Can higher social status of competitors cause decision makers to commit more errors?

Li Qian Tay (liqian.tay@research.uwa.edu.au)
School of Psychological Science
University of Western Australia

Abstract

The ability to make good decisions is critical in life. Although anecdotal and preliminary evidence suggests that social comparison could impair decision making, surprisingly little attention has been paid to such dynamics within cognitive science. The present study aimed to address this gap by exploring, via a sample of 1.5 million chess games and a fuzzy regression discontinuity design, whether higher status of competitors could cause individuals to commit more errors. Critically, chess data includes overt symbols of social status, viz. titles conferred at arbitrary thresholds of ratings that represent playing strength, and an objective measure of errors could be calculated by contrasting the moves that players chose in games against the optimal moves determined by powerful chess engines. I found no evidence that the mere presence of status titles impacted error rates.

Keywords: decision making; error rate; cognitive psychology, social psychology, regression discontinuity design; chess

Introduction

Decision making is ubiquitous in daily lives and cognitive science has revealed important insights into this critical process over the last decades (e.g., Kahneman & Tversky, 1974; Mata et al., 2018; Trueblood et al., 2021). Computational models have led to significant theoretical advances regarding the latent processes that guide individuals' decisions. For example, the drift diffusion model is a common evidence accumulation model that formalises drift rate, the speed with which evidence accumulates in a decision maker, and response boundaries that quantify the amount of evidence required prior to decision (Ratcliff et al., 2016). However, much of the research has focussed on factors at the individual level, leaving the potential impact of social dynamics on decision making underexplored (Kish-Gephart, 2017).

In real-world scenarios, such as sports or business negotiations, decisions often involve multiple parties and may be competitive in nature. Social psychologists have therefore long proposed a key role for social comparison, the tendency for individuals to self-evaluate not just with objective information or criteria but also in comparisons to others (e.g., Festinger, 1954). For example, social comparison has been said to drive competitiveness and negative emotions, including envy and even fatigue (e.g., Latif et al., 2021). Other work found that the introduction of overt social comparison mechanisms to classrooms, such as reward badges and leaderboards, could reduce students' intrinsic motivation (Hanus & Fox, 2015; Orosz et al., 2013). Nonetheless, many of these studies relied on self-report questionnaires or decision tasks with hypothetical scenarios, where validity rests on the assumption that individuals can adequately introspect and provide honest appraisals. Notwithstanding the advantages of this approach, such an assumption may not always be tenable, particularly if the choices that individuals make are not incentivized and both rewards and costs are hypothetical (see also Hertwig et al., 2018; Loomis, 2011).

Given the above, I opted to exploit online chess data to offer incremental progress within this area of social cognition. Briefly, chess is a two-player competitive game, in which each player controls a set of 16 white or black pieces that can move across a checkerboard. The aim is to capture the opponent's pieces, and ultimately their king. In traditional, offline games, it is typical to quantitatively estimate player skill through rating systems (e.g., Elo, Glicko-2; Glickman, 2012). For instance, when players compete in matches or tournaments, such as those held by FIDE (The International Chess Federation), their ratings increase after wins and decrease after losses, scaled by the rating of the player and the ratings of the competitors. Critically, players above certain ratings may hold official titles (e.g., Candidate Master, FIDE Master, International Master, Grandmaster). Some of these titles are obtained automatically by reaching predefined thresholds (e.g., rating ≥ 2200 for Candidate Master and rating ≥ 2300 for FIDE Master); others require achievement of norms at FIDE-rated tournaments (e.g., performance of ≥ 2600 against opponents with average rating ≥ 2380 in FIDE-rated tournaments is needed to achieve Grandmaster norms).

The ratings of online chess follow similar conventions, with players gaining or losing ratings after games. Although online chess does not confer titles, traditionally titled players can verify their identities to replicate their titles online. These titles are then displayed alongside the verified players' usernames (which may or may not be their real names) in all games. This means that in the present research, I was able to

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

use the presence of a title as an overt indicator of higher status. Moreover, the advent of powerful chess engines (i.e., computer programs that can run analyses of chess positions) allowed for an objective measure of errors. Specifically, I focussed on the average centipawn losses of players in each game. A centipawn is a unit of measurement in chess representing 1/100 of a pawn, the weakest piece in the game, and centipawn losses could be calculated by contrasting a player's move choice in each position against the ideal choice determined by chess engines. This is a common measure in professional chess and allows for a nuanced measure of errors beyond outcomes (e.g., two players can draw a game, but one could still have committed more errors and thus have a higher average centipawn loss than the other).

