Female chess players outperform expectations when playing men

Tom Stafford
Department of Psychology, University of Sheffield
Sheffield, S1 2LT, United Kingdom

Abstract

"Stereotype threat" has been offered as a potential explanation of differential performance between men and women in some cognitive domains. Questions remain about the reliability and generality of the phenomenon. Previous studies have found that stereotype threat is activated in female chess players when they are matched against male players. I use data from over 5.5 million games of international tournament chess and find no evidence of a stereotype threat effect. In fact women players outperform expectations when playing men. Further analysis shows no influence of degree of challenge, nor of player age, nor of prevalence of female role models in national chess leagues on differences in performance when women play men versus when they play women. Though this analysis contradicts one specific mechanism of influence of gender stereotypes, the persistent differences between male and female players suggest that systematic factors do exist and remain to be uncovered.

Introduction

Gender differences in cognition

The topic of sex differences in cognition evokes strong reactions, including accusations of sexism, essentialism, political correctness or the denial of human nature (Fine, 2010; Pinker, 2003; Halpern et al., 2007). As psychological scientists, we know that the reality of any observed sex difference is one issue, and the causal pathways leading to any observed sex differences is another. Simply put, we cannot infer from a real difference between the sexes that this difference is inevitable, immutable or inborn (Mameli & Bateson, 2011; Griffiths, Machery, & Linquist, 2009). To diagnose any difference as innate we would need clarity on the mechanisms producing that difference; mechanisms which potentially span genetic

---

In press at Psychological Science. This is the accepted, Author produced, text. This research was supported in part by Leverhulme Trust Project Grant on Bias and Blame (RPG-2013-326).
The author declares no conflicts of interest with respect to the authorship or the publication of this article.
Correspondence concerning this article should be addressed to Tom Stafford, Department of Psychology, University of Sheffield, Sheffield, S1 2LT, UK Contact: t.stafford@sheffield.ac.uk

inheritance, developmental influences, the interactions of genetics with the environment and the ongoing influences of adult society on cognitive performance.

Possible environmental influences on sex differences in cognition come in different flavours. There are those which affect the development of skills and preferences across the lifespan; those which, through cultural ideas of gender, affect others' judgement; and those which affect our own behaviour. Demonstrating the reality, or lack of reality, of one potential mechanism doesn't speak to the reality of the others. Nevertheless, if we are to win an accurate account of the emergence of sex differences in cognition each potential mechanism needs to tested and verified.

Stereotype Threat

One notable psychological phenomenon which can influence performance on cognitive tests is that of 'stereotype threat', whereby an individual's awareness of a negative stereotype influences their performance (Inzlicht & Schmader, 2012). This was originally proposed for African Americans and intelligence test performance (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, Steele, & Quinn, 1999).

Stereotype threat has been offered as part explanation for sex differences on cognitive tasks (e.g. Fine, 2010). The suggested mechanisms for the effect are plausible – increased anxiety, performance monitoring and/or negative thought suppression which creates additional working memory load (Beilock, Rydell, & McConnell, 2007; Schmader, Johns, & Forbes, 2008) – but it is important to recognise that a) establishing the reality of even a true effect in laboratory conditions is not straightforward and b) regardless of the reality of stereotype effect there are other reasons for sex differentiated performance (cf Sackett, Hardison, & Cullen, 2004).

Stereotype Threat & publication bias

Recent analyses have suggested that the literature on stereotype threat suffers from publication bias (Flore & Wicherts, 2015; Stricker, 2008; Ganley et al., 2013; Doyle & Voyer, 2016). If studies reporting a positive effect are more likely to be published then this will exaggerate the true size and robustness of stereotype threat. Despite this, other meta-analyses have attested to the reality of the effect (Lamont, Swift, & Abrams, 2015; Nguyen & Ryan, 2008; Doyle & Voyer, 2016). One 2016 review states "Stereotype-threat effects are generally robust, with moderate to small effect size" (Spencer, Logel, & Davies, 2016, p.418).

An approach which may complement experimental studies of stereotype threat is to investigate its impact on cognitive performance outside the lab. This also makes it possible to assess the importance of stereotype threat amidst the myriad influences on behaviour in daily life. Field studies make it possible to access vastly increased statistical power over typical experimental studies.

Chess

Chess has an illustrious history within cognitive science (Newell, Shaw, & Simon, 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 is their cognitive capacity in planning, and their ability to reason 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.

