time-constrained expert problem solving: Evidence from grand-master-level chess. *Psychological Science*, 7, 52–55.

Hsu, F.-H., Anantharaman., T., Campbell, M., & Nowatzyk, A. (1990). A grandmaster chess machine. *Scientific American*, 263, 44–50.

Lehmann, E. L. (1983). *Theory of point estimation.* New York: John Wiley & Sons.

Statistical Sciences, Inc. (1995). *S-Plus user's manual, version 3.3 for Windows.* Seattle: Statistical Sciences, Inc.

Thompson, K. (1982). Computer chess strength. In M. R. B. Clarke (Ed.), *Advances in computer chess 3.* (pp. 55–56). Oxford: Pergamon.

Author Note

Mark E. Glickman, Department of Mathematics. Christopher F. Chabris, Department of Psychology.

Christopher F. Chabris was supported by an NSF Graduate Fellowship.

We thank Christopher Avery, Neil Charness, and Andrew Metrick for their comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Mark E. Glickman, Department of Mathematics, Boston University, 111 Cummington Street, Boston, MA 02215; e-mail: *mg@math.bu.edu*.

Footnotes

1. Although our focus in this article is on the variability inherent in estimates made from chess rating data, one of the most overlooked features of Elo-type ratings is that they are themselves estimates of unknown strength or ability parameters, and are therefore as much subject to variability as any other measures of human performance. Glickman (1995, 1996) discusses how to measure the uncertainty of ratings and incorporate that information into rating calculations to increase accuracy and predictive power. CG and GS do not consider the variability associated with their raw data; indeed, it would be difficult for them to do so, given that the standard Elo rating system ignores the issue. Fortunately, their conclusions would not be significantly altered by accounting for reasonable estimates of rating variability. However, future researchers should consider this factor carefully when drawing inferences from rating data.

2. A rating difference of 200 points is significant because it predicts a 3-1 victory margin for the superior player in a match (e.g., Elo, 1986), which would be considered a decisive result. (For example, all recent world championship matches have been decided by smaller margins.) Thus, if Kasparov lost 200 points of strength under clock-simultaneous conditions, it would be fair to conclude that the lost thinking time affected his play significantly. It is interesting to note that research in computer chess (e.g., Thompson, 1982; Hsu, Anantharaman, Campbell, and Nowatzyck, 1990) has equated a 200-point advantage to the approximate benefit derived from searching one ply (one move for one side) deeper in the game tree, and that this additional search typically increases the time spent by a factor of 4–6 – the same time constraints imposed on Kasparov. So the GS study also offers some support for the idea that forward search is crucial to chess skill, though again the variability precludes any definitive judgments.