Woman Against Woman?: An Experimental Approach to Single-Gender Competitive Performance

Victoria Cragg Peterson

February 19, 2018
Carleton College, Department of Economics
Senior Integrative Exercise
Advisors: Jonathan Lafky & Jenny Bourne

ABSTRACT

Although the literature in experimental economics on gender dynamics is growing, there are few single-gender experiments. Competition between men and women is a popular topic for study because of its implications toward real-world issues of workplace gender diversity. Little research has been done, however, on how women compete in single-gender scenarios. Single-gender competition among women is an under-researched facet of the literature. Expansion in this field is important to better understand gender differences. This experiment follows from the design of Niederle and Vesterlund (2006) but samples only college-aged women in a self-selection tournament game. This study finds not only that forced competition does not increase performance in women, but also that agency of self-selection has a measured effect on performance. This study also finds that under-confidence is an important consideration for tournament entry and performance in women.

I. Introduction:

Although literature exists exploring competitive tendencies between men and women, the field lacks research surrounding how women compete directly with other women. Women are sometimes perceived as willing to compete against one another in negative ways, acting out through indirect aggression (Gordon, 2015; Campbell, 2004). Women celebrities are chastised for criticizing one another, especially when such criticisms can hurt another's career. At the same time, some push to create a community of women seeking to empower other women, not to compete but instead to collaborate. Two narratives emerge; one that pits women against one another, and one where women do not compete at all.

Experimental evidence has so far been inconclusive on how women compete, and to what extent. This experiment assesses a previously unaddressed topic, that of single-gender self-selected competition in women. Building on existing experimental designs, this research explores what happens when all men are removed from the experimental setting and women are first forced to compete, then later given a choice to compete. This study finds that agency, the ability to choose, affects a woman's performance in competition and, surprisingly, that women are under-confident in their abilities. These findings are extremely significant on the individual level, but are also important to consider for structuring policy and institutions to best support women.

II. Review of Existing Literature and Underlying Theory

The majority of assessments on how women compete result from experiments with mixed gender groups. One explanation for why women do not prefer to compete focuses on differences in ability between genders. The explanation that men perform inherently better than women is the hypothesis most easily discredited. In non-competitive experiments men and women perform equally well at the same task, meaning that women are not inherently worse at performing tasks than men (Gneezy et al. 2003; Niederle and Vesterlund, 2006; Gneezy and Rustichini, 2004).

Although stereotypes indicate that men may be more effective workers in fields of finance, research suggests no gender difference in the performance of fund managers. Portfolios managed by women perform just as well as portfolios managed by men (Atkinson et. al, 2003). Furthermore, for the same reason women may be overlooked for career advancement (being overly risk-averse), women may manage stock portfolios more effectively than men. Risk-aversion leads to more prudent investment decisions and a lack of gross overconfidence means that women do not make as many risky investments or trades as men do (Jianakoplos and Bernasek ,1998; Levy et al, 1999). One study finds that women trade stock 45 percent less frequently than men, but that men make more inefficient and overconfident investments (Barber and Odean, 2001). On the whole, these portfolios tend to even out, but large risks can lead to inordinately large payoffs that women who make small risks won't see.

It is also possible that these perceived differences in ability between men and women are a result of a self-fulfilling stereotype threat. Stereotype threat occurs when a person or group conforms to commonly held negative stereotypes (Spencer et al, 1999). When asked to complete the same real-effort task, like computational math, men and women do not differ in ability (Gneezy et al. 2003; Niederle and Vesterlund, 2006; Gneezy and Rustichini, 2004). Despite the

lack of a measurable difference in performance in these real-effort tasks, women still face discrimination especially in traditionally masculine fields, and are often expected to perform worse than men because of these stereotypes (Moss-Racusin et al., 2012).

Discrimination causes hardship that can add to preexisting stresses of participating in a competition. For this reason, women may elect to not compete, especially in tasks where they have been stereotyped to lose. Unwillingness to compete, or lack of performance while competing, makes a compelling argument for why women choose not to enter, or are overlooked in, highly competitive fields (Niederle and Vesterlund, 2006). This explanation of disinterest in competition asks the question "Are women unwilling to compete?" This question has important implications for women entering the job market.