Chess is heavily male dominated both in terms of the absolute number of male players and in terms of male representation among the best chess players. The stereotypical chess grandmaster is undeniably a man, and – due to the face-to-face nature of tournament play – it is difficult for gender not to be salient when a female chess player competes with man. If the stereotype threat phenomenon is robust and general then we should be able, with the right analysis, to observe it operating in chess.

Previous research has explored a number of possible competing explanations for the under-representation of women in chess (Chabris & Glickman, 2006; Bilalić, Smallbone, McLeod, & Gobet, 2009). In chess, both observational (Rothgerber & Wolsiefer, 2014) and experimental studies (Maass, D'Ettole, & Cadinu, 2008) appear to confirm the existence of stereotype threat. Rothgerber and Wolsiefer (2014), looking at 219 female chess players, report (p.79) that "Stereotype threat susceptibility was most pronounced in contexts that could be considered challenging: when playing a strong or moderate opponent". Maass and colleagues (2008) ran a study using internet chess where the perceived gender of opponents was experimentally manipulated with 42 female participants. When they believed they were playing an opponent of the opposite gender female players were less likely to win. If these findings apply widely to chess performance they have the potential to systematically undermine the performance of female players.

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 achievements in chess, but this 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. Apart from their importance to understanding chess, these data also provide an opportunity to interrogate a real world domain for the reality, or not, of the effects of gender on performance, including any stereotype threat effects.

Data and method

The data comprise records of $9,662,202$ games of standard tournament chess, played between January 2008 and August 2015. There are also records of $461,637$ FIDE rated players ($56,474$, 12.2%, women. The average birth year for these players was 1983, with an average age of 31.5 years (standard deviation 19.28) at the time the games were played.

In recent years an increasing number of younger players have joined the rating system, expanding the number of rated players and lowering the average rating.

For each player the data consists of a unique player ID, date of birth, gender, nationality and 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 ratings 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 between any two rated players and as a way of ranking any player against the historical community of all players contained within the system. Because of this it is possible to compare players who have never played, and may not even be contemporaneous.

When analysing game outcomes, I analysed only games of standard tournament chess between players who both possessed FIDE ratings and were active during the 92 month period for which I have data. This left $5,558,110$ games, from $150,977$ male players and $16,158$ female players.

To investigate the possibility of stereotype threat, I compared women's performance when playing against a man, and when playing against another woman, to the expected outcome from when a man plays against a man. An advantage of chess is that we are able to precisely gauge the challenge presented by individual games to each player, via comparison of player Elo ratings. As well as looking at the difference in outcome by gender of opponent, I also investigated whether player age and prevalence of other female chess players affects outcome.

Analysis scripts are available at `https://osf.io/aeksv`, as well as a sample of 5% of players represented in the full game-by-game dataset. For commercial reasons this full raw dataset is not available at the point of writing. I do provide the full (summary) data which supports the key analysis presented here. Whilst I acknowledge that it not appropriate to use null hypothesis significance testing (NHST) to guide interpretation of my data, I do report the p-values of standard null hypothesis tests in places. This is to ease comparison for readers familiar with NHST; such readers will note that no p-values I report are marginal. Everything which might be considered 'significant' is extremely significant, everything which is not significant is resolutely not significant.

Results

Differences in ratings

In the player record, the average FIDE rating of men was 2070 (standard deviation 186), and for women 1978 (standard deviation 195). This difference is statistically significant, $t(460345) = 35.51, p < 0.001$. For reference, a rating above 2500 is associated with Chess Grandmaster level (at this level 98.9% of players in these data were male). The ratio of the standard deviations of ratings for women to men was 1.05, showing higher variability in women's ratings (as with Chabris & Glickman, 2006).

Differences in by-game performance

These data also allow us to look at how individual game performance is affected by player characteristics. The Elo system provides a predicted outcome for any match based

on the rating difference between the two players. Figure 1 shows the observed relationship between rating difference and game outcome for games featuring men only. The rating difference is the rating of the player playing White minus the rating of player playing Black. For outcome, a win for the player playing White is coded as 1, a win for the player playing Black as 0, a draw as 0.5.