I asked one simple question: Could the mere presence of a symbol indicative of competitors' higher social status cause decision makers to commit more errors?

Method

Data was retrieved from the open-source chess platform Lichess (Lichess, 2023). I focussed on all standard chess games played on the platform in September 2020. This was chosen to balance between achieving a large sample with minimising download and decompression time, given no a priori reason to believe that any impact of month and year on the research question exists. I excluded Bullet chess, in which players only have a maximum of 1 minute for the entire game, as such games tend to be characterised by chaotic gameplay due to the paucity of time, and which I reasoned as likely tapping into different cognitive processes than games with longer time controls. Further, due to computational constraints, I focussed only on games and positions with readily available Stockfish evaluations in the Lichess database. Stockfish is a free and open-source chess engine that has repeatedly won the global Chess Engine Competition (Stockfish, 2022). This resulted in a final sample size of 1,511,826 games.

Design

I employed a fuzzy regression discontinuity design. Briefly, a regression discontinuity design is an observational causal inference method (or, quasi-experimental method) characterised by a running variable, a treatment, and an outcome variable. The running variable in a fuzzy regression discontinuity design is assumed to increase the probability of the treatment past an arbitrary threshold along the running variable, but certain units below the threshold could still receive the treatment and certain units above the threshold might not receive the treatment (vs. sharp regression discontinuity designs, whereby the probability of treatment jumps from 0 to 1 immediately past threshold). Conceptually, the design works by acting as a "local randomised experiment", with units of analysis just below or above the arbitrary threshold that are comparable, except for the difference in probability of treatment. Any difference in outcomes for units around this threshold would then be estimated as the causal impact of the treatment (see Cattaneo et al., 2015). Similar designs have been used to examine the causal effects of social services programs, school improvement grants, alcohol consumption, remedial education, etc. (e.g., Carpenter & Dobkin, 2009; Dragoset et al., 2017; Varacca, 2022).

Critically, within the current context, chess ratings could be thought of as a running variable and chess titles as the treatment. Given the minimum rating requirement for a Candidate Master title by FIDE is 2200, players rated above 2200 should have an increased probability of having a title on Lichess. This would be the case even though not all title holders verify their identities to hold a title on Lichess. I thus exploited a local randomness around rating = 2200, where players with and without titles would be most comparable. Two separate sets of analyses were run, one treating players with black pieces as the competitor (players with white pieces as the decision maker) and the other treating players with white pieces as the competitor (players with black pieces as the decision maker). Across analyses, the running and treatment variables were always the rating and title of the competitor, and the outcome variable was the error rate of the decision maker.

Measures

Chess Rating. Ratings of players on Lichess follow the Glicko-2 system (Glickman, 2012). All players, upon creation of an account on the online server, start with rating = 1500, except for if they otherwise hold verified identities with FIDE or national titles. As with traditional chess, wins in games increase ratings and losses decrease ratings. The rating system thus offers an objective and transparent quantification of playing strength against competitors, as the ratings of players are prominently displayed in all games. Such ratings have previously been used to study the effects of expertise (e.g., cognitive abilities of high-rated players vs. low-rated players in the German Chess Database; Vaci & Bilalic, 2017). Here, as aforementioned, the rating was used as the running variable.

Social Status. The mere presence of a title (e.g., Candidate Master, FIDE Master, International Master, Grandmaster) was operationalised as indicative of high status. Status was therefore a binary variable, representing treatment.

Error Measure. Errors were measured via average centipawn loss by players across black and white pieces for a given game. I extracted pre-existing cloud analyses by Stockfish on Lichess and then calculated average centipawn losses for all moves in each game in the sample by averaging over every deviation detected by the chess engine. Higher average centipawn losses represented greater rates of error.

Results

As is customary, I first regressed the error measure on the treatment, competitor status. Results suggest that having a higher-status competitor was associated with greater error rates for when the decision maker played as white,

$t(1511824) = 6.91$, $p < .001$, or black, $t(1511824) = 12.60$, $p < .001$. I also regressed the error measure on the competitor's ratings, and again the competitor's rating significantly predicted decision makers' errors when the decision maker played as white, $t(1511824) = 88.21$, $p < .001$, or black, $t(1511824) = 109.84$, $p < .001$. Nonetheless, these patterns could be explained by confounds that impact both the likelihood of players having titles and making mistakes (e.g., latent abilities of the players; Lichess's pairing system that preferentially allocates players of similar ratings; other unobserved variables).