It is generally accepted that in the hiring or promotion processes at a large company some amount of competition exists that a potential candidate must undergo to pursue a position. Many corporations have highly competitive hiring and promotion processes involving many interviews and steps. At some point during the hiring or promotion process a candidate is directly, or indirectly competing with another candidate for a single position. As adults, women are generally found to underperform in competitive environments compared to men, even if they perform well in a similar non-competitive environment (Gneezy et al. 2003). This finding might lead one to believe that women are perhaps inherently less competitive than men, or that women are conditioned to be less competitive. Because literature on single-gender competition is uncommon in experimental economics, it is useful to consider findings from mixed-gender experiments to understand what factors contribute to a woman's decision to compete.

To test this theory of learned aversion to competition one would look at the performance of children in competitive environments. As children age into adolescence, gender differences

arise in competitive performance. Girls from single-gender schools outperform girls from coeducational schools in tournament settings, suggesting that the girls' underperformance during competition is based on social conditioning rather than biological differences (Booth and Nolen, 2009). The task conducted in the Booth and Nolen (2009) study is independent of stereotype threat (the participants solved mazes under a time constraint), unlike previous studies conducted. Although running is a simple and familiar task for children, it can be considered a boy's activity, which can confound the observed results. When using this running task in a coeducational school, girls do not perform as well in a competitive environment as boys do (Gneezy and Rustichini, 2004). Although these environments see no change in the girls' performance, studies conducted in different countries find the opposite, suggesting that women from different countries behave according to different social norms.

In some countries no concrete difference is present in performance across genders. In Sweden, when traditional "girls'" activities such as skipping rope are added into the competitive set, no difference in performance arises. This similar performance by boys and girls suggests that the social conditioning women face is specific to the environment in which they are placed. (Dreber et al. 2011). Both studies, Booth and Nolen (2009) and Dreber et al. (2011), suggest that the lack of elevated performance in competitive environments for women is based on conditioning rather than biological predispositions. Although these studies highlight conditioning as a reason for the gender gap in performance, they do not explore the mechanism of such a gap (Booth and Nolen, 2009; Dreber et al, 2011).

One hypothesized cause of this underperformance in competitions and lack of willingness to compete is that women could be under-confident in their abilities. It is not that women are unconfident, but rather that they are less overconfident than men (Pulford and Colman, 1996).

This weakened overconfidence in women means that high-performing women enter competitive environments less frequently than lower-performing men (Niederle and Vesterlund, 2006). It is important to make the distinction that in these experimental settings women are confident in their ability to complete a task, framed without competitive language. The women are never asked explicitly if they have "won" a tournament or competition, but rather are asked to assess their absolute ability in a task. This framing is important because a women might shy away from competition if she believes that her participation will impose a negative externality on another participant, like "hurt feelings" (Niederle and Vesterlund, 2006). Because women are also demonstrated to be overconfident in their abilities, but to a lesser extent than men, it is likely that another explanation is at play for why women shy from competition.

A supporting explanation as to why a woman would not enter a competitive environment is that women are more risk-averse than men. Heightened risk-aversion can mean women do not want to compete when the outcome of the competition is uncertain. Although it has been suggested that female risk-aversion is an inherent survival instinct, it is also a learned phenomenon (Campbell 2004). It is an oversimplification to assert that all women, however, are risk-averse. Swedish women who are board directors are found to be more risk-loving than the men, showing that women in Sweden are not conditioned in the same way as other women (Adams and Funk, 2009). This contradiction, coupled with the study on Swedish children running races (Dreber et al. 2011) gives even more credence to the argument of social conditioning: that some societies create a gender gap in competitiveness and others do not. Although interesting, the argument that women have evolved to be risk-averse is not testable in an experimental setting, and is offered here only as a support to the social conditioning hypothesis.

