No stereotype threat effect in international chess

Tom Stafford (t.stafford@sheffield.ac.uk)
Department of Psychology, University of Sheffield
Sheffield, S102TP, United Kingdom

Abstract

We examine data from over 6.6 million games of tournament chess between players rated by the international chess authority, FIDE. Previous research has focussed on the low representation of women in chess. We replicate and extend previous analysis (Chabris and Glickman, 2006) on an international level. We find no support for differential variability, differential drop-out between male and female players, or social context (in the form of proportion of female players at a national level) as drivers of drivers of male-female differences. Further, we examine games between mixed and same gender pairs for evidence of a 'stereotype threat' effect. Contrary to previous reports, we find no evidence of stereotype threat. Though this analysis contradicts one specific mechanism whereby gender stereotype may influence players, the persistent differences between male and female players suggests that systematic factors do exist and remain to be uncovered.

Keywords: learning; chess; skill acquisition; expertise;

Introduction

Chess has an illustrious history within cognitive science (Newell et al., 1958; Chase & Simon, 1973; Charness, 1992), providing a paradigmatic example of cognitive skill, and a testbed for theories of skill acquisition and performance. Aside from its worldwide popularity, and historical and cultural interest, chess has the advantage of being a skill with minimal perceptual or motor requirements, the upper-bound on an individual's performance being their cognitive capacity in planning, and reasoning through the complex space of possible moves. Chess also has the advantage that players are rated using the Elo system (Elo, 1978), which updates according to a player's success or failure in games against other rated players. This provides an objective measure of skill, which is not directly contaminated by the subjective perception of observers.

The chess playing community is predominantly male. Previous research has explored a number of possible competing explanations for the under-representation of women in chess (Chabris & Glickman, 2006; Bilalić et al., 2009). In their study, Chabris and Glickman (2006) looked at 250,000 US tournament players and found that men's ratings were not more variable than women's and that younger players did not learn at different rates according to gender. They found that men's ratings were higher than women's in areas where there were fewer women players (and not so in areas, defined by zip code, where at least 50% of younger players are female) — a result which is suggestive of an effect of social context on players' performance or skill acquisition which interacts with gender.

One notable psychological phenomenon which can influence performance is that of 'stereotype threat', whereby an individuals' awareness of a negative stereotype influences their performance. This was originally proposed for intelligence test performance and African Americans (Steele & Aronson, 1995), and has since been extended to other domains, most pertinently for our purposes to women and performance in non-stereotypically feminine domains of achievement, such as mathematics (Spencer et al., 1999).

In chess, both observational (Rothgerber & Wolsiefer, 2014) and experimental studies (Maass et al., 2008) appear to confirm the existence of stereotype threat. Rothgerber and Wolsiefer (2014) looked at 219 female chess players and found evidence for stereotype threat in the field — these players under-performed relative to their rating when they played male players. They report (p.79) that "Stereotype threat susceptibility was most pronounced in contexts that could be considered challenging: when playing a strong or moderate opponent". It should be noted that their sample was very young (5 – 15 years). Maass and colleagues (2008) ran an study using internet chess, where the perceived gender of opponents was experimentally manipulated with 84 participants (half male, half female, mean age 33.5). When believing they were playing an opponent of the opposite gender female players were less likely to win. If these findings generalise widely to chess performance they have the potential to systematically undermine the performance of female players.

A recent analysis suggests that publication bias has exaggerated the reality of the stereotype threat phenomenon (Flore & Wicherts, 2015) (see also (Stricker, 2008; Ganley et al., 2013)). In a similar vein, there is a possibility that a methodological confound is responsible for some positive replications of the effect, at least in the domain of mathematics and a gender stereotype threat effect (Stoet & Geary, 2012).

So although an obvious disparity exists in participation rates between men and women, there is uncertainty over the mechanisms by which this is perpetuated. In particular, the phenomenon of stereotype threat offers a specific psychological mechanism whereby cultural stereotypes and the existing relative paucity of female role models can interact with gender to hamper women's achieve-

ments in chess, but has not been convincingly established for a wide age range playing at the higher levels of the game. This is what this study set out to do.

