Team Gender Diversity in High-Stake Competitions

Redzo Mujcic and Kenan Kalayci

July 2016

Abstract

We examine the link between gender diversity and performance outcomes in team competitions with large financial stakes, using new data on more than 2,300 teams from a series of high-profile television game shows. The collected data come from three diverse field settings including an actual business environment (*The Apprentice*), a natural survival contest (*Survivor*), and a high-pressure cooking competition (*Hell's Kitchen*). Our overall results show all-female and female-dominated teams to be as successful as their male counterparts in each competition. Team success rates also depend on task type, with notable differences in win rates observed in favour of all-male teams for projects which demand creativity on *The Apprentice*. The probability of winning an intellectual challenge on *Survivor* is found to be proportional to the share of women in a team, while male-dominated teams perform better in physical challenges. Lastly, we report male and female team leaders to have equal success in high-stake business competitions, independent of the team gender mix or type of project undertaken. These findings provide some of the first evidence about the role of gender diversity in competitive field settings with unusually large payoffs.

Keywords**:** gender diversity; teams; high stakes; leadership; game shows.

1. Introduction

Despite the elimination of a gender gap in higher educational attainment and significant convergence in labour market participation rates, females are still vastly underrepresented in senior roles in corporations, academia, and the government (the "leaky pipeline" problem). This problem has previously been attributed to discrimination and bias against females (Goldin and Rouse 2000; Moss-Racusin et al. 2012), lifestyle and personal choices of women (Ceci and Williams 2011), and gender differences in human capital (Blau and Kahn 2000). An alternative explanation for the lack of diversity in senior positions is the hypothesis that females do not perform as well as males under pressure and that they are more averse to competition than men (Gneezy et al. 2003; Niederle and Vesterlund 2007; Antonovics et al. 2009; Shurchkov 2012; Buser et al. 2014; Flory et al. 2014; Sutter and Glätzle-Rützler 2015; Buser 2016).

While gender differences regarding high-pressure competitions between individuals is important, it is unclear how these differences carry over to economic outcomes given that a majority of labour production in professional settings takes place in teams (Hamilton et al. 2003; Lazear and Shaw 2007) and competition is usually between teams rather than between individuals. A handful of recent studies have explored the link between gender diversity and team outcomes in a limited range of decision settings such as corporate boards (Adams and Ferreira 2009; Ahern and Dittmar 2012), laboratory experiments (Dufwenberg and Muren 2006; Pearsall et al. 2008; Ivanova-Stenzel and Kuebler 2011), and business games involving university students (Apesteguia et al. 2012; Hoogendoorn et al. 2013). Most of these studies report mixed-gender teams to perform better than single-gender teams on average.[1]

Little is however known about the role of gender diversity in high-pressure team competitions with large financial stakes, where participants experience life-changing personal and professional career outcomes. In such situations, individual performances, team dynamics, and overall outcomes can be quite different (Baumeister 1984; Ariely et al. 2009; Beilock 2010; Kamenica 2012). That the literature is rather scarce on this topic is perhaps unsurprising given, for example, the large associated costs of conducting controlled experimental studies with substantial monetary rewards on offer.

[1] Bear and Woolley (2011) and Azmat (2014) survey the empirical and experimental literature on gender diversity in teams. Other recent studies have also examined the effects of group composition and gender diversity on outcomes, other than overall performance, such as willingness to compete in schools (Booth and Nolen (2012), expression of minority views (Amini et al. 2016), and pricing decisions in actual markets (List et al. 2016). For studies focusing on individual versus team preferences and decision-making, see e.g. Kocher and Sutter (2005); Charness and Sutter (2012); Maciejovsky et al. (2013); Balafoutas et al. (2014); Ambrus, Greiner, and Pathak (2015).

The present study empirically examines patterns in team gender composition and performance outcomes using a series of well-known and previously unexplored game show competitions with unusually high stakes: namely *The Apprentice*, *Survivor*, and *Hell's Kitchen*. In each competition, contestants are assigned into teams to compete across a range of actual tasks and projects for an opportunity to win up to one million dollars in prize money. While only a single contestant is the eventual winner, a major component of each competition, and hence contestant progression, is centred on teamwork and collective outcomes. Based on team performance, one or more individual contestants are regularly eliminated from the competition, leading to fluctuations in the size and composition of teams. Similarly, team gender composition is frequently altered when new teams are formed at the discretion of the ultimate judging panel. We use such variation in team compositions to study the association between the gender mix of more than 2,300 teams and their relative success rates.

The data we collect allow us to study the optimal gender composition of teams in three diverse field settings including an actual business environment (*The Apprentice*); a natural survival environment (*Survivor*); and a high-pressure cooking competition (*Hell's Kitchen*). The specific and varying team challenges undertaken within each of these competitions enable us to additionally look at the nature and type of tasks which are more successfully performed by some team genders than others.

To test for this, we are able to group the team tasks undertaken within each competition into separate categories according to particular and sometimes opposing skill sets. For instance, the team projects performed by contestants on *The Apprentice* can be split into three main categories including (i) tasks which predominantly involve designing and developing ideas for new products and services (e.g., creating a new mobile phone application), an activity which demands creative thinking and entrepreneurial skills from team members; (ii) projects based around buying and selling (trading) goods and services (e.g., purchasing supplies and selling lemonade on the streets of New York), where the final results are simply measured in terms of overall sales and profits made; and (iii) tasks of advertising and coordinating marketing events for existing or upcoming goods and services (e.g., facilitating an advertising campaign for Lamborghini cars), which are also judged by the total number of potential customers attracted. In a similar however less precise way, the team challenges undertaken and observed on *Survivor* can be categorized into predominately *physical* (e.g., obstacle courses, manoeuvring heavy objects, running/swimming contests) and *intellectual* tasks (e.g., solving puzzles, memory games, quizzes). Lastly, since *Hell's Kitchen* is a competition which focuses purely on cooking skills, here we are only able to split the performed team tasks by the relative level of within-competition stakes and pressure experienced by contestants; where, for example, the regular ‘dinner service’ task is much more demanding and highly scrutinized by the ultimate judge, both in terms of the quantity and quality of the final product, compared to the other team challenge performed on the show. Similar increases in within-competition stakes and psychological pressure are observable on *Survivor*, namely across the ‘reward’ and ‘immunity’ challenges; with team success in the latter ensuring progression to the next stages of the competition.

Another interesting and novel empirical angle which we are able to explore, specifically using data from *The Apprentice*, relates to the interaction between team gender composition and the sex of the project manager. That is, do some teams perform better under a male or a female leader? While male leadership still continues to be the norm in most societies and organisations (Eagly and Carli 2003; Van Vugt et al. 2008), there is a clear lack of empirical evidence about the relative performance outcomes of male versus female-headed teams, especially when the gender composition of the followers is varied (Eagly et al. 1995). To this end, we also consider the following research questions: Are male project managers more successful than female project managers in business environments with high stakes? And, are any found sex differences in leadership driven by the gender mix of the group or the type of business project undertaken?

While the present study is one of the first to consider the gender composition of teams in high-stake television competitions, the use of game show data is not new within the social sciences. Past research has turned to such natural experiments (Harrison and List 2004) to elicit a range of individual-level decisions and preferences including risk attitudes (Gertner 1993; Metrick 1995; Beetsma and Schotman 2001; Post et al. 2008), gender and race discrimination (Levitt 2004; Antonovics et al. 2005, 2009; Dilks et al. 2010), cooperation (List 2004, 2006; Belot et al. 2010; Oberholzer-Gee et al. 2010; Van den Assem et al. 2012), and strategic reasoning (Bennett and Hickman 1993; Berk et al. 1996; Tenorio and Cason 2002; Van Dolder et al. 2015). In addition to the clearly observable and well-defined choice problems, a key attraction and methodological feature of such settings is the substantial size of the monetary payoffs. Game shows also differ from laboratory and other experimental environments in terms of participant selection, task familiarity, and public scrutiny levels. Our study utilises game show data in a different way than previously done: by focusing on team, rather than individual participant, outcomes.

Previous empirical work on gender diversity and team performance outcomes has mostly focused on a limited range of laboratory experiments and business games involving student populations (e.g., Dufwenberg and Muren 2006; Apesteguia et al. 2012; Hoogendoorn et al. 2013). Although some of the already considered payoffs are not trivial, such as improved entrepreneurial career prospects upon graduation (e.g., Hoogendoorn et al. 2013), both the pecuniary and non-pecuniary incentives found in our present set of contexts are much greater (e.g., prize money of up to one million US dollars, as well as worldwide publicity). Such rewards also arguably better reflect the pressures of real-life workplaces than, say, the common experimental laboratory.

Similar to most experimental work in the social sciences, the selection of game show participants, and its potential effects on the generalisability of any empirical findings, is often questioned. That is, random individuals first audition for the show; producers then select the participants which they wish to have on the show; and finally participants are matched with their initial opponents or allocated into teams. While the underlying selection procedures and television producer preferences are unknown, one objective feature of game show settings is that the resulting demographic characteristics of participants approximate those of the general (middle-class) population much more closely than in most lab or field studies.[2] This element allows us to consider more representative teams of individuals than previously possible.

Contrary to most studies based on low-stake environments, we find female-dominated teams and female leaders to perform as well as their male counterparts in high-stake competitions, with some team gender differences evident across task types. Such team-level findings extend a burgeoning literature focusing on gender differences in preferences and outcomes at the individual-level, where women are generally reported to be outperformed by men (Gneezy et al. 2003; Booth 2009; Niederle and Vesterlund 2011; Buser et al. 2014). Our work especially complements the growing number of empirical studies on gender differences in response to increased individual payoffs (e.g., Camerer and Hogarth 1999; Antonovics et al. 2009; Azmat et al. 2015).