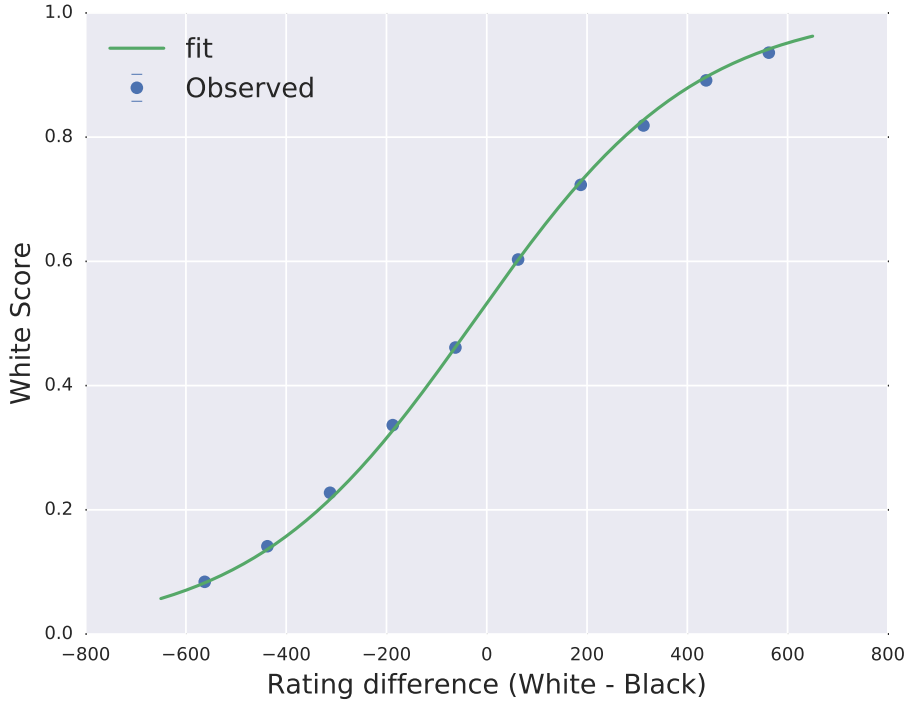

Figure 1. Difference in player rating against average game outcome (4,659,239 games from male-only competitors). 95% confidence intervals shown but not visible at this resolution.

As expected, there is a clear relationship between the relative player rating 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. In order to subsequently calculate predicted outcome for any rating difference I fitted a logistic function to the observed data, for games featuring male players only.

I coded all the games in the 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'). The difference in rating allows us to precisely operationalize the challenge presented by each game. If stereotype threat is most likely to manifest in "challenging situations" (Rothgerber & Wolsiefer, 2014) then this would be when playing someone of a higher rating. International chess tournaments are certainly challenging, and the difference in Elo rating allows us to gauge precisely the challenge presented within any particular pairing.

Using the function derived from MM games (see above), I calculated the difference from predicted outcome for every game. Calculating the average difference from expected outcome for both FM and MF games (reversing the sign for MF games, so that, for both