Data and method

Our data comprise 6,641,158 games played by 461,637 FIDE rated players, of which 56,474 (12.2%) are women. Of these players, 176,583 were active during the 92 month period for which we have data, and for these players we have statistics on each game they played (and subsequent rating changes). The average birth year for these players was 1983, with a standard deviation of 19.78.

For each player the data consists of a unique player ID, date of birth, gender, nationality and a details of the games they played (including the piece colour they played as - White or Black - who they played against, the tournament this was part of, and the outcome). The data also contains all players' official FIDE rating calculated according to the Elo system. This system updates players' ratings according to game outcomes and acts both as a prediction system for the outcome of a match of any two rated players and as a way of ranking any player against the historical community of all players contained within the system. Because Elo ratings are updated after each game, it is possible to compare players who have never played, and may not even be contemporaneous.

Our analytic strategy is to first confirm for our international sample the differences found in the US sample studied by Chabris and Glickman (2006), and then to explore in greater detail the possible influence of social context — and particularly the availability of female role models - has on learning and performance. Specifically, this means to confirm a difference in ratings between male and female players, which cannot be attributed to differential variability in ratings or drop-out. Next, to see if drop-out and rating differences between young male and female players can be attributed to difference in proportion of female players at a country level. Finally, we investigate in detail the possibility of stereotype threat, a candidate phenomenon for reduced female performance in stereotypically male domains. The advantage of chess is that we are able to precisely gauge the challenge presented by individual games to each player, via comparison of player ratings.

Analysis scripts are available at `https://osf.io/aeksv`. For commercial reasons the full dataset is not available at the point of writing.

Results

Differences in ratings and drop-out

The average FIDE rating of men is 1880 (standard deviation 295, and for women 1700 (standard deviation 318). This difference is statistically significant, $t(176096) = 74.31, p < 0.0001$). For reference, a rating above 2500 is associated with Chess Grandmaster level.

The ratio of the standard deviation ratings for women to men was 1.08. Like Chabris and Glickman (2006), this shows higher variability in women's ratings, offering no support to the hypothesis that males are more variable in their ability.

One possible explanation for the male advantage in ratings is that women drop out of chess at higher rates than men. Following Chabris and Glickman (2006), we looked at young players (aged 5 to 25, birth years 1987–2007) who were active in the third year covered by our data, tracking them to see if they remained active in the remaining years covered by our data.

The average number of days until one of these players became inactive was 553 (for 28,190 men) and 560 (for 6,510 women, a significant difference ($t(34698) = 2.42, p = 0.016$). Clearly higher drop-out rate among potentially strong female players is not a plausible explanation for any male advantage in ratings.

Differences by country

One proposed mechanism by which a gender difference may be created, perpetuated or exacerbated is the minority status of women chess players. Considerable variation exists in the proportion of chess players who are female by nation (minimum: 2.0%, Denmark; maximum: 33.9%, Mozambique). The proportion of female chess players at competition level could conceivably influence younger female players, by providing role-models, mentors, or merely by lessening their perceived minority status. To investigate this issue, we compared the difference between male and female ratings, and between male and female time-until-drop-out, against the proportion of female chess players in each country.

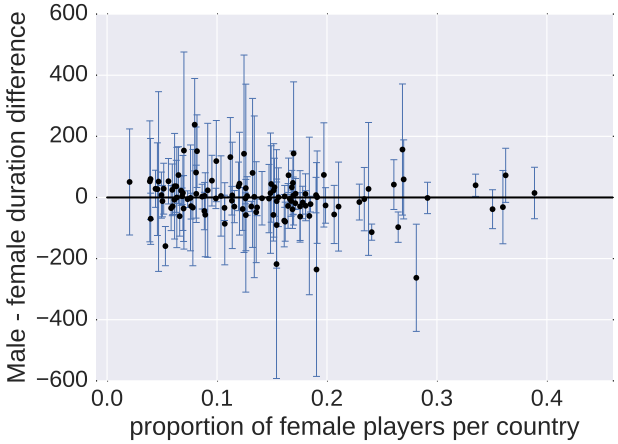

Figure 1: Difference in duration-until-inactive between young male and female players. 95% confidence intervals shown.

As shown in Figure 1, there is no discernible relation between proportion of female players in country and the difference in drop out rates for young male and female players (Pearson's $r = -0.15$, $p = 0.10$).