2. Data on Game Show Competitions

2.1 Data Collection

We collected consistent data on participating contestants, team compositions, and competition outcomes from a total of 397 episodes (37 seasons) of *The Apprentice* produced in 6 different geographical regions (Asia: 1 season, Australia: 5 seasons, Ireland: 4 seasons,

[2] For further discussions of this methodological issue, see Harrison and List (2004) and Van den Assem et al. (2012).

New Zealand: 1 season, United Kingdom: 11 seasons, United States: 15 seasons); 269 episodes (42 seasons) of *Survivor* also from 6 different countries (Australia: 2 seasons, Philippines: 2 seasons, South Africa: 3 seasons, Sweden: 4 seasons, United Kingdom: 2 seasons, United States: 29 seasons); and 189 episodes (17 seasons) of *Hell's Kitchen* from 3 different countries of production (Finland: 1 seasons, Italy: 2 seasons, United States: 14 seasons). Overall, this translates to a total sample of *N*=794 team-level observations for *The Apprentice*; *N*=944 for *Survivor*; and *N*=636 for *Hell's Kitchen*. Details about each competition structure and game show rules are provided in the Supplementary Appendix.

As a primary source to collecting the data, we carefully coded episode-by-episode summaries of the above seasons for each competition. These are clearly documented and available online at corresponding official show websites and related archives (see Supplementary Appendix for specific links to data sources). In order to confirm that the listed information was accurately stored, we also employed research assistants to view a large random selection of actual episodes of the televised shows.

Tables 1 and 2 report the frequency distributions of observed team sizes and gender compositions in each of the three competitions. The five distinct team gender categories listed in Table 1 ('All male', 'Majority male' (>50% share of males, but not all male), 'Equal mix' (50% share of males), 'Majority female' (<50% share of males, but not all female), and 'All female') form our main set of independent indicator variables. This range of team gender compositions enables us to clearly infer the changes in team performance outcomes as we sequentially move from all-male to all-female teams, and vice versa. Consistent across all three competitions, the size of teams ranges from two to ten, with the majority of teams having between four and eight members (see Table 2).[3]

2.3 Dependent Variable: Probability of Team Success

The main outcome or dependent variable that we consider in each of the three competitions is the winning percentage of teams (i.e., success rate). This probability of winning is simply equal to the mean value of a binary indicator variable which takes on a value of 1 if the team had won the task or challenge in the given competition, and zero otherwise. This average team performance and success variable however does not strictly equal 0.50 for all team observations and competitions as the competing teams are in some cases both deemed as the winners or losers of a challenge. This is especially true for the observed probability of

[3] Alternatively, if we restrict the team gender composition variable to only three separate but less precise indicators, such as 'low share' of women (<40%); 'medium share' of women (between 40% and 60%); and 'high share' of women (>60%), we reach the same empirical conclusions.

winning on *Hell's Kitchen* which is equal to 0.44 due to several instances of both teams performing badly during a dinner service task and hence being declared as losers. On the other hand, the average success level on the *Survivor* competition is approximately 0.51 (this point estimate is due to a three-team concept implemented in the Philippine version of the show, season three, where the top two best performing teams would both win the team challenge). The same is however not true for observed success rates on *The Apprentice,* with a clear winner and loser being chosen following a team project task (win rate = 0.50). The main focus of our empirical analysis is then to condition this probability of winning a task on the gender composition of teams (as categorised in Table 1).

2.4 Contestant Characteristics

Tables A1-A3 in the Supplementary Appendix summarise individual-level contestant characteristics for *The Apprentice*, *Survivor*, and *Hell's Kitchen*, respectively. In all three competitions, the average contestant was white and slightly above 30 years of age. While the oldest participants on *The Apprentice* and *Survivor* were 75 years of age, those on *Hell's Kitchen* were somewhat younger at 51 years. Information on each participant's occupational or job title was also available for *The Apprentice* and *Hell's Kitchen* competitions, but not for *Survivor*. We used these occupational titles to roughly approximate the degree of relevant experience or skill level of individual contestants. These values are coded from 0 to 3, where a zero is given to a contestant with an occupational title which is highly unrelated to the typical task undertaken on the given game show (such as a 'stockbroker' on *Hell's Kitchen),* while a value of 3 is assigned to individual contestants with highly relevant occupational backgrounds (such as an 'executive chef' on *Hell's Kitchen*). As noted earlier, these sampled contestants, and the observed variation in their background characteristics, form a more representative cross-section of the general population than the university students employed in most experimental studies (in which possible selection effects may similarly have potential, however unclear, influences on the empirical findings).

2.5 Team-Level Characteristics

The above individual-level job titles are then used to generate the average skill level of participating teams in *The Apprentice* and *Hell's Kitchen* competitions, as further summarised in Tables A4-A6. Alternatively, as noted in Table A5, to approximate the skill level of teams on *Survivor*, we use the average number of individual challenges or tasks won by team members; where information on such individual achievements and parts of the show is also readily available. The latter within-competition measure of skill can thus be viewed as being endogenous and hence more accurate than external proxies such as the current job titles of contestants.

The other team-level characteristics reported in Tables A4-A6 comprise the extra control variables in our formal regression analyses. In addition to capturing the average age and skill level of teams, we also measure (i) the racial diversity of teams, as approximated by the share of white team members; and (ii) hometown or regional homophily, defined as the highest share of team members from the same geographical region (captured mainly at the state and, where possible, city level). Information on the latter independent variable is however unavailable for about half of the contestants on *The Apprentice* and close to ten percent of the contestants on *Survivor*. In any case, our formal regression analysis shows this variable (similar to the other demographic controls) not to be an important factor for team performance outcomes.

3. Results

3.1 Overview of Main Findings

The key findings from the three studies are illustrated in Figure 1; which relates team gender composition to the observed probability of team success.

Study 1. Based on the raw data for *The Apprentice* (Figure 1A), there are no apparent differences in success rates between all-male and all-female teams (53% *vs* 48%, *P* = 0.444, two-sided test), and similarly between majority male and majority female teams (52% *vs* 51%, *P* = 0.955). Teams with an equal mix of men and women have the lowest observed likelihood of winning a project task at 43%. This success rate is about 10 percentage points lower than that of all-male teams (*P* = 0.139); 9 percentage points lower than that of majority male teams (*P* = 0.128); 8 percentage points lower than that of majority female teams (*P* = 0.138); and 5 percentage points lower than the success rate of all-female teams (43% *vs* 48%, *P* = 0.458). However, as shown in the parentheses above, none of these between-team differences reach statistical significance at conventional levels.

The robustness of the above descriptive results is further tested with more formal econometric analyses in Table 3, where we regress the probability of winning a project task on the set of gender composition dummy variables (with the ‘equal mix’ team gender category being the omitted/reference group). We are also able to control for a number of other team-level attributes including the average age of teams; racial diversity of teams; average skill level of teams; the gender of the project manager, as well as indicators for the country of production; season; and episode number. Once we account for such variables (in addition to the potential clustering of unobserved factors within each country, season, and episode grouping), the estimated magnitudes and differences in success rates between the equal gender mix and other team gender configurations still persist, however at fairly low levels of statistical significance.

Study 2. Figure 1B shows team success rates on *Survivor* to be very similar in magnitude across the different gender compositions: all-male teams have the highest success rate (58%), closely followed by all-female teams (54%); then by equal gender mix teams (52%); majority male teams (51%); and majority female teams (49%). None of these pairwise differences are statistically significant at conventional levels (with $P > 0.400$ for the largest observed difference of 9 percentage points between all-male and majority-female teams). This finding is further supported by the regression estimates in Table 4 (models 1 to 3); where none of the included gender composition dummies reach statistical significance. Overall, the gender mix of teams does not seem to matter for overall team performance in this high-stakes natural survival competition.

Study 3. Differences in performance outcomes by team gender mix are most apparent on *Hell's Kitchen*, with equal gender mix teams being associated with the highest probability of success (Figure 1C). Specifically, teams consisting of an equal share of men and women have a 27 percentage points higher success rate than all-male teams (69% *vs* 42%, $P = 0.040$); 29 percentage points higher than majority male teams (69% *vs* 40%, $P = 0.026$); 20 percentage points higher, but statistically indifferent, than majority female teams (69% *vs* 49%, $P = 0.144$); and 26 percentage points higher than all-female teams (69% *vs* 43%, $P = 0.049$). All-female teams are found to perform similarly to all-male teams ($P > 0.800$). Also, while majority female teams appear to have a 9 percentage point higher probability of success than majority male teams, this observed difference is not significant ($P > 0.200$).

Regression results in Table 5 show the above conclusions to hold after we account for other team-level characteristics (models 1 and 2), with only the reported difference between equal-gender mix and majority-female teams reaching statistical significance following the inclusion of competition-related variables (model 3). A test of equality between coefficients on the majority male and majority female team gender indicators suggests similar overall performances between these two groups (i.e., we cannot reject the null hypothesis of equality between the two estimated coefficients; $P > 0.270$). Overall, balanced-gender teams are estimated to be significantly more likely to win a team challenge on *Hell's Kitchen* than any of the other team genders.