I thus turned to the regression discontinuity analysis. I focused on inference under local randomization and analysed the data using the R package *rdlocrand* (Cattaneo et al., 2022; see also Cattaneo et al., 2016). Figure 1 presents data points around the threshold of rating = 2200.

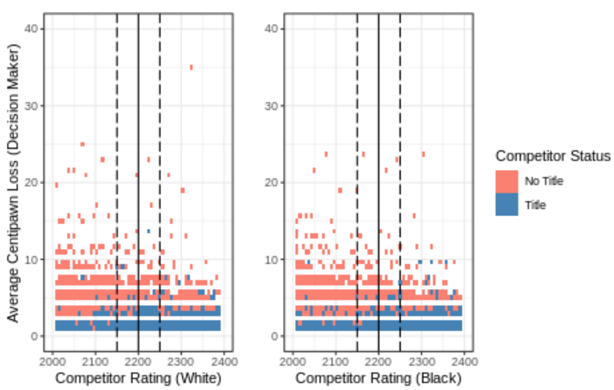

Figure 1: Binned data; solid lines represent threshold; dashed lines represent selected window.

Figure 1 suggests that no discontinuity in outcomes exists around the threshold. To test this, the window for analysis with as-if randomisation was specified as rating = 2150 to 2250 (i.e., ±50 the threshold). Intention-to-treat analyses revealed non-significant results for when the decision maker played as white, $t(19946) = 0.012$, $p = .458$, or black, $t(19677) = 0.001$, $p = .366$. Note that the lack of pre-treatment covariates precluded data-driven selection of the rating window. However, results were insensitive to window sizes up to ratings ≈ 2050 to 2350, where differences turned statistically significant, but at which point the local randomisation assumption was less credible.

Discussion

The ability to make good decisions is amongst the most important skills an individual can have. However, even though decisions in the real world are seldom made in isolation but can rather involve a range of social dynamics, the potential effects of such dynamics on the quality of decisions have garnered relatively little attention in cognitive psychology. In this study, I sought to address this gap by testing the possibility that the social status of competitors can impact individuals' error rates in a consequential decision-making context using online chess data.

Contrary to expectations, I found no evidence that the mere presence of a status symbol caused individuals to commit more errors. This challenges the simple narrative that social dynamics are ubiquitously impactful, whilst bringing to attention the need to conduct causal analysis based on objective assessments in addition to correlational analysis with self-reported data. Nonetheless, I caution against drawing strong conclusions based on exploratory results and instead point to a few plausible explanations that warrant further investigation. For one, Lichess's pairing system that preferentially allocates players of similar ratings and the regression discontinuity design used in the current study relies on a small window around the rating = 2200 threshold. One possibility, given current results, is therefore that social comparison is impactful only when status differentials are subjectively large (see also Demakakos et al., 2008). Players around the threshold in the current study may already consider themselves to have high status, particularly if they are amateur rather than professional players. These players may also be especially motivated or focussed when facing high-status competitors. This could have countered any potential negative effect of the titles.

Second, one a-priori benefit of using chess data is the separation of playing strength, as indicated by chess ratings, and social status, as indicated by titles (even if higher playing strength predicts higher status). If, however, social status exerts impacts only when individuals lack objective assessments, the prominent display of competitors' chess ratings in games could have again neutralised the titles' effect. Further, there is the issue of decision duration. To illustrate, consider that social comparisons were found to lower students' intrinsic motivation (e.g., Orosz et al., 2013). If such effects only compound over days, weeks, or even months and years, the focus on chess games may simply have too limited a window of analysis. Taken together, I suggest that follow-up research should thus: (1) investigate the potential moderating effects of subjective status differential and motivation (e.g., by systematically varying the size of actual status differences and incentives), (2) disentangle subjective and objective indicators of abilities and status (e.g., by systematically varying the display of objective performance indicators and status symbols), (3) adopt a longitudinal or mixed-method stance (see also Timans et al., 2019); and consider alternative operationalisations of social status (e.g., differing levels in a socioeconomic or organizational hierarchies, which do not necessarily involve ability).