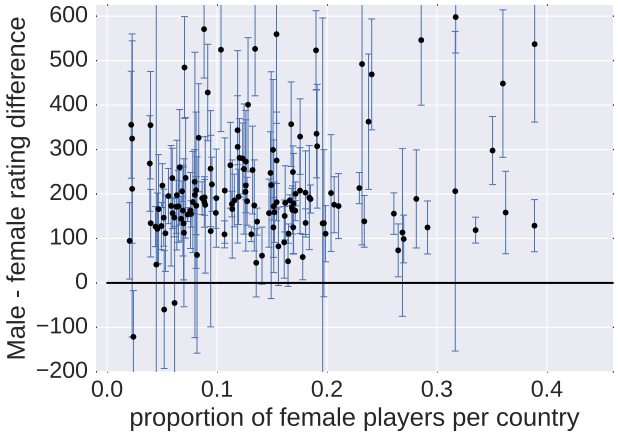

Figure 2: Difference in rating between young male and female players. 95% confidence intervals shown.

As shown in Figure 2, there is not a strong relation between proportion of female players in country and the difference in rating for young male and female players. If anything, countries with a higher proportion of female players see a larger disparity between male and female ratings (Pearson's $r = 0.21$, $p < 0.012$, compare to Chabris and Glickman, 2006, figure 4).

Differences in learning rate

So far our analysis has been restricted to using a player's best rating, but inspecting the change in players' ratings over time also allows the investigation of certain questions. Especially for younger players, we would expect performance to improve with practice, and so be reflected in an increase in rating over time. Previous research has asked if performance at chess is affected by gender, it is also possible to ask if learning is affected by gender. Anything that affects learning will have an influence on performance, by necessity.

Due to uncertainty over the precise model which best fits the learning curve (Gaschler et al., 2014; Gallistel et al., 2004) we fitted change in performance with a simple linear regression for each player, using their Elo rating at each time point. We restrict our analysis here again to young players (born 1987–2007) and to those who played at least 10 games during the period covered by the data.

The average slope was higher for males (0.101/day, n=21263) than for females (0.064/day, $n = 5265$); a highly significant difference $t(26526) = 15.95, p < 0.00001$. For reference, this represents a modest average yearly increment of 36.9 versus 23.4 Elo points.

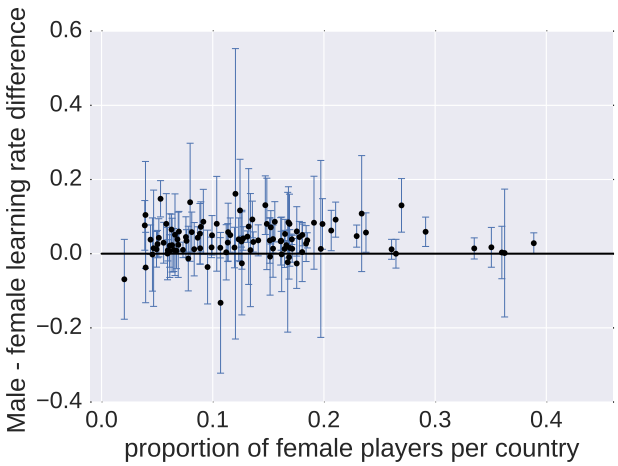

Figure 3: Difference in average slope of linear fit to ratings over time between male and female players. 95% confidence intervals shown but not visible at this resolution.

Figure 3 shows how the differing proportion of female players relates to the average difference in slope between male and female players. As with drop-out, there is no clear relation between the proportion of female players and how young female players learn with respect to young male players (Pearson's $r = 0.07$, $p = 0.48$), although we note that as with the Chabris and Glickman (2006) analysis of US zip codes, those countries with the highest proportion of female players ($> 30\%$) are conspicuous in showing no strong evidence of a difference between learning rates for young male and female players.

Differences in by-game performance

Our data also allows us to look at how individual game performance is affected by player characteristics. Figure 4 shows observed relationship between rating difference (rating of player playing White − rating of player playing Black) against outcome (coding a win for player playing White as 1; win for player playing Black as 0; draw as 0.5).

As we expect, there is a clear relationship between the relative player ratings and game outcome. Note that at around 0 difference in player ratings the average outcome is above 0.5 — showing, as is widely known, that the White player has an advantage.