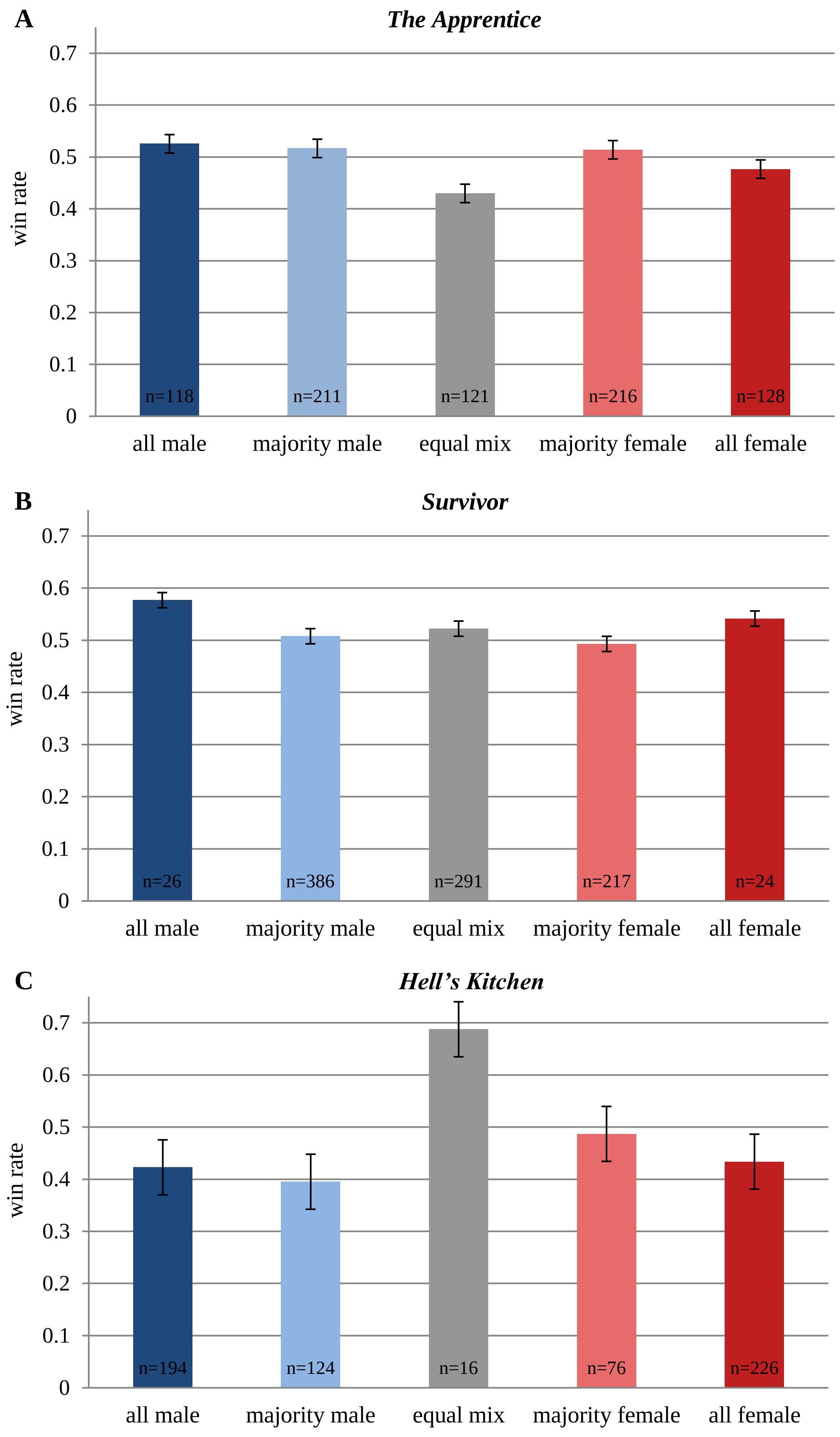

Figure 1: Gender Composition and Team Outcomes, by competition **(A)** *The Apprentice*; **(B)** *Survivor*; and **(C)** *Hell's Kitchen.* Number of observations per team gender category is denoted by *n*. Error bars indicate ± 1 SE.

3.2 Team Success Rates by Gender Composition and Task Type

We next consider whether the above observed patterns between team gender composition and performance outcomes differ by the type or nature of task undertaken.

Study 1. Figure 2A presents team success rates on *The Apprentice* by team gender mix and project type. All-male teams are found to be considerably more successful than all-female teams in projects involving the design and creation of new business products (win rate: 70% *vs* 25%, *P* = 0.002, two-sided test). Similarly, teams comprised only of men also have a higher probability of winning a design task compared to majority male (52%); majority female (53%); and balanced-gender (46%) teams. Only the observed difference between all-male and balanced-gender teams is however statistically significant (*P* = 0.074), with the remaining two disparities in win percentages being insignificant at conventional levels (*P* > 0.130 for each pairwise, two-sided, comparison). All-female teams thus have significantly worse outcomes compared to any of the other team genders (*P* < 0.099 for each comparison). Formal regression estimates in Table 6 show these initial outcome gaps between single-gender and balanced-gender teams to disappear once we account for other controls and clusters. However, significant differences in the estimated probability of success between all-male and all-female teams still remain (i.e., we reject the null hypothesis of equality between the corresponding coefficients; 0.25 *vs* -0.23, *P* = 0.043). To further and more directly demonstrate the latter disparity, Supplementary Table A7 reports the same probit regression equations but with the all-male team gender category this time omitted as the reference group. In line with the summary results above, all-female teams are estimated to be as much as 47 percentage points less likely to succeed in the creation of new commercial products and ideas on *The Apprentice* than all-male teams (*P* = 0.032).

There are also some apparent differences in outcomes across team genders for classic business trading tasks, where teams are required to buy and sell actual products in order to make profits. The middle-panel of Figure 2A indicates both majority male and majority female teams to have similarly higher success rates in buying and selling tasks (53% and 55%, respectively) relative to teams with an equal gender mix (40%). Interestingly, single-gender teams are found to perform somewhere between majority-gender and balanced-gender teams, however again these raw differences do not reach statistical significance. The above findings are mostly corroborated by the regression estimates shown in the middle-panel of Table 6, where majority male teams are estimated (in the full-specification) to have a 19 percentage point higher probability of winning a trading task compared to teams with an equal-gender mix (*P* = 0.059). An almost identical marginal effect is estimated for majority female teams (20 percentage points, *P* = 0.066). Moreover, there are no statistically significant differences in the estimated likelihood of success between equal-gender mix and single-gender teams (*P* > 0.300), after controlling for other observable factors and potential clusters.

Finally, in terms of marketing projects undertaken on *The Apprentice*: teams that have a higher mix of males are found to perform equally as well as female-dominated teams, with corresponding win percentages of 51% (all male), 50% (majority male) and 56% (all female), respectively. On the other hand, balanced gender teams (44% win rate) and majority female (46% win rate) teams are found to be slightly less successful. However, none of these differences in team success rates are statistically significant at conventional levels (*P* > 0.240 for each pairwise comparison). The above findings are also confirmed by the regression estimates reported in Table 6 and Table A7. Overall, we find no significant differences in team outcomes by gender composition for promotional and marketing-oriented projects.

Study 2. In Figure 2B, we observe a monotonic increase in the probability of winning an intellectual task on *Survivor* as the share of women in a team increases; with all-female teams having the highest success rate (63%) and all-male teams having the lowest rate (38%). The opposite finding however holds for team challenges of a physical nature: all-male teams win 67% of physical tasks, compared to 50% by all-female teams. The most notable, and only statistically significant, difference in success rates is observed between all-male and majority female teams. The former team gender mix is found to be around 22 percentage points more likely to win a physical challenge (*P* = 0.083). The probit regression results in Table A8 show the above inferences to also hold once we correct for other measured factors and cluster effects.

Within-competition stakes. Figure 3A presents team success rates by gender composition and type of challenge on *Survivor* (reward *versus* immunity). Immunity challenges are of higher intrinsic value for contestants since the winning team members are safe from elimination and hence guaranteed progression to the next stage of the competition. As a result, there is relatively more at stake during an immunity challenge. From the raw data, we find asymmetric patterns in team outcomes across the two types of challenges: all-female teams perform the best in reward challenges, with the remaining team genders having lower but quite similar success rates (however, these differences fail to reach statistical significance). On the other hand, all-male teams tend to have most success in immunity challenges. For instance, majority female teams are estimated to be 23 percentage points less likely to win an immunity challenge compared to all-male teams (*P* = 0.092), once we control for a host of other factors (Table A9). The latter pattern in team outcomes suggests male teams to perform better than female teams under pressure (i.e., as the importance of the task increases) in this high-stake competitive environment.

Study 3. In terms of similar increases in within-competition stakes on *Hell's Kitchen*, we find the probability of winning the 'dinner service' task to be lower across each team gender (Figure 3B), compared to that during the 'team challenge' task. This is mainly due to the increased level of difficulty and overall task completion rates. Nonetheless, there are no changes in relative success rates (by team gender mix) across the team challenges, with balanced-gender teams always having the highest win percentage (however, we do note that the extremely small number of equal-gender mix teams observed during the 'team challenge' task makes these findings incompletely valid).

3.3 Sex of Team Leader and Performance Outcomes

Our final set of research questions concerned the sex of the team leader and associated competition outcomes. To shed light on this topic, we turn to the summary results from *The Apprentice* as depicted in Figure 4. Overall team success rates are found to be highly similar under male and female leaders (Figure 4A), with both sexes winning at a rate close to 50% (51% *vs* 49%, $P > 0.560$).

Figure 4B shows this finding to be robust to the type of business task performed, with no significant leadership gender differences in product design, trading, or promotional tasks ($P > 0.395$ for each pairwise comparison).

Figure 4C summarises team success rates conditional on the sex of the project manager and the gender composition of teams. While there is a slight indication that all-female teams perform better under male project managers (i.e., the win percentage of all-female teams is equal to 55% under a male leader and 48% under a female leader), the gender of the project manager seems to play no major role in team success rates (with each of the paired comparisons being statistically indifferent; $P > 0.500$). Overall, male and female team leaders are found to perform with equal success in high-stake business competitions, independent of the team gender mix or the type of project undertaken.