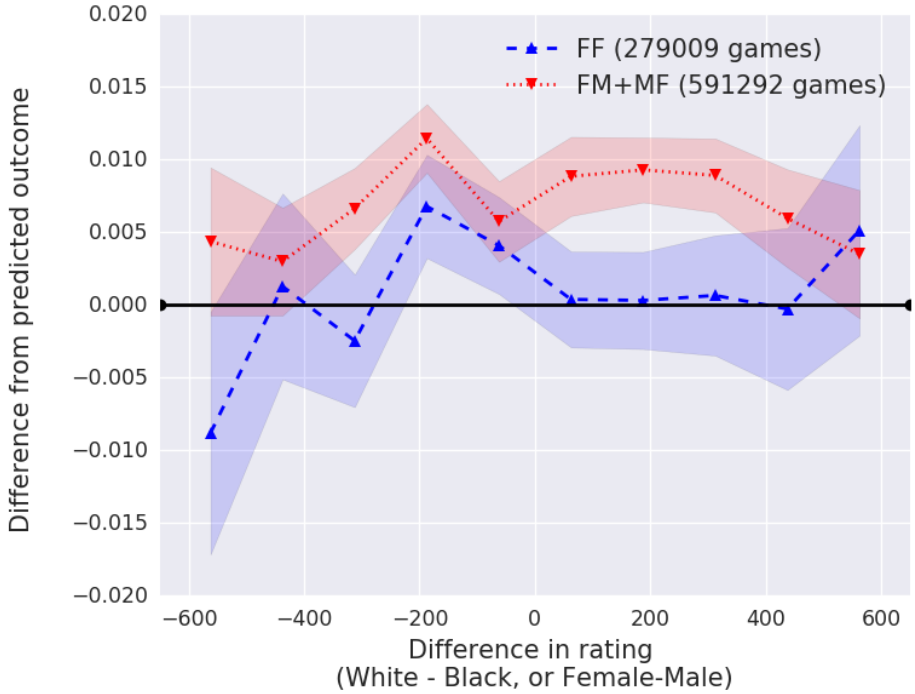

Figure 2. How player gender pairing affects game outcome (5,558,110 games total). Baseline expectation, from analysis of MM games, shown in black. Shaded regions show 95% confidence intervals.

FM and MF games a negative number represents a worse than expected outcome for the female player) tells us how female players perform, relative to expectations, when facing a male player. I did this across the range of possible rating differences for players, using a binning width of 125 Elo points. The results are shown in Figure 2. Note that this figure shows the variation around the function shown in Figure 1: by removing the variation due to rating difference it allows us to focus on the other factors which influence game outcome.

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 (the baseline) or when a woman plays a woman ('FF'). Graphically, this should appear as a lower curve for the 'FM+MF' group. In particular 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).

The opposite is the case – female players outperformed expectations when facing male players, across the whole range of rating differences. Note the scale on this figure: a difference of 0.01 from the predicted outcome is a 1% increment in the probability of winning a game, or one extra win in one hundred games, compared to the baseline expectation. The observed average for mixed pairs was *above* the average for same-sex pairs (both MM and FF). This is the *opposite* of a stereotype threat effect, reflecting a lift in female chess players' performance when playing a male opponent, above their rating-predicted performance.

Another angle on these data is to look for 'upsets' – games with a strong favourite

Table 1
Regression results predicting size of stereotype threat effect across individuals

	coef	**std err**	**t**	**P> |t|**	**[95.0% Conf. Int.]**
Intercept	0.0065	0.631	0.010	0.992	[-1.230, 1.243]
BirthYear	1.489e-05	0.000	0.047	0.963	[-0.001, 0.001]
Fprop	-6.3471	4.538	-1.399	0.162	[-15.242, 2.547]
BirthYear:Fprop	0.0031	0.002	1.364	0.173	[-0.001, 0.008]

(based on Elo ratings) in which the favourite lost[1]. I took a rating difference of 500 Elo points as an arbitrary threshold for defining games with a strong favourite (note from Figure 1 that this rating difference predicts a victory for the stronger player with ~95% probability). Of such games, between male players ('MM') 3.18% resulted in upsets, and between female players ('FF') 2.83% resulted in upsets. The number of upsets was higher for mixed pairs ('FM' or 'MF' pairs, p< 0.0001 using Fisher's Exact Test). Of those games between mixed pairs where the female player was overmatched, upsets occurred 3.70% of the time. Of those games between mixed pairs where the male player was overmatched, upsets occurred 3.51% of the time. Although upsets are numerically more likely to favour the female player this is not statistically significant (p= 0.562 using Fisher's Exact Test).

To confirm the 'negative stereotype threat' pattern, I switched to using the individual players as a base unit of analysis rather than games. The advantage of this is that it better controls for confounding factors, such as a change in both the rating and gender proportion of players across time (e.g. that more women and more weaker players are entering the international chess ratings). Using each player as her own control, I calculated the difference between actual game outcome and expected game outcome given the relative rating of the players, for both games where she played another woman ('vsF') and for those where she played a man ('vsM').

Over all Female players the average stereotype threat effect was 0.014, which is significantly different from zero (95% CI 0.010,0.017), and which was again a reverse of the classic stereotype threat effect[2]. Figures 3 and 4 show that there is no systematic variation in the size of the stereotype threat by proportion of female players in different national chess leagues, or by birth year of the player[3]. To confirm this I fitted a regression model predicting the size of the stereotype threat effect for each female player from player birth year and proportion of female players in their country of origin ('Fprop'), as well as the interaction. Estimates of the influence of these factors all overlapped with zero, as shown in Table 1, based on an overall model which explained little of the variance ($R^2 = 0.003$, $F(3, 12687) = 13.72, p < 0.001$).

[1] I thank the reviewers for this suggestion.