We can ask how this relationship changes for higher and lower rated competitors. To do this we split the games into five groups (quintiles) according to the average rating of the two players, so that the 1st quintile contains games in which the average rating of the two players is in the top 20% of all games, and so on. To highlight how quintile affects outcome, it is helpful to

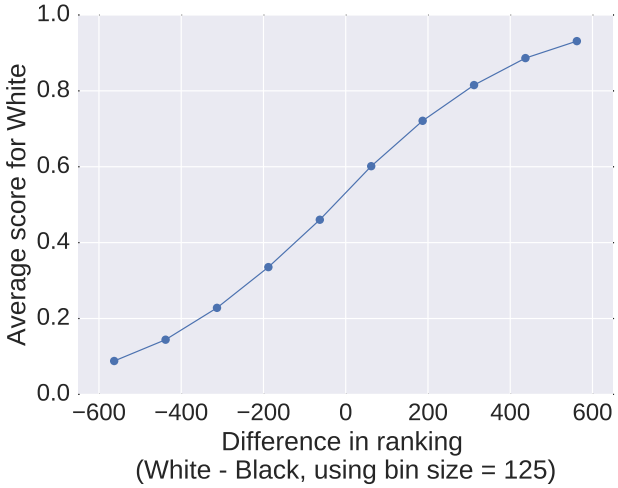

Figure 4: Difference in player rating against average game outcome (6,641,158 games). 95% confidence intervals shown.

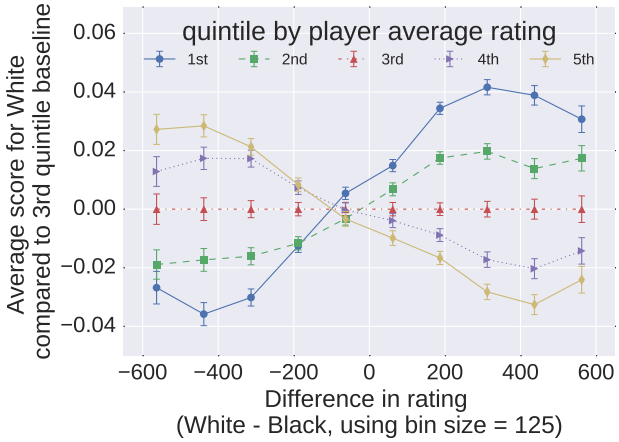

Figure 5: Effect of player average rating on game outcome (6,641,158 games). 95% confidence intervals shown.

'normalise' outcome against the effect of rating difference. To do this we plot the relative change in outcome compared to some baseline. In this case, the baseline we use will be the middle (3rd) quintile. The result is shown in Figure 5.

Figure 5 shows the variation around the basic pattern that is shown in Figure 4 which occurs as you move from the games between the lowest rated players (5th quintile) to the highest rated players (1st quintile). For the two higher quintiles, the curve is above the 3rd quintile when White's rating exceed Black's, and below that when Black's exceeds White's. This shows that even for the same absolute difference in rating, when the player's ratings are, on average, higher, the probability of the higher rated player winning is greater. The opposite is true for lower rated players – the same absolute difference in rating is less predictive of outcome. Alternatively put, the higher rated players can make more advantage of any absolute rating difference (note also that the effect shows at the 0 rating difference point, meaning that the probability of a White win when the players are evenly matched moves further away from 50% as the players' ratings goes up – higher rated players can make more of a small advantage).

Finally, we can look to see if gender affects the results. Previously female players have been shown to be at a relative disadvantage when they believed they were playing men (Maass et al., 2008), a finding which is in line with the literature on 'stereotype threat', whereby highlighting a person's membership of a minority group can affect their performance to be in line with stereotypes associated with that group (in this case, the stereotype that women are not as good at chess). In international chess the gender of your opponent is highly obvious. We code all the games in our data set according to whether they are played between two men ('MM'), two women ('FF') or mixed gender pairings, with a woman playing White ('FM') or Black ('MF'). A previous observational report of stereotype threat in chess (Rothgerber & Wolsiefer, 2014) suggested that this effect would be most likely in "challenging situations" and when playing someone of a higher grade. International chess tournaments are certainly challenging, and our previous analysis is suitable for showing how any effect changes with player rating relative to their opponent.