4. Discussion

The use of teams in organisations and other productive environments has increased rapidly over the past few decades. One team attribute of general importance for team performance is gender diversity. This paper studied the relationship between gender composition and team outcomes in a series of competitive environments with large monetary and career changing stakes. It did so by analysing new game show data on approximately 2,300 teams and their success rates.

Our results show that gender diversity does not play a strong role in team outcomes in large-stake business (*The Apprentice*) and natural survival (*Survivor*) competitions, however is associated with better outcomes in high-pressure cooking competitions (*Hell's Kitchen*). In the latter setting, while female-dominated teams do perform as well as their male counterparts, teams with an equal share of men and women perform best. This suggests the optimal team gender mix to be context-dependent even in environments where the intensity of competition and individual rewards are unprecedentedly high.

We also find performance outcomes by team gender to be task-specific in the studied high-stake business competition. Sizable and robust differences in outcomes are apparent between all-male and all-female teams for projects which demand creativity on *The Apprentice*, while similar conclusions do not hold for typical trading or promotional tasks. Although such findings could imply that male teams are simply more creative when it comes to business-oriented projects, we alternatively note the possible existence of a gender bias in the attribution of creativity. Recent experimental and archival evidence by Proudfoot et al. (2015) shows women to be less likely than men to have their creative ideas acknowledged and rewarded, even when they produce identical output. This gender bias in perceived creativity is shown to persist across a range of professional settings and evaluators, with judged creativity being strongly associated with stereotypically masculine traits including ambitious, competitive, daring, and courageous appearances.

Our reported estimates and findings are robust to the inclusion of team skill and life experience levels (as approximated by the age and relevant occupational levels of contestants). While the used skill measures are not ideal, and arguably quite inaccurate for such uncommon game show settings faced by participants, we improve on this by capturing endogenous (within-competition) ability levels of contestants and teams in the *Survivor* competition. In the latter analysis, we especially note the found monotonic rise in team success rates during more intellectually-oriented tasks as the share of women within a team increases. This empirical pattern is consistent with that described in Woolley et al. (2010); where teams of laboratory subjects perform better in intellectual tasks as relatively more women are included, even after controlling for individual intelligence levels. Such findings suggest that a group's collective intelligence and collaboration level can potentially be influenced by the chosen gender composition.

Finally, the result that female team leaders are as successful as male team leaders in competitive business environments (after controlling for team gender mix and task type) provides some rare empirical evidence on the issue, and gives support for the increased promotion of women into high-stake leadership roles. To this end, a gradual exposure process of competitive teams to female leaders is one potential strategy to demolish the historic norm of male leadership. Nevertheless, as recent experimental studies demonstrate, mandating female leadership or imposing strict gender quotas may not necessarily lead to better outcomes for women in the short-run due to the possibility of backlash and peer sabotage (Gangadharan et al. 2015; Leibbrandt et al. 2015). Furthermore, female leaders may have doubts about the awaiting group's willingness to follow, especially in mixed-gender and female-minority environments (Dasgupta et al. 2015; Grossman et al. 2015).

The present work provides some of the first evidence on team gender composition and outcomes in a unique set of competitive settings where the potential payoffs are substantially increased. The results suggest the influence of gender diversity on team performance to broadly differ from the narrow range of experimental and observational contexts studied so far. Moreover, our gender findings from actual team competitions counter much of the existing results based on between-individual competitions in which female participants are usually found to perform worse than their male counterparts. Although our study has the advantage that the teams we observe are not endogenously formed by the contestants themselves (e.g., Apesteguia et al. 2012), these teams are also not randomly assigned (e.g., Hoogendoorn et al. 2013). In most cases, participants are first allocated into teams based on some observable individual characteristic (such as gender) and then further assigned according to the ultimate judge's preference. Only in a very small number of instances is a team's gender mix altered due to an exogenous event such as a personal injury. Thus, the found relations between team gender composition and performance outcomes cannot be seen as being strictly causal. One promising approach for future research on gender diversity is the use of field experiments in similar real-world settings; however the issue of providing life-changing or large enough stakes for the studied subjects would still remain, especially in Western societies. We hope that the current study provides an example and further stimulates ideas on how available archival data, such as that from game shows, can be used in different ways to study important gender issues in teams and overall facilitate a greater variety of field settings.

References

Adams RB, Ferreira D (2009) Women in the boardroom and their impact on governance and performance. *J Financ Econ* 94(2):291–309.

Ahern KR, Dittmar AK (2012) The changing of the boards: The impact on firm valuation of mandated female board representation. *Q J Econ* 127(1):137–197.

Ambrus A, Greiner B, Pathak PA (2015) How individual preferences are aggregated in groups: An experimental study. *J Public Econ* 129: 1–13.

Amini M, Ekström M, Ellingsen T, Johannesson M, & Strömsten F (2016) Does gender diversity promote nonconformity? *Manage Sci* forthcoming.

Antonovics K, Arcidiacono P, Walsh R (2005) Games and discrimination: Lessons from the Weakest Link. *J Hum Resour* 40(4):918–947.

Antonovics K, Arcidiacono P, Walsh R (2009) The effects of gender interactions in the lab and in the field. *Rev Econ Stat* 91(1):152–162.

Apesteguia J, Azmat G, Iriberri N (2012) The impact of gender composition on team performance and decision making: Evidence from the field. *Manage Sci* 58(1):78–93.

Ariely D, Gneezy U, Loewenstein G, Mazar N (2009) Large stakes and big mistakes. *Rev Econ Stud* 76(2):451–469.

Azmat G (2014). Gender diversity in teams. *IZA World of Labor* 29:1–10.

Azmat G, Calsamiglia C, Iriberri, N (2015) Gender differences in response to big stakes. *J Eur Econ Assoc* forthcoming.

Balafoutas L, Kerschbamer R, Kocher M, & Sutter M (2014) Revealed distributional preferences: Individuals vs teams. *J Econ Behav Organ* 108: 319–330.

Baumeister RF (1984) Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *J Pers Soc Psychol* 46(3):610–620.

Bear JB, Woolley AW (2011) The role of gender in team collaboration and performance. *Interdiscip Sci Rev* 36(2): 146–153.

Beetsma RM, Schotman PC (2001) Measuring risk attitudes in a natural experiment: data from the television game show Lingo. *Econ J (London)* 111(474): 821–848.

Beilock S (2010). Choke: *What the secrets of the brain reveal about getting it right when you have to*. Simon and Schuster.

Belot M, Bhaskar V, Van de Ven J (2010) Promises and cooperation: Evidence from a TV game show. *J Econ Behav Organ* 73(3):396–405.

Bennett RW, Hickman KA (1993) Rationality and the ‘Price is Right’. *J Econ Behav Organ* 21(1):99–105.

Berk JB, Hughson E, Vandezande K (1996) The price is right, but are the bids? An investigation of rational decision theory. *Am Econ Rev* 86(4): 954–970.

Blau F, Kahn L (2000) Gender differences in pay. *J Econ Perspect* 14(4):75–99.

Booth AL (2009) Gender and competition. *Labour Econ* 16(6): 599–606.

Booth AL, Nolen PJ (2012) Choosing to compete: How different are girls and boys? *J Econ Behav Organ* 81:542–555.

Buser T (2016) The impact of losing in a competition on the willingness to seek further challenges. *Manage Sci* forthcoming.

Buser T, Niederle M, Oosterbeek H (2014) Gender, competitiveness and career choices. *Q J Econ* 129(3): 1409–1447.

Camerer CF, Hogarth RM (1999) The effects of financial incentives in experiments: A review and capital-labor-production framework. *J Risk Uncertainty* 19(1): 7–42.

Ceci SJ, Williams WM (2011) Understanding current causes of women's underrepresentation in science. *Proc Natl Acad Sci USA*, 108(8):3157–3162.

Charness G, Sutter M (2012) Groups make better self-interested decisions. *J Econ Perspect* 26(3): 157–176.

Dasgupta N, Scircle MM, Hunsinger M (2015) Female peers in small work groups enhance women's motivation, verbal participation, and career aspirations in engineering. *Proc Natl Acad Sci USA* 112(16):4988–4993.

Dilks LM, Thye SR, Taylor PA (2010) Socializing economic theories of discrimination: Lessons from Survivor. *Soc Sci Res* 39(6):1164–1180.

Dufwenberg M, Muren A (2006) Gender composition in teams. *J Econ Behav Organ* 61(1):50–54.

Eagly AH, Carli LL (2003) The female leadership advantage: An evaluation of the evidence. *Leadersh Q* 14(6):807–834.

Eagly AH, Karau SJ, Makhijani MG (1995) Gender and the effectiveness of leaders: a meta-analysis. *Psychol Bull* 117(1): 125–145.

Flory JA, Leibbrandt A, List JA (2015) Do competitive workplaces deter female workers? A large-scale natural field experiment on job entry decisions. *Rev Econ Stud* 82(1):122–155.

Gangadharan L, Jain T, Maitra P, Vecci J (2015) Social norms and governance: The behavioral response to female leadership. *Eur Econ Rev* forthcoming.

Gertner R (1993) Game shows and economic behavior: risk-taking on "Card Sharks". *Q J Econ* 108(2):507–521.

Gneezy U, Niederle M, Rustichini A (2003) Performance in competitive environments: Gender differences. *Q J Econ* 118(3):1049–1074.

Goldin C, Rouse, C (2000) Orchestrating impartiality: The impact of "blind" auditions on female musicians. *Am Econ Rev* 90(4):715–741.

Grossman PJ, Komai M, Jensen JE (2015) Leadership and gender in groups: An experiment. *Can J Econ* 48(1):368–388.

Hamilton BH, Nickerson JA, Owan H (2003) Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *J Polit Econ* 111(3):465–497.