[2] Additionally the female advantage for mixed-sex games was confirmed by multilevel modelling combining player and game-level factors. Details are in the online materials `https://osf.io/aeksv`

[3] Additional analysis confirmed that there was significant spread in the proportion of games female players had played against other women. Across this range there was some evidence that the female advantage when playing men did not hold for players who had played greater than 65% of their games against other women. Details are in the online materials `https://osf.io/aeksv`

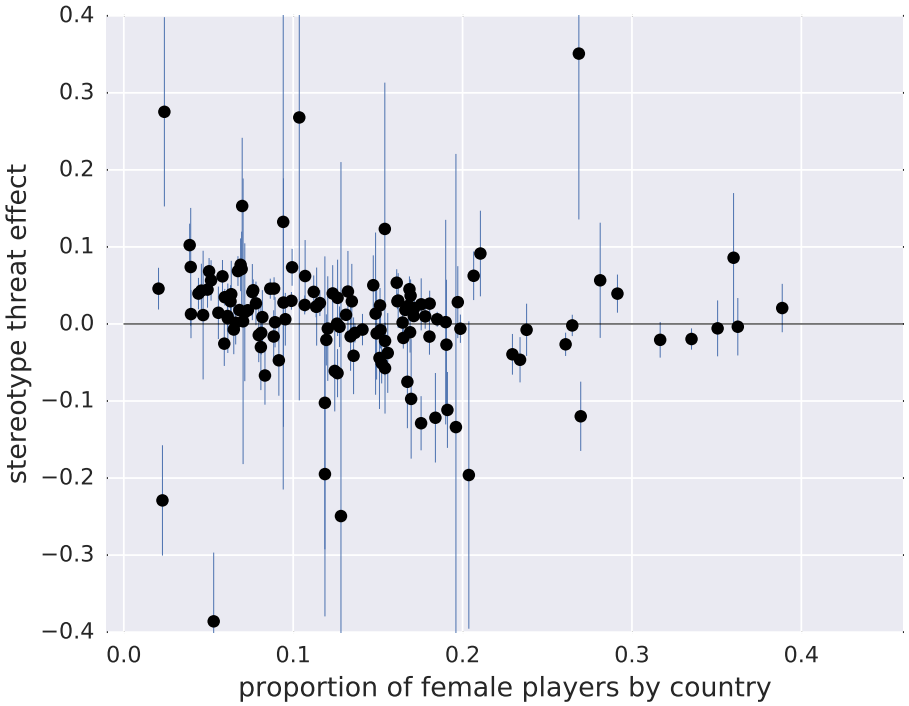

Figure 3. Stereotype threat effect, average by country. 95% confidence intervals shown.

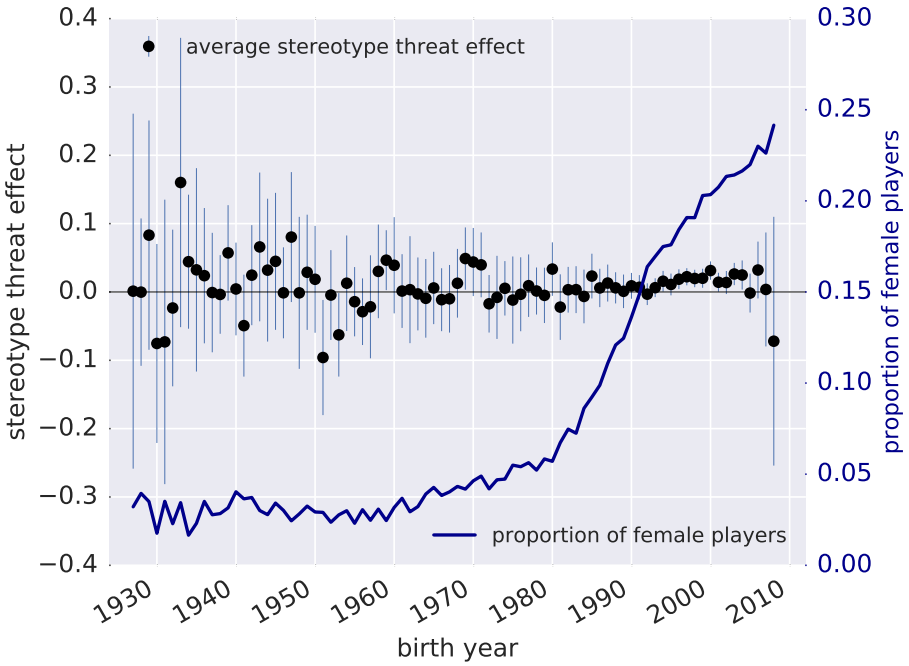

Figure 4. Stereotype threat effect, average by birth year (dots, left axis). 95% confidence intervals shown. Right axis shows proportion of female players in dataset for that birth year (continuous line).