As with the analysis by player average rating, we normalise the outcome by comparing our results for player rating difference against a baseline. In this case we choose a baseline of the games when two men played ('MM'). A stereotype threat effect should reduce the probability of a woman winning when she plays a man, compared to when a man plays a man or a woman plays a woman. Graphically, this should appear as a lower curve for the 'FM' group (where White is a woman), and a higher curve for the 'MF' group (where White is a man). In particular, following Rothgerber's suggestions, we would expect that this effect would manifest most strongly when a woman plays a superior opponent (so in the negative portion of x-axis for the 'FM' group and the positive portion of x-axis for the 'MF' group). The results are shown in Figure 6

As can be seen in Figure 6, there is no stereotype threat effect observable in international chess. If anything, the opposite pattern is found – both mixed gender pairs show a relatively enhanced probability of White winning when White is the lower rated player, and a relatively diminished probability of White winning when

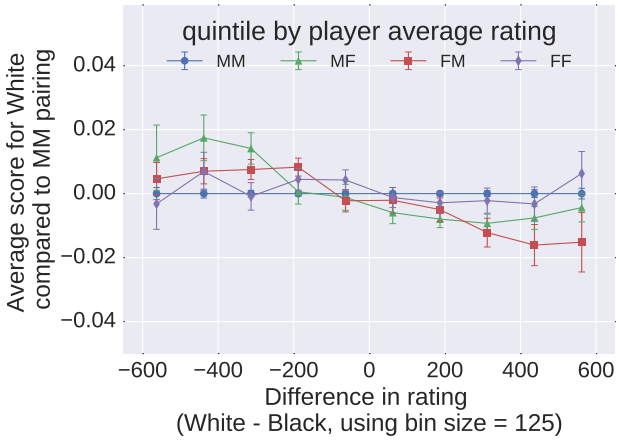

Figure 6: How player gender pairing affects game outcome (6,624,273 games). 95% confidence intervals shown.

White is the higher rated player, regardless of whether White is a man or a women. Recall that a stereotype effect should be the opposite of that seen here for the 'FM' and 'MF' pairings — diminished chance of White winning when White is a woman and the lower rated player (the 'FM' curve should move into the bottom left quadrant, which it doesn't), and enhanced chance of White winning when White is a man and the higher rated player (the 'MF curve should move into the top right quadrant, which it doesn't). By comparing this effect to Figure 4 you can see that it is the same as the effect of lower rated player-pairs overall (e.g. the 4th and 5th quintiles by pair average rating).

Discussion

We replicate and extend previous analyses of differences between male and female chess players. Like Chabris and Glickman we find no support for the idea that differential drop-out might explain sex differences in achievement in the sport. Those authors concluded that social context may be an important factor, based on a by-state analysis of US players. Our analysis used an international sample and examines by-country differences in proportion of female players. We find no evidence that this proportion influences the differences between men and women in ratings, drop-out or learning rate. Our analysis uses the international population of players, so we might expect greater cultural variation than between US zip codes (as in Chabris and Glickman's analysis). Further, our data allows us to explicitly test the idea of stereotype threat, one candidate mechanism by which social context may effect performance. This has not been done before for chess. We find no support for this phenomenon, contrary to previously published reports. We note that this is consistent with other studies of stereotype threat in high-stakes real-world settings (Stricker, 2008).

We use a large population for our analysis, rather than a sample of tens or hundreds (Rothgerber & Wolsiefer, 2014; Maass et al., 2008). It may be that the older age of our sample, the higher playing standard and/or the greater pressure of international competition induces a professionalism among players that protects against stereotype threat. Although this gives our results a strong validity in terms of the population of FIDE rated chess players, it does mean that we must recognise the unusually highly rated nature of our population compared to those used in some other studies of chess players, and particularly of younger chess players. Working with very large datasets introduces some new opportunities for the cognitive scientist (Stafford & Dewar, 2014; Stafford & Haasnoot, under review). Observational studies, however large, necessarily have reduced power of causal inference compared to experimental studies. Counterbalancing this is undeniable relevance of any phenomenon observable real-world data such as that used here.