Harrison GW, List JA (2004) Field experiments. *J Econ Lit* 42(4): 1009–1055.

Hoogendoorn S, Oosterbeek H, Van Praag M (2013) The impact of gender diversity on the performance of business teams: Evidence from a field experiment. *Manage Sci* 59(7):1514–1528.

Ivanova-Stenzel R, Kübler D (2011) Gender differences in team work and team competition. *J Econ Psychol* 32(5):797–808.

Kamenica E (2012) Behavioral economics and psychology of incentives. *Annu Rev Econom* 4(1):427–452.

Kocher MG, Sutter M (2005) The decision maker matters: Individual versus group behaviour in experimental beauty-contest games. *Econ J (London)* 115(500): 200–223.

Lazear EP, Shaw KL (2007) Personnel economics: The economist's view of human resources. *J Econ Perspect* 21(4):91–114.

Leibbrandt A, Wang LC, Foo C (2015) Gender quotas, competitions, and peer review: Experimental evidence on the backlash against women. CESifo Working Paper Series No. 5526.

Levitt SD (2004) Testing theories of discrimination: Evidence from Weakest Link. *J Law Econ* 47(2):431–453.

List JA (2004) Young, Selfish and Male: Field evidence of social preferences. *Econ J (London)* 114(492):121–149.

List JA (2006). Friend or foe? A natural experiment of the prisoner's dilemma. *Rev Econ Stat* 88(3):463–471.

List JA, Neilson WS, Price MK (2016) The effects of group composition in a strategic environment: Evidence from a field experiment. *Eur Econ Rev* forthcoming.

Maciejovsky B, Sutter M, Budescu DV, Bernau P (2013) Teams make you smarter: How exposure to teams improves individual decisions in probability and reasoning tasks. *Manage Sci* 59(6): 1255–1270.

Metrick A (1995) A natural experiment in "Jeopardy!". *Am Econ Rev* 85(1):240–253.

Niederle M, Vesterlund L (2007) Do women shy away from competition? Do men compete too much?. *Q J Econ* 122(3):1067–1101.

Niederle M, Vesterlund L (2011) Gender and competition. *Annu Rev Econ* 3(1): 601–630.

Oberholzer-Gee F, Waldfogel J, White MW (2010) Friend or foe? Cooperation and learning in high-stakes games. *Rev Econ Stat* 92(1): 179–187.

Pearsall MJ, Ellis AP, Evans JM (2008) Unlocking the effects of gender faultlines on team creativity: Is activation the key?. *J Appl Psychol* 93(1):225–234.

Post T, Van den Assem MJ, Baltussen G, Thaler RH (2008) Deal or no deal? Decision making under risk in a large-payoff game show. *Am Econ Rev* 98(1):38–71.

Proudfoot D, Kay AC, Koval CZ (2015) A gender bias in the attribution of creativity: Archival and experimental evidence for the perceived association between masculinity and creative thinking. *Psychol Sci* 26(11):1751–1761.

Shurchkov O (2012) Under pressure: Gender differences in output quality and quantity under competition and time constraints. *J Eur Econ Assoc* 10(5):1189–1213.

Sutter M, Glätzle–Rützler D (2015) Gender differences in the willingness to compete emerge early in life and persist. *Manage Sci* 61(10): 2339–2354.

Tenorio R, Cason TN (2002) To spin or not to spin? Natural and laboratory experiments from The Price is Right. *Econ J (London)* 112(476):170–195.

Van den Assem MJ, Van Dolder D, Thaler RH (2012) Split or steal? Cooperative behavior when the stakes are large. *Manage Sci* 58(1):2–20.

Van Dolder D, Van den Assem MJ, Camerer C, Thaler RH (2015) Standing united or falling divided? High stakes bargaining in a TV game show. *Am Econ Rev P&P* 105(5):402–407.

Van Vugt M, Hogan R, Kaiser RB (2008) Leadership, followership, and evolution: some lessons from the past. *Am Psychol* 63(3): 182–196.

Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW (2010) Evidence for a collective intelligence factor in the performance of human groups. *Science* 330(6004):686–688.

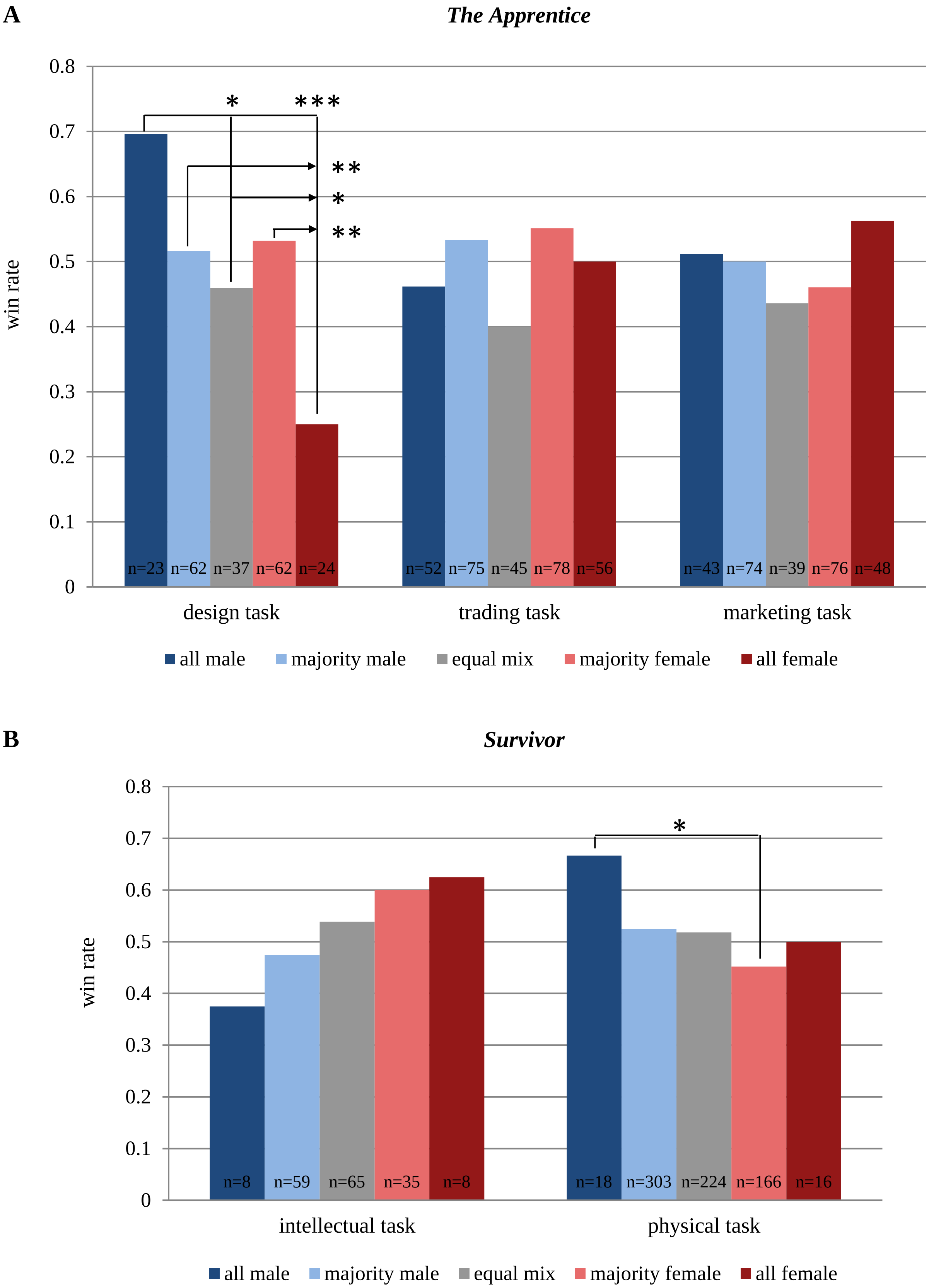

Figure 2: Gender Composition and Team Outcomes on **(A)** *The Apprentice*, and **(B)** Survivor, by project/task type. Number of observations per team gender category is denoted by *n*. **P* < 0.10; ***P* < 0.05; ****P* < 0.01, based on two-sided test of differences between proportions.