Discussion

Our data allow us to explicitly test for the operation of stereotype threat, in this particular domain, as one candidate mechanism by which social context may affect performance. Contrary to previously published reports (which used smaller samples, and a narrower range of abilities; Rothgerber & Wolsiefer, 2014; Maass et al., 2008), stereotype threat does not appear to affect chess at this level. Female players, far from suffering a stereotype threat, display a boost in performance when playing men compared to playing women.

I note that tournament chess is a different task from those which were used to establish the stereotype threat phenomenon. In particular, for any rated player, chess will be a highly familiar task and task novelty has been shown to interact with stereotype threat via arousal (O'Brien & Crandall, 2003; Ben-Zeev, Fein, & Inzlicht, 2005). So, paradoxically, it could be that stereotype related anxiety raises performance, protecting against 'threat' effects in these data[4]. It may be that the older age of the sample, the higher playing standard and/or the greater pressure of international competition induces a professionalism among players that also protects against stereotype threat.

If stereotypes are not negatively affecting female players' performance against male players in chess, what mechanisms are producing the difference for mixed pairs compared to single-sex pairs? One plausible mechanism is a degree of male *under-performance* rather than female over-performance. This could be due to male underestimation of female opponents, misplaced chivalry, or 'choking' due the ego-threat of being beaten by a women (Baumeister, 1984). I note a recent analysis of grand-slam tennis which suggests that men may be particularly vulnerable to choking (Cohen-Zada, Krumer, Rosenboim, & Shapir, 2017). The analysis of upsets supports this idea. It seems more likely that any psychological factor would cause a favourite to throw a game with an unwise move, than that an underdog would be able to play a whole game at the level required to overcome a large rating difference disadvantage[5].

The question of the under-representation of women in chess remains unsolved. I 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, Gula, & Bilalić, 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, Slezak, Cecchi, & Sigman, 2014; Vaci & Bilalić, 2016), and I see this

[4]I thank Steve Spencer for pointing this possibility out.

[5]I thank Steve Spencer and Roy Baumeister for suggesting this analysis.

study as part of that trend. Future work with these data has great potential for investigating differences in change in expertise, as well as performance. 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 I am happy to make the analysis scripts available immediately at `https://osf.io/aeksv`, along with a subset of the data and with full summary data supporting the regression analysis, and with the full raw, game by game, data available in time.

The current study shows that the stereotype threat phenomenon has boundary conditions. A proviso is that the analysis requires one to accept the operationalisation used here – that of contrasting games where female players play male opponents with those where female players play female opponents. It may be, of course, that stereotype threat affects female chess players in different ways. Such a broader view of the phenomenon has many advantages (Lewis & Sekaquaptewa, 2016). Nonetheless, in the current study we looked, with a very highly powered statistical lens, at female performance in a highly gender stereotyped domain, using the advantage of a large sample to look in exactly the place where, from a reading of the literature, we would expect to find stereotype threat if it existed (younger players, and female players relatively deprived of role models). The evidence suggests no stereotype threat effect, with – in fact – a small effect in the opposite direction.

Other studies of stereotype threat in high-stakes real-world settings are not consistent (Stricker & Ward, 2004; Stricker, 2008; Walton & Spencer, 2009). For example, one field study failed to show the stereotype threat effect, showing that gender priming could lift girls' scores on educational tests (Wei, 2012). Another field study replicated the effect in the original domain (black students and math performance), but failed to find evidence of the effect in the domain of gender (Stricker, Rock, & Bridgeman, 2015). Obviously there is significant work to do on defining the conditions under which we can expect stereotype threat to manifest.