Importantly, beyond the empirical result and theoretical discussion, the current study also contributes methodologically to the literature. Although the use of chess to study cognitive processes has a long history in cognitive science (e.g., Van Harreveld et al., 2007), to the best of my knowledge the current study is the first to utilise a regression discontinuity design to examine causal effects. Considering that there is a general, but unfounded, reluctance to employ non-experimental causal methodologies in psychology (e.g., Grosz et al., 2020), the current study represents yet another incremental progress in this area. This is important because many factors, whether social or otherwise, might not be manipulable due to ethical or feasibility considerations, and researchers often must rely on proxies that require further assumptions. For instance, it remains unclear whether asking study participants to imagine the government's implementation of a policy accurately reflects the cognitive processes involved for when the government actually implements the policy (see also hypothetical bias; e.g., Loomis, 2011). Indeed, whilst the current study focussed on chess, researchers should find that there exist a range of other data sources of psychological and public interests that are conducive for such analyses (e.g., crime reports, mobility data, natural disasters, and social media data; also see Marinescu et al., 2018).

To conclude, I employed a novel approach to study social cognition. Results suggest that the relationship between the competitors' higher status and decision makers' errors, if any, is likely complex. Although further research is required, this calls into question the simpler narrative that social dynamics are ubiquitously impactful. It is hoped the current study will contribute to increasingly nuanced discussion about the potential influence of social factors during decision making.

References

Carpenter, C., & Dobkin, C. (2009). The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics, 1*(1), 164-182.

Cattaneo, M. D., Frandsen, B. R., & Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference, 3*(1), 1-24.

Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2016). Inference in regression discontinuity designs under local randomization. *The Stata Journal, 16*(2), 331-367.

Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2022). *rdlocrand: Local Randomization Methods for RD Designs*. https://cran.r-project.org/package=rdlocrand

Demakakos, P., Nazroo, J., Breeze, E., & Marmot, M. (2008). Socioeconomic status and health: the role of subjective social status. *Social science & medicine, 67*(2), 330-340.

Dragoset, L., Thomas, J., Herrmann, M., Deke, J., James-Burdumy, S., & Lee Luca, D. (2019). The impact of school improvement grants on student outcomes: Findings from a national evaluation using a regression discontinuity design. *Journal of Research on Educational Effectiveness, 12*(2), 215-250.

Festinger, L. (1954). A theory of social comparison processes. Human relations, 7(2), 117-140.

Glickman, M. E. (2012). *Example of the Glicko-2 system.* Boston University, 1-6.

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science, 15*(5), 1243-1255.

Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & education, 80*, 152-161.

Hertwig, R., Wulff, D. U., & Mata, R. (2019). Three gaps and what they may mean for risk preference. *Philosophical Transactions of the Royal Society B, 374*(1766), 20180140.

Kish-Gephart, J. J. (2017). Social class & risk preferences and behavior. *Current opinion in psychology, 18*, 89-92.

Latif, K., Weng, Q., Pitafi, A. H., Ali, A., Siddiqui, A. W., Malik, M. Y., & Latif, Z. (2021). Social comparison as a double-edged sword on social media: The role of envy type and online social identity. *Telematics and Informatics, 56*, 101470.

Lichess. (2023). *Open-source chess server*. https://Lichess.org.

Loomis, J. (2011). What's to know about hypothetical bias in stated preference valuation studies?. *Journal of Economic Surveys, 25*(2), 363-370.

Marinescu, I. E., Lawlor, P. N., & Kording, K. P. (2018). Quasi-experimental causality in neuroscience and behavioural research. *Nature human behaviour, 2*(12), 891-898.

Mata, R., Frey, R., Richter, D., Schupp, J., & Hertwig, R. (2018). Risk preference: A view from psychology. *Journal of Economic Perspectives, 32*(2), 155-172.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in cognitive sciences, 20*(4), 260-281.

Stockfish. (2023). *Open-source chess engine.* https://stockfishchess.org/

Timans, R., Wouters, P., & Heilbron, J. (2019). Mixed methods research: what it is and what it could be. *Theory and Society, 48*, 193-216.

Trueblood, J. S., Heathcote, A., Evans, N. J., & Holmes, W. R. (2021). Urgency, leakage, and the relative nature of information processing in decision-making. *Psychological Review, 128*(1), 160.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science, 185*(4157), 1124-1131.

Vaci, N., & Bilalić, M. (2017). Chess databases as a research vehicle in psychology: Modeling large data. *Behavior research methods, 49*, 1227-1240.

Van Harreveld, F., Wagenmakers, E. J., & Van Der Maas, H. L. (2007). The effects of time pressure on chess skill: an investigation into fast and slow processes underlying expert performance. *Psychological research, 71*, 591-597.