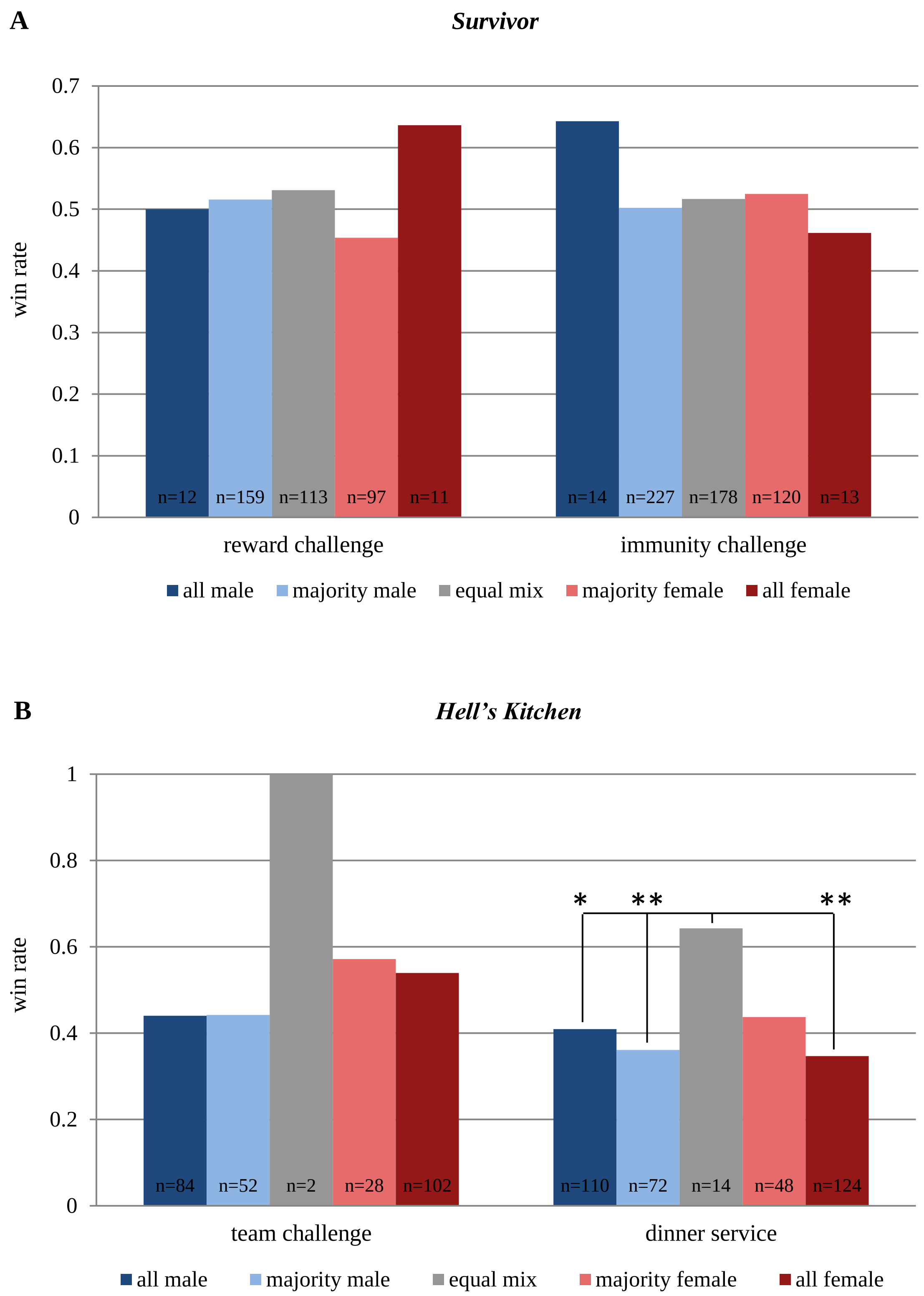

Figure 3: Gender Composition and Team Outcomes on **(A)** *Survivor*, and **(B)** *Hell's Kitchen*, by level of within-competition stakes. 'Immunity' challenge in subfigure **(A)** and 'dinner service' challenge in subfigure **(B)** have higher relative within-competition stakes (as experienced by contestants). Number of observations per team gender category is denoted by *n*. **P* < 0.10; ***P* < 0.05; ****P* < 0.01, based on two-sided test of differences between proportions.

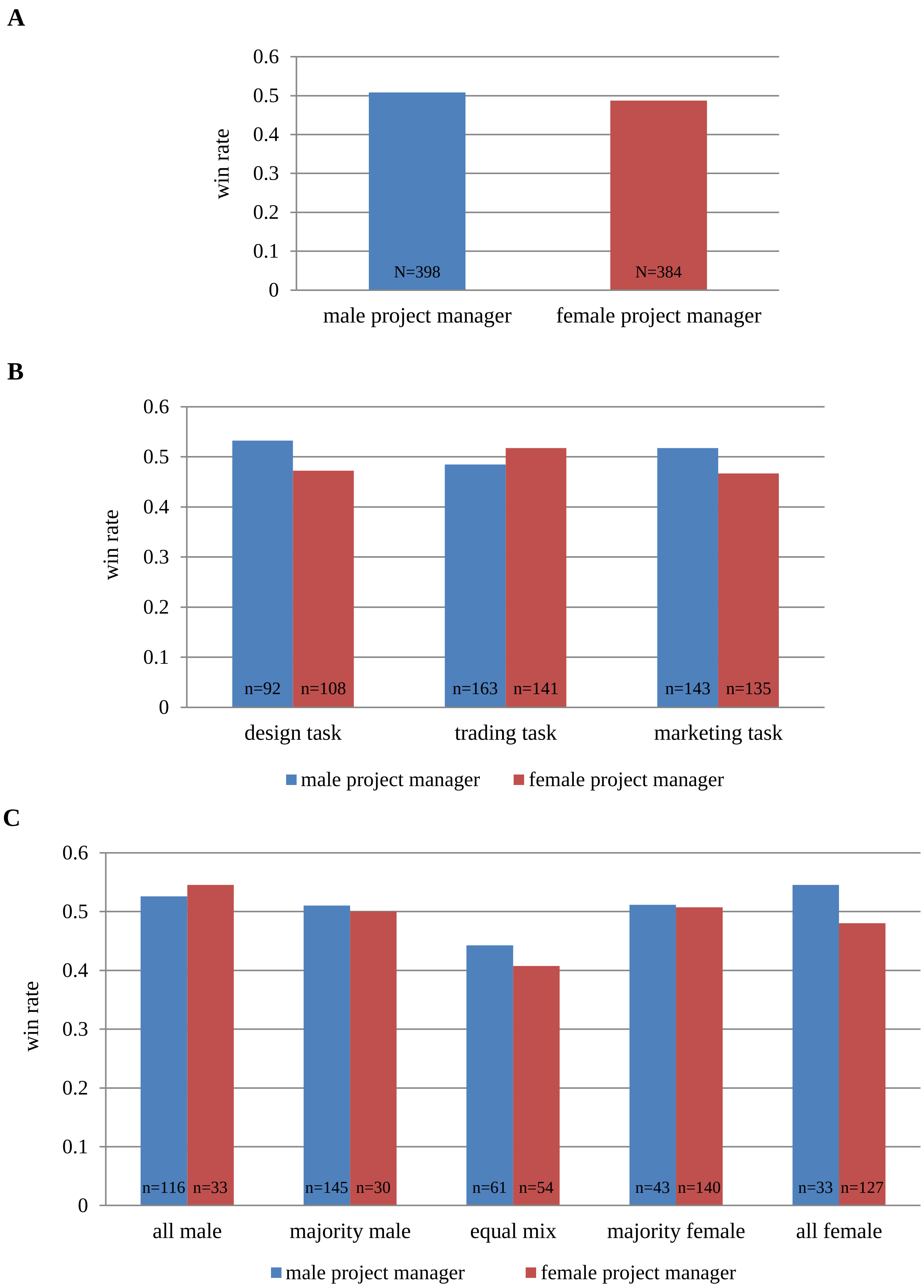

Figure 4: Sex of the project manager (team leader) and team outcomes on *The Apprentice*. **(A)** Success rate by sex of the project manager. **(B)** Success rate by sex of the project manager and team gender composition. **(C)** Success rate by sex of the project manager and business project type. Number of observations per gender category is denoted by *n*.

Table 1: Frequency Distributions of Team Gender Mix, by competition

	The Apprentice		***Survivor***		***Hell's Kitchen***	
Gender mix	Frequency	Percent	Frequency	Percent	Frequency	Percent
All male	118	14.86	26	2.75	194	30.50
Majority male	211	26.97	386	40.89	124	19.50
Equal mix	121	15.24	291	30.83	16	2.52
Majority female	216	27.20	217	22.99	76	11.95
All female	128	16.12	24	2.54	226	35.53
Total	**794**	100.00	**944**	100.00	**636**	100.00

Table 2: Frequency Distributions of Team Size, by competition

	The Apprentice		***Survivor***		***Hell's Kitchen***	
Team size	Frequency	Percent	Frequency	Percent	Frequency	Percent
2	39	4.91	2	0.21	1	0.16
3	121	15.24	6	0.64	63	9.91
4	152	19.14	52	5.51	132	20.75
5	147	18.51	128	13.56	122	19.18
6	131	16.50	216	22.88	119	18.71
7	107	13.48	220	23.31	91	14.31
8	70	8.82	169	17.90	58	9.12
9	24	3.02	121	12.82	40	6.29
10	3	0.38	30	3.18	10	1.57
Total	**794**	100.00	**944**	100.00	**636**	100.00

Table 3: Estimated Probability of Team Success on *The Apprentice*

	(1)		(2)		(3)	
	b	*P*	*b*	*P*	*b*	*P*
All male	0.096	0.118	0.092	0.165	0.103	0.157
Majority male	0.087	0.113	0.086	0.127	0.099*	0.100
Equal mix (*reference*)						
Majority female	0.084	0.145	0.092	0.128	0.109*	0.092
All female	0.047	0.430	0.064	0.339	0.094	0.209
Average age			0.002	0.536	0.009	0.262
Average skill level			0.022	0.599	-0.025	0.718
Proportion white			-0.034	0.678	-0.073	0.631
Male project manager			0.011	0.812	0.017	0.730
Other controls	No		No		Yes	
Observations	794		782		782	

Notes: Probit regression model. Average marginal effects (*b*), with corresponding p-values (*P*). Robust standard errors clustered at the Country, Season, and Episode level. Dependent variable equals 1 if team won the task/challenge, 0 otherwise. 'Equal mix' team-gender category is the omitted/reference group. Included set of team-level characteristics are average age; average skill/relevant occupational level; proportion of white team members; and gender of the project manager. *Other controls* capture show-related variables including season, episode, and country of production fixed effects, task type indicators, and an indicator for the celebrity version of the show. **P* < 0.10; ***P* < 0.05; ****P* < 0.01.