The question of the under-representation of women in chess remains unsolved, we have merely provided evidence that stereotype threat is an unlikely mechanism for sustaining any difference in male-female ratings once players have achieved a standard that allows them to hold a FIDE rating. Some researchers (Bilalić et al., 2009; Charness & Gerchak, 1996) suggest that the gender difference at the top of the distribution is a natural consequence of different participation rates — in other words, that the low number of women in the highest echelons of chess is the simple result of the much larger number of men in the population of chess players from which the best players are drawn. It is certainly a problem that analysis of rated players limits the conclusions that can be drawn because we are in effect only looking at a subset of all possible players (Vaci et al., 2014). From this perspective the difference in participation between men and women in chess itself may be the primary factor to be explained, rather than any difference in ratings or maximal achievement (which may be explained sufficiently by differential participation).

Recently, chess has been a focus for large scale analytics. (Howard, 2006; Chassy & Gobet, 2015; Leone et al., 2014), and we see this study as part of that trend. Future work with this data has great potential for investigating differences in change in expertise, as well as performance. Highly relevant is the observation that tournament games are actually the most significant events in rating improvement (Howard, 2012, 2013). Future work on chess is sure to focus on within-game dynamics as well as the dynamics of ratings. To the end of promoting integration of existing work and further exploration of the rich data provided by FIDE chess ratings we are

happy to make the analysis scripts for the current analysis available immediately at `https://osf.io/aeksv`), and the full data available in time.

References

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 of London B: Biological Sciences*, *276*(1659), 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*(12), 1040–1046.

Charness, N. (1992). The impact of chess research on cognitive science. *Psychological research*, *54*(1), 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*, 46–51.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, *4*(1), 55–81.

Chassy, P., & Gobet, F. (2015). Risk taking in adversarial situations: Civilization differences in chess experts. *Cognition*, *141*, 36–40.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? a meta-analysis. *Journal of school psychology*, *53*(1), 25–44.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: implications of a quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(36), 13124–13131.

Ganley, C. M., Mingle, L. A., Ryan, A. M., Ryan, K., Vasilyeva, M., & Perry, M. (2013). An examination of stereotype threat effects on girls' mathematics performance. *Developmental psychology*, *49*(10), 1886.

Gaschler, R., Progscha, J., Smallbone, K., Ram, N., & Bilalić, M. (2014). Playing off the curve-testing quantitative predictions of skill acquisition theories in development of chess performance. *Frontiers in psychology*, *5*.

Howard, R. W. (2006). A complete database of international chess players and chess performance ratings for varied longitudinal studies. *Behavior research methods*, *38*(4), 698–703.

Howard, R. W. (2012). Longitudinal effects of different types of practice on the development of chess expertise. *Applied Cognitive Psychology*, *26*(3), 359–369.

Howard, R. W. (2013). Practice other than playing games apparently has only a modest role in the development of chess expertise. *British journal of psychology*, *104*(1), 39–56.

Leone, M. J., Slezak, D. F., Cecchi, G. A., & Sigman, M. (2014). The geometry of expertise. *Frontiers in psychology*, *5*.

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*(2), 231–245.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Chess-playing programs and the problem of complexity. *IBM Journal of Research and Development*, *2*(4), 320–335.

Rothgerber, H., & Wolsiefer, K. (2014). A naturalistic study of stereotype threat in young female chess players. *Group Processes & Intergroup Relations*, *17*(1), 79–90.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology*, *35*(1), 4–28.

Stafford, T., & Dewar, M. (2014). Tracing the trajectory of skill learning with a very large sample of online game players. *Psychological science*, *25*(2), 511–518.

Stafford, T., & Haasnoot, E. (under review). Confirming and quantifying sleep consolidation in skill learning: a field study using an online game.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of african americans. *Journal of personality and social psychology*, *69*(5), 797.

Stoet, G., & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, *16*(1), 93–102.

Stricker, L. J. (2008). The challenge of stereotype threat for the testing community. In *Presidential address to the division of evaluation, measurement, and statistics. 2007 american educational research association annual meeting.*

Vaci, N., Gula, B., & Bilalić, M. (2014). Restricting range restricts conclusions. *Frontiers in psychology*, *5*.

Acknowledgements

The "Sonas 92" dataset used in this analysis was prepared by Jeff Sonas of Sonas Consulting (jjeff@sonasconsulting.com). Without his generosity and advice this study would not have been possible. Thanks also to Alberto Ara, Stephen Want and to three anonymous reviewers and the editor for their detailed feedback.