Comparing Extreme Members is a Low-Power Method of Comparing Groups: An Example Using Sex Differences in Chess Performance

Mark E. Glickman, Ph.D.[1, 2]

Christopher F. Chabris, Ph.D.[3]

[1] Center for Health Quality, Outcomes and Economic Research, a VA HSR&D Center of Excellence, Bedford MA.

[2] Boston University School of Public Health, Department of Health Policy and Management, Boston MA.

[3] Department of Psychology, Union College, Schenectady, NY.

Author Affiliations

Mark E. Glickman, Ph.D.

Center for Health Quality, Outcomes, and Economic Research

Edith Nourse Rogers Memorial Veterans Hospital

200 Springs Road (152)

Bedford, MA 01730

Tel: 781-687-2875

Fax: 781-687-3106

Email: mg@bu.edu

Christopher F. Chabris, Ph.D.

Department of Psychology

Union College

807 Union Street

Schenectady, NY 12308

Email: chabris@gmail.com

Four years after he won the world chess championship, Garry Kasparov was quoted as saying "there is real chess and there is women's chess … chess does not fit women properly" (Chelminski, 1989). It is true that no woman has ever come close to winning the world chess championship, and that men vastly outnumber women at the highest levels of chess achievement. However, it is also obvious that men outnumber women at all levels in chess, and this difference in overall "participation rates" (the proportion of all men and women who choose to enter competitive chess) has been cited to explain the difference in high achievement (Charness & Gerchak, 1996; Howard, 2005; Chabris & Glickman, 2006). The question of sex differences in achievement is equally salient in other fields where more men reach the top levels than women, such as academia, business, and the law.

Chess is an excellent domain in which to study predictors of performance because of its relatively objective rating system developed by Arpad Elo (1986), which assigns to each player in official competitions a numerical value representing his or her "strength." The larger the rating difference between two players, the better the higher-rated player is expected to score in a match between them. Bilalić, Smallbone, McLeod, and Gobet (2009; hereafter BSMG) use chess rating data in an intriguing way to address the question of sex differences in chess ability. BSMG develop an approximation to calculate the expected value of the k-th highest value in a sample of n from a normal distribution, and use the result to compare chess ratings between top males and females. Their work extends a result by Charness and Gerchak (1996), who derive an approximation to the expected maximum of a sample from a normal distribution. BSMG apply their approximation to the top 100 male and female German tournament chess players, and this analysis shows that it is difficult to conclude that men are better than women on average even though the best men have much higher ratings than the best women. BSMG correctly observe

that the larger male German population of chess players would, simply by chance, produce better players even when the averages are the same. They go one step further and conclude that above the levels of the 80th-best men and women players, women are actually higher-rated than expected relative to men based on the sample sizes. They argue, like Charness and Gerchak, that one must account for participation rates when comparing the best achievers before generating new hypotheses (e.g., differences in cognitive ability or training regimens) to explain performance differences between groups.

The main drawback of BSMG's analyses is that they do not account for the inherently high variability of the extreme values in a sample. While differences in the highest chess rating between men and women 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. By design, comparing only the highest achievers is a low-power procedure that is not likely to produce useful results.

To see why this is so, suppose in a sample of n observations from a population with continuous probability distribution function F (and density f), we wish to find the approximate distribution of the highest values. Instead of using BSMG's approximation for the k-th largest value of a normally distributed sample, we can use an asymptotic normal approximation to the distribution of the t-th fractile, X_t, where $t = (n-k)/n$. The approximate distribution of X_t is normal, and is given by

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

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 (e.g., Lehmann, 1983; Bickel & Doksum, 1977). For

most common distributions, the value of $F^{-1}(t)$ can be calculated numerically using standard statistical software packages, such as R (R Development Core Team, 2008).

As an example, consider BSMG's comparison of the 5th best male to the 5th best female. By inspection of the graph in their Figure 2, the observed rating difference is about 290 in favor of the male player. The authors assume that the distribution of all players' ratings is normal, following $N(1461, 342^2)$, and that the numbers of male and female chess players are 113,386 and 7,013 respectively. From our formula, the approximate distribution of the 5th order statistic for men is $N(2802, 36.9^2)$ and for women it is $N(2552, 44.2^2)$, so that the distribution of the difference is $N(250, 57.6^2)$; note that the variance of the difference is the sum of the individual variances. Thus an approximate 95% confidence interval for the rating difference between the 5th best male and female is 137.1–362.9, which is arguably too wide an interval to serve as a diagnostic for whether men are stronger chess players than women *on average*.