Table 4: Estimated Probability of Team Success on *Survivor*

	(1)		(2)		(3)	
	b	*P*	*b*	*P*	*b*	*P*
All male (*reference*)						
Majority male	-0.069	0.579	-0.058	0.639	-0.075	0.546
Equal mix	-0.055	0.663	-0.042	0.736	-0.077	0.547
Majority female	-0.084	0.514	-0.080	0.531	-0.101	0.439
All female	-0.035	0.869	-0.040	0.850	0.006	0.977
Average age			-0.001	0.969	0.002	0.811
Average skill level			0.036	0.244	0.034	0.515
Proportion white			0.048	0.330	0.338*	0.069
Highest proportion from same region					0.207	0.352
Immunity challenge					0.001	0.859
Other controls	No		No		Yes	
Observations	944		944		878	

Notes: Probit regression model. Average marginal effects (*b*), with corresponding p-values (*P*). Robust standard errors clustered at the Country, Season, and Episode level. Dependent variable equals 1 if team won the task/challenge, 0 otherwise. 'All male' team gender category is the omitted/reference group. Included team-level characteristics are team average age; average skill level; proportion of white team members; and highest share of team members from the same geographical region/hometown. *Other controls* capture show-related variables including season, episode, and country of production fixed effects; a challenge type (immunity) indicator variable; nature of task indicators (intellectual/physical); and an indicator for the celebrity version of the show. *P < 0.10; **P < 0.05; ***P < 0.01.

Table 5: Estimated Probability of Team Success on *Hell's Kitchen*

	(1)		(2)		(3)	
	b	*P*	*b*	*P*	*b*	*P*
All male	-0.267***	0.008	-0.257***	0.016	-0.308**	0.025
Majority male	-0.295***	0.005	-0.301***	0.006	-0.379***	0.004
Equal mix (*reference*)						
Majority female	-0.204*	0.081	-0.215*	0.075	-0.279**	0.050
All female	-0.256***	0.010	-0.286***	0.007	-0.295**	0.027
Average age			-0.013	0.151	-0.012	0.369
Average skill level			-0.007	0.847	0.068	0.244
Proportion white			0.060	0.599	0.165	0.353
Highest proportion from same region			-0.159	0.302	-0.477*	0.076
Dinner service task					-0.135***	0.000
Other controls	No		No		Yes	
Observations	636		636		636	

Notes: Probit regression model. Average marginal effects (*b*), with corresponding p-values (*P*). Robust standard errors clustered at the Country, Season, and Episode level. Dependent variable equals 1 if team won the task/challenge, 0 otherwise. 'Equal mix' team gender category is the omitted/reference group. Included team-level characteristics are team average age; average skill/relevant occupational level; proportion of white team members; and highest share of team members from the same geographical region/hometown. *Other controls* capture show-related variables including season, episode, and country of production fixed effects; a task type (dinner service) indicator variable; and an indicator for the celebrity version of the show. *P < 0.10; **P < 0.05; ***P < 0.01.

Table 6: Estimated Probability of Team Success on *The Apprentice*, by task type

	Design task				Trading task				Marketing task			
	b	*P*	*b*	*P*	*b*	*P*	*b*	*P*	*b*	*P*	*b*	*P*
All male	0.235*	0.054	0.247	0.125	0.062	0.521	0.088	0.432	0.076	0.463	0.087	0.538
Majority male	0.055	0.587	0.063	0.622	0.133	0.126	0.186*	0.059	0.064	0.501	0.060	0.580
Equal mix (*reference*)												
Majority female	0.070	0.508	0.078	0.563	0.151	0.108	0.199*	0.066	0.025	0.800	0.074	0.530
All female	-0.220*	0.076	-0.230	0.159	0.100	0.274	0.117	0.333	0.126	0.212	0.284**	0.038
Other controls	No		Yes		No		Yes		No		Yes	
Observations	208		200		306		304		280		278	

Notes: Probit regression model. Average marginal effects (*b*), with corresponding p-values (*P*). Robust standard errors clustered at the Country, Season, and Episode level. Dependent variable equals 1 if team won the task/challenge, 0 otherwise. ‘Equal mix’ team gender category is the omitted/reference group. *Other controls* capture team-level characteristics such as average age; average skill/relevant occupational level; proportion of white team members; and gender of the project manager. Also included are season, episode, and country of production fixed effects, and an indicator for the celebrity version of the show. *P < 0.10; **P < 0.05; ***P < 0.01.

Supplementary Appendix A – *Mujcic and Kalayci (2016)*

A1. Structure of Game Show Competitions
A2. Archival Data Sources
Supplementary Tables A1-A9

A1. Structure of Game Show Competitions

Study 1: The Apprentice. Our first study of gender composition and team performance was based on a sample of *N*=794 team-level observations from *The Apprentice*. This well-known reality television competition involves two teams of contestants that compete in a range of business-oriented projects for the ultimate prize of a one-year US$250,000 starting contract to manage a business enterprise owned by the host and ultimate judge of the competition (e.g., Donald Trump in the US version of the show).

Contestants are initially divided into two teams of equal size (at first, and in most seasons, by gender). Each episode the teams are assigned a team project task to undertake and compete in. This task typically examines a given business skill such as (i) creating and designing new products or services; (ii) buying and selling goods or services; and (iii) advertising and promoting existing and upcoming goods or services. Each team selects a 'project manager' who is responsible for leading and managing the team for the duration of the project. The two teams are also regularly visited and monitored by advisors employed by the ultimate judge.

Project task outcomes are measured from a pure business sense: the creativity and potential appeal of a new product; the total sales and profits made; or the quality and reach of an advertising campaign. Overall, the team with the best relative result is declared as the winner and is rewarded with a positive experience ranging from fine-dining to a tour of another successful corporation.

The losing team remains in the boardroom to discuss their efforts and failures with the judging panel. The team project manager is then asked to return with a set number of other team members (usually two, but sometimes also one or three), which he/she believes were the worst performers. In this final stage, one of the contestants is fired and eliminated from the competition. The ultimate judge can however deviate from the standard procedure by overruling the losing project manager on which team members to bring back to the boardroom; or by simultaneously firing more than one contestant; and even eliminating contestants before the final stage/boardroom meeting.

Team compositions are therefore altered each time a contestant is fired, as well as when team changes are directly enforced by the ultimate judge. The latter adjustment usually takes place multiple times in a single season of the show. Such changes are enforced to enable the ultimate judge to assess the performance of contestants in various team configurations, and in some instances to correct for highly unbalanced team sizes (such as when one team repeatedly loses and quickly depreciates in numbers). We use the above changes in the size and gender mix of teams as the main source of variation for our independent variable(s) of interest.

***Study 2: Survivor*.** The second study used *N*=994 team-level observations extracted from the popular reality television competition *Survivor*, a reality competition television series in which 16 to 20 contestants compete in a natural remote location (such as a tropical island or jungle) for a cash prize of one million dollars.

The individual contestants are split into two teams, or 'tribes', and compete in a range of intellectual and physical challenges to either gain a reward (such as food, hunting tools, fire, and other luxuries) or immunity (being safe from elimination and progressing to the next round). The team that loses the immunity challenge has to anonymously vote out one of its members, making the team (arguably) weaker in future challenges. Initial teams are either gender balanced or, less often, divided by gender. The gender mix of teams evolves with each stage of the competition as, for example, (i) individual contestants are eliminated; (ii) new teams are assigned by show directors; and (iii) contestants are forced to leave due to injury or other personal reasons. The latter act is however only observed on a handful of occasions in the total sample.

When about half of the contestants remain, teams are combined into a single group and the contestants start to compete (head-to-head) as individual players. In this later stage, winning 'immunity' saves individuals from elimination; for which votes are cast by a jury made up of previously eliminated contestants. The final three contestants participate in a final 'Tribal Council' meeting to convince the jury as to why they deserve to win the entire competition.

As there is a strategic element to the game, Dilks et al. (2010) analyse the voting patterns of contestants on *Survivor* (for 17 seasons of the US version) and report evidence of a significant bias in the relative number of votes against female, non-white, and mature (> 40 years of age) contestants during the first half of the competition (i.e., when contestants compete as teams and collective victories lead to rewards and guaranteed progression).

Dilks et al. however find the opposite to be true in the second half of the competition: when contestants compete individually they vote relatively more against younger, male, and white contestants. The authors argue such findings to be consistent with statistical discrimination on the basis of required skill and competence at the different stages of the competition.

Team challenges undertaken during the first haft of the show are of main interest to us; where, based on the above research, teams may not always be exogenously formed but are made up of the most competitive or able participants. The reward and immunity challenges consist of predominantly intellectual and physical tasks. More cognitively demanding tasks include activities such as solving puzzles; memory games; geographical guidance, and creating the design for shelter, rafts, and rescue signals. On the other hand, physical challenges are based around endurance, lifting, balancing, eating, running and swimming activities. Each team challenge has set rules and scoring systems. For a majority of team challenges, the winner is objectively determined by simply looking at, for example, which group completes the task more quickly. In other tasks (which are typically more intellectual in nature and require creativity), a panel of external judges decides on the winning team.

Study 3: Hell's Kitchen. The third, and final, study gathered data (*N*=636) on team gender and performance outcomes from *Hell's Kitchen*, a reality television cooking competition involving 12 to 20 contestants who compete for the title of head chef at an exclusive restaurant owned by the host chef and ultimate judge (Gordon Ramsey in the UK and US versions of the show), in addition to a cash prize of US$250,000.

The contestants compete in two teams which are initially split by gender. Each episode consists of a team (reward) challenge and a dinner service, after which one of the contestants is eliminated.

The team challenge is based around a cooking task (such as ingredient preparation, or a tasting test) for which the winner is determined by either a set scoring system or judging panel, and is rewarded with a luxury experience. On the other hand, the losing team is punished by having to perform unpleasant tasks such as cleaning the kitchen areas and preparing ingredients for upcoming dinner services.