Working with very large datasets introduces some new opportunities for the cognitive scientist (Stafford & Dewar, 2014; Goldstone & Lupyan, 2016). Experimental and observational studies complement each other. They have different advantages, such as allowing strong causal inference for experimental studies, or more easily allowing high statistical power for observational studies. They also train our scientific imaginations in different ways. Experimental studies encourage us to focus on isolated causal factors. Observational studies encourage us to see all factors in the context of other factors (Stafford & Haasnoot, 2017). Observing a phenomenon 'in the wild' provides a strong validation of the generality and robustness of an effect. Lab studies of stereotype threat have illustrated one mechanism by which social attitudes may create discrimination. This study of one social attitude in one domain – gender stereotypes in chess – does nothing to disprove the reality of discrimination generally, but it does suggest that this one mechanism, stereotype threat, may be more limited in its applicability than one might conclude from reading the experimental literature alone.

References

Baumeister, R. F. (1984). Choking under pressure: self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of personality and social psychol-*

ogy, *46*(3), 610–620.

Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General*, *136*(2), 256.

Ben-Zeev, T., Fein, S., & Inzlicht, M. (2005, March). Arousal and stereotype threat. *Journal of Experimental Social Psychology*, *41*(2), 174–181. doi: 10.1016/j.jesp.2003.11.007

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*, *7*(1), 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.

Cohen-Zada, D., Krumer, A., Rosenboim, M., & Shapir, O. M. (2017). Choking under pressure and gender: Evidence from professional tennis. *Journal of Economic Psychology*, *61*, 176–190. doi: 10.1016/j.joep.2017.04.005

Doyle, R. A., & Voyer, D. (2016). Stereotype manipulation effects on math and spatial test performance: A meta-analysis. *Learning and Individual Differences*, *47*, 103–116.

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

Fine, C. (2010). *Delusions of gender: The real science behind sex differences*. Icon.

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.

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.

Goldstone, R. L., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in cognitive science*, *8*(3), 548–568.

Griffiths, P., Machery, E., & Linquist, S. (2009). The vernacular concept of innateness. *Mind & Language*, *24*(5), 605–630.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological science in the public interest*, *8*(1), 1–51.

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.

Inzlicht, M., & Schmader, T. (2012). *Stereotype threat: Theory, process, and application.* Oxford University Press.

Lamont, R. A., Swift, H. J., & Abrams, D. (2015). A review and meta-analysis of age-based stereotype threat: Negative stereotypes, not facts, do the damage. *Psychology and Aging*, *30*, 180–193.

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

Lewis, N. A., & Sekaquaptewa, D. (2016). Beyond test performance: A broader view of stereotype threat. *Current Opinion in Psychology*, *11*, 40–43.

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.

Mameli, M., & Bateson, P. (2011). An evaluation of the concept of innateness. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *366*(1563), 436–443.

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.

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*, 1314–1334.

O'Brien, L. T., & Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin*, *29*(6), 782–789. doi: 10.1177/0146167203029006010

Pinker, S. (2003). *The blank slate: The modern denial of human nature.* Penguin.

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

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American-White differences on cognitive tests. *American Psychologist*, *59*(1), 7.

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological review*, *115*(2), 336.

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual review of psychology*, *67*, 415–437.

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. (2017). Testing sleep consolidation in skill learning: a field study using an online game. *Topics in Cognitive Science*, *9*, 485–496.

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.

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.*

Stricker, L. J., Rock, D. A., & Bridgeman, B. (2015). Stereotype threat, inquiring about test takers' race and gender, and performance on low-stakes tests in a large-scale assessment. *ETS Research Report Series*, *2015*(1), 1–12.

Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance1. *Journal of Applied Social Psychology*, *34*(4), 665–693.

Vaci, N., & Bilalić, M. (2016). Chess databases as a research vehicle in psychology: Modeling large data. *Behavior research methods*, 1–14.

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

Walton, G. M., & Spencer, S. J. (2009). Latent Ability: Grades and Test Scores Systematically Underestimate the Intellectual Ability of Negatively Stereotyped Students. *Psychological Science*, *20*(9), 1132–1139. doi: 10.1111/j.1467-9280.2009.02417.x

Wei, T. E. (2012). Sticks, stones, words, and broken bones: New field and lab evidence on stereotype threat. *Educational Evaluation and Policy Analysis*, *34*(4), 465–488.

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 from the Cognitive Science Society conference for their detailed feedback. Tim Heaton provided advice on multilevel modelling. Eleanor Kent helped with proofreading. Reviewers Roy Baumeister and Steven Spencer helped improve the manuscript to its current form.