A related problem is revealed when we compare the 100th best male and female German tournament chess players. According to the BSMG approximation formula, the 100th best male should be rated near 2495.6 and the 100th best female should be rated near 2066.1, a difference of 429.5 (it is unclear to us how BSMG arrived at the value of 440 mentioned on p. 1162). The observed difference appears to be about 380, based on the graph in their Figure 2. Using our formula, the 100th best male rating has a $N(2530.6, 10.0^2)$ distribution, and the 100th best female rating has a $N(2210.0, 13.4^2)$ distribution, so that the male–female difference follows a $N(320.6, 16.7^2)$ distribution, with a 95% confidence interval of 287.9–353.3. Relative to our mean, here again it appears as though men outperform women (significantly), which is the reverse of the conclusion presented by BSMG.

In fact, this conclusion is not justified either, because it is sensitive to an unchecked and potentially false assumption. Underlying the calculations made by both BSMG and ourselves is the assumption that chess ratings are distributed normally. This is a crucial assumption, and one that is arguably not satisfied by actual chess rating systems. The apparent justification for assuming a normal distribution in BSMG's analysis is in their Figure 1, which shows a superimposed normal density function having similar features to the empirical rating distribution. It is difficult to determine from this graph whether the right tail of the rating distribution is normal (a normal probability plot might help address this question), but there is nothing in the statistical architecture of chess rating systems that favors ratings being distributed normally (see Glickman, 1995, for a detailed discussion of this issue).

To demonstrate the extent to which the assumed distribution can affect conclusions about the comparison between the top men and women, assume that German chess federation ratings have the mean and standard deviation specified by BSMG, but that the ratings follow a t-distribution with some specified degrees of freedom. Histograms of data coming from a t-distribution and a normal distribution would be virtually indistinguishable, but a t-distribution has tails that are sufficiently heavy to affect the analysis of the extremes. Such t-based models are becoming increasingly popular for robust data analyses (e.g., see Lange et al., 1989). Using our formula, we calculated the estimated ratings of the 100th best male and female assuming chess ratings truly followed a t-distribution with 15 degrees of freedom instead of a normal distribution. These ratings would follow $N(2752.7, 19.3^2)$ for the male and $N(2243.2, 17.7^2)$ for the female, so that the difference would follow $N(509.5, 26.2^2)$. Our normal distribution calculation resulted in a mean of 320.6, which is 188.9 less than the estimate based on the t-distribution. This very large discrepancy stems entirely from the different assumptions about the

distribution of ratings. Unless the analyst is sure about this distribution, specifically at the right tails, any statistical comparison between top order statistics is highly uncertain not only because extremes tend to vary greatly, but also because the assumed distribution of the data may be incorrect.

If one's goal is to detect average differences among groups, one should choose procedures that are based on less variable statistics than an analysis of extremes, and ones that are more robust to distributional assumptions. An obvious candidate is the sample mean, which is considerably less variable than high-order statistics. Even using lower order statistics, such as the top 10th or 20th percentile of the sample, would reduce the variability appreciably relative to the ones used by BSMG. Using the mean, or even the lower percentiles of the empirical distribution, is also much less sensitive to distributional assumptions than is using the highest values. We took this approach to examine sex differences in chess ability among 250,000 U.S. rated players; we found that the male mean was significantly higher than the female mean, but that this difference itself might result from the much larger number of boys than girls who enter competition (Chabris & Glickman, 2006; see also Maass et al., 2008).

The greater objectivity of Elo-type ratings as compared to other measures of relative ability (peer evaluations, impact analyses, patents, 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. Researchers using chess ratings as data to answer questions about patterns of human performance should keep in mind that this variability is greatest for extreme values in a distribution, and that the extremes are also very sensitive to small changes in the underlying form of the distribution. Accordingly, though the conclusion

BSMG arrived at could be correct, the procedures they followed do not have the statistical power to support it.

References

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

Bilalić, M., Smallbone, K., McLeod, P., & Gobet, F. (2009). Why are (the best) women so good at chess? Participation rates and gender differences in intellectual domains. *Proceedings of the Royal Society B, 276,* 1161–1165.

Chabris, C.F., & Glickman, M.E. (2006). Sex differences in intellectual performance: Analysis of a large cohort of competitive chess players. *Psychological Science*, 17, 1009–1107.

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.

Chelminski, R. (1989). Playboy interview: Garry Kasparov. *Playboy,* November. [http://www.playboy.com/articles/garry-kasparov-1989-interview/index.html]

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.

Howard, R.W. (2005). Are gender differences in high achievement disappearing? A test in one intellectual domain. *Journal of Biosocial Science, 37,* 371–380.

Lange, K.L., Little, R.J.A., & Taylor, J.M.G. (1989) Robust statistical modeling using the t distribution. *Journal of the American Statistical Association, 84,* 881–896.

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

Maass, A., D'Ettole, C., & Cadinu, M. (2008). Checkmate? The role of gender stereotypes in the ultimate intellectual sport. *European Journal of Social Psychology, 38,* 231–245.

R Development Core Team (2008) *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. [http://www.R-project.org]

Acknowledgments

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