For the dinner service task, the two teams compete in taking actual orders and preparing a three-course meal for around 100 guest diners, with each team operating from their own fully-equipped kitchen area. The dinner service tasks encompass much higher levels of pressure than team challenges due to expected high standards, constant quality checks, and speed of service requests by the ultimate judge. Each team is also closely monitored and scrutinised by employed assistants/supervisors. While the aim is to complete each dinner service, this does not always occur due to teams breaking down. In such instances, the ultimate judge shuts down the kitchen areas and informs the teams to come back for the elimination stage. The winning team is then announced, while the losing team is asked to decide which individual team members should be up for elimination. The ultimate judge then eliminates a member of the losing team from the competition.

The size and gender composition of the initial single-gender teams thus varies based on the sequence of team results. When the same team consistently loses the dinner service task and thus begins to be outnumbered by the opposing team, the ultimate judge selects members of the dominant team to switch across and balance the team numbers. Even when the competing teams are of equal size, the ultimate judge frequently alters the two teams in order to observe how the contestants perform under varied team compositions. Additionally, in a very minor number of cases (n<10), individual contestants leave the competition due to personal reasons, or are ruled out due to physical injury. Given such small frequencies, we are unable to empirically exploit, for example, the event of a serious injury to contestants as a potential exogenous shock to team gender composition.

A2. Archival Data Sources

Publically available information and archives related to the analysed reality television competitions are accessible from the following online sources and links therein:

Study 1: The Apprentice

http://www.nbc.com/the-apprentice

http://www.bbc.co.uk/programmes/b0071b63

https://en.wikipedia.org/wiki/The_Apprentice_(TV_series)

https://www.youtube.com/user/ApprenticeNBC

https://www.youtube.com/show/theapprentice/videos

Study 2: Survivor

http://www.cbs.com/shows/survivor

https://en.wikipedia.org/wiki/Survivor_(U.S._TV_series)

https://en.wikipedia.org/wiki/Survivor_(TV_series)

http://survivor.wikia.com/wiki/Main_Page

https://www.youtube.com/user/SurvivorOnCBS

Study 3: Hell’s Kitchen

http://www.fox.com/hells-kitchen

https://en.wikipedia.org/wiki/Hell%27s_Kitchen_(U.S._TV_series)

https://en.wikipedia.org/wiki/Hell%27s_Kitchen_(UK_TV_series)

Table A1: Summary of Contestant Characteristics, *The Apprentice* (N=571)

Variable	**Description**	**Mean**	**SD**	**Min**	**Max**
Age	Years of Age	33.43	10.17	19	75
Gender	= 1 if Male	0.50	0.50	0	1
Skill / Relevant Occupational level	= 0 if Low Level 3 if High Level	1.09	0.85	0	3
Race	% White	0.85			
	Black	0.06			
	Asian	0.03			
	Other	0.06			
Country of production	% Asia	0.02			
	Australia	0.11			
	Ireland	0.11			
	New Zealand	0.02			
	United Kingdom	0.31			
	United States	0.43			

Notes: The number of individual contestants in the studied sample is denoted by *N*.

Table A2: Summary of Contestant Characteristics, *Survivor* (N=735)

Variable	**Description**	**Mean**	**SD**	**Min**	**Max**
Age	Years of Age	33.04	10.04	18	75
Gender	= 1 if Male	0.51	0.50	0	1
Skill Level (within-competition)	Number of individual challenges won	0.86	1.41	0	8
Race	% White	0.79			
	Black	0.09			
	Asian	0.08			
	Other	0.04			
Country of production	% Australia	0.04			
	Philippines	0.05			
	South Africa	0.07			
	Sweden	0.10			
	United Kingdom	0.04			
	United States	0.70			

Notes: The number of individual contestants in the studied sample is denoted by *N*.

Table A3: Summary of Contestant Characteristics, *Hell's Kitchen* (N=281)

Variable	**Description**	**Mean**	**SD**	**Min**	**Max**
Age	Years of Age	30.59	6.20	19	51
Gender	= 1 if Male	0.51	0.50	0	1
Occupation / Relevant occupational level	= 0 if Low Level 3 if High Level	1.65	0.88	0	3
Race	% White	0.85			
	Black	0.07			
	Asian	0.02			
	Other	0.06			
Country of production	% Finland	0.05			
	Italy	0.12			
	United States	0.83			

Notes: The number of individual contestants in the studied sample is denoted by *N*.

Table A4: Summary of Team Characteristics on *The Apprentice*

Variable	**Description**	**N**	**Mean**	**SD**	**Min**	**Max**
Age	Average age of team members	794	32.96	7.47	23.8	58.33
Skill level	Average skill level of team members based on relevant occupation / experience (e.g., company director, marketing executive, real estate broker, law student)	794	1.10	0.67	0	3
Shared race	Proportion of white team members	794	0.85	0.20	0	1
Shared hometown	Highest proportion of team members from the same geographical region (hometown)	396	0.33	0.16	0.11	1

Table A5: Summary of Team Characteristics on *Survivor*

Variable	**Description**	**N**	**Mean**	**SD**	**Min**	**Max**
Age	Average age of team members	944	32.61	4.00	23	50.8
Skill level	Average number of individual challenges won by team members during the competition	944	1.04	0.62	0	3.33
Shared race	Proportion of white team members	944	0.78	0.27	0	1
Shared hometown	Highest proportion of team members from the same geographical region (hometown)	878	0.31	0.17	0.10	1

Table A6: Summary of Team Characteristics on *Hell's Kitchen*

Variable	**Description**	**N**	**Mean**	**SD**	**Min**	**Max**
Age	Average age of team members	636	30.08	2.44	25	36.33
Skill level	Average skill level of team members based on relevant occupation / experience (e.g., executive chef, line cook, stockbroker, stay-at-home father)	636	1.76	0.47	0.33	2.8
Shared race	Proportion of white team members	636	0.84	0.18	0.25	1
Shared hometown	Highest proportion of team members from the same geographical region (hometown)	636	0.34	0.12	0.13	0.75

Table A7: Estimated Probability of Team Success on *The Apprentice*, by task type (*alternative team gender reference category*)

	Design task				Trading task				Marketing task			
	b	*P*	*b*	*P*	*b*	*P*	*b*	*P*	*b*	*P*	*b*	*P*
All male (*reference*)												
Majority male	-0.180	0.117	-0.183	0.227	0.072	0.391	0.098	0.318	-0.012	0.898	-0.028	0.827
Equal mix	-0.236*	0.053	-0.247	0.125	-0.062	0.522	-0.087	0.434	-0.076	0.463	-0.088	0.538
Majority female	-0.163	0.157	-0.168	0.287	0.090	0.284	0.111	0.326	-0.051	0.591	-0.014	0.914
All female	-0.446***	0.010	-0.477**	0.032	0.038	0.774	0.029	0.859	0.051	0.719	0.196	0.262
Other controls	No		Yes		No		Yes		No		Yes	
Observations	208		200		306		304		280		278	

Notes: Probit regression model. Average marginal effects (*b*), with corresponding p-values (*P*). Robust standard errors clustered at the Country, Season, and Episode level. Dependent variable equals 1 if team won the task/challenge, 0 otherwise. ‘All male’ team gender category is the omitted/reference group. *Other controls* capture team-level characteristics such as average age; average skill/relevant occupational level; proportion of white team members; and gender of the project manager. Also included are season, episode, and country of production fixed effects, and an indicator for the celebrity version of the show. *P < 0.10; **P < 0.05; ***P < 0.01.

Table A8: Estimated Probability of Team Success on *Survivor*, by task nature

	Intellectual task				Physical task			
	b	*P*	*b*	*P*	*b*	*P*	*b*	*P*
All male (*reference*)								
Majority male	0.100	0.569	0.108	0.644	-0.142	0.279	-0.149	0.230
Equal mix	0.163	0.338	0.119	0.510	-0.149	0.259	-0.167	0.186
Majority female	0.225	0.222	0.157	0.558	-0.215	0.113	-0.242*	0.062
All female	0.250	0.445	0.134	0.689	-0.167	0.447	-0.149	0.484
Other controls	No		Yes		No		Yes	
Observations	175		175		727		727	

Notes: Probit regression model. Average marginal effects (*b*), with corresponding p-values (*P*). Robust standard errors clustered at the Country, Season, and Episode level. Dependent variable equals 1 if team won the task/challenge, 0 otherwise. ‘All male’ team gender category is the omitted/reference group. *Other controls* capture team-level characteristics such as average age; average skill level; and proportion of white team members. Also included are season, episode, and country of production fixed effects; a challenge type (immunity) indicator variable; and an indicator for the celebrity version of the show. *P < 0.10; **P < 0.05; ***P < 0.01.

Table A9: Estimated Probability of Team Success on *Survivor*, by challenge type

	Reward challenge		Immunity challenge	
	b	*P*	*b*	*P*
All male (*reference*)				
Majority male	0.020	0.900	-0.170	0.191
Equal mix	0.029	0.861	-0.188	0.149
Majority female	-0.052	0.757	-0.230*	0.092
All female	0.195	0.471	-0.209	0.392
Other controls	Yes		Yes	
Observations	372		530	

Notes: Probit regression model. Average marginal effects (*b*), with corresponding p-values (*P*). Robust standard errors clustered at the Country, Season, and Episode level. Dependent variable equals 1 if team won the task/challenge, 0 otherwise. 'All male' team gender category is the omitted/reference group. *Other controls* capture team-level characteristics such as average age; average skill level; and proportion of white team members. Also included are season, episode, and country of production fixed effects; nature of task indicators (intellectual/physical); and an indicator for the celebrity version of the show. **P* < 0.10; ***P* < 0.05; ****P* < 0.01.