In experiments comparing competitive tendencies between genders experimental design plays an integral role in determining what hypothesis the experiment can actually address. For example, in a coed experiment where women and men are asked to compete against one another, results cannot claim that women do not want to compete ever. These experiments can claim only that women do not want to compete with men. The gender of the competitor matters. It has been suggested that women will perform differently when men are present, potentially because they feel as if they are competing for male attention and must outperform other women (Gneezy and Rustichini, 2004; Campbell, 2004). It is important to consider the details of these environments and the competitive conditions because results can vary depending on the experimental environment. Although each experiment adds to the literature on gender and competition, they all answer slightly different questions. As mentioned previously, the country in which the experiment takes place has an effect on findings, as well as the gender of competitors and the format of the competition.

It is relatively unclear if women are more willing to compete with other women than with men. One study allows subjects to choose the gender of their competitor, but does not measure an increase in women's performance, even when they choose to compete with other women. This finding supports the theory that the gender of the competitor could be independent of competitive tendencies (Gupta et al. 2013). Although, when allowed to choose the gender they compete against, women do compete more frequently, their performance in the competition does not improve – the men still outperform the women while competing. Even though more women are willing to compete with other women than with men, no concrete explanation exists for the persistent gap in competitive performance. Another study finds, however, no such gap in competition between men and women (Eriksson et al. 2009). The Eriksson study (2009) allows

participants to choose whether or not they compete, but gives participants no information about the gender of their competitor. It is difficult therefore to refute the findings of the Gupta study (2013) because the experimental designs, although both examining self-selection into competition, are too different.

Similarly, when competition is forced, women competing in all-women groups perform better than they do when competing against men, but they still do not perform as well as men do in either group (Gneezy et al. 2003). In these instances, participants can see the other participants but are never formally notified of the gender of their groupmates. In similar settings, but under self-selection, women still do not want to compete. The Gneezy (2003) experiment is very similar in design to the Niederle and Vesterlund experiment (2006), with the main difference in design being the ability for participants to self-select into a competitive tournament during the experiment. The addition of self-selection allows the experimenter to examine tournament entry as a facet of competitive preferences that cannot be addressed with forced competition (Niederle and Vesterlund, 2006). Even in these instances of self-selecting into competitive environments, women still prefer not to compete.

Adult women, who have presumably spent a lifetime being conditioned to be less competitive, would be expected to perform similarly to the girls in the Gneezy and Rustichini (2004) study. Despite expectations, no clear consensus exists on how women compete, both in how they perform in forced competitions and how they self-select into competitions. It is unclear why women both underperform in and shy away from competition with men (Niederle and Vesterlund, 2006). Similarly, no clear answer prevails for how women perform in competitions strictly with other women, with no men present at all. The next step for research in this field is to explore how women compete in groups made up of only other women when they are given the

option to self-select into a competitive environment. By building on the design of the Niederle and Vesterlund study (2006), using a self-selection tournament style experiment, I will attempt to further answer the question of how women compete against other women.

III. Experimental Design

The participants in this experiment are all women students at **<Removed for MEA Competition>** aged roughly 18-23. Participants are selected from an initial survey in order to allow participants to self-identify their gender. Self-identification is important to this experiment to remove the risk of misgendering, and is done using the survey to avoid demand effects of asking a room full of participants to divide based on gender. Demand effects occur when participants try to alter their behavior based on what they believe is appropriate for the experiment. Participants are scheduled for experiment days over the course of one week. All experiments take place in the same room. Gender is not discussed at any point during the experiment except in the initial interest survey. Group sizes are restricted to four or eight participants so as not to let participants suspect that only women are being sampled. Participant attendance is confirmed several times via email before the start time of each experiment but no information about the experiment is given.

This experiment closely follows the design of the 2006 Niederle and Vesterlund study with the major modification that the participants are only women, whereas the original design samples both men and women. The sampling of only women contributes to the answering of the question "Do women shy from competition even in single-gender environments?" The experiment is conducted in three phases, each phase containing the same task. In this experiment, the real task is to sum sets of two-digit numbers using pen and paper. Participants are given four

minutes to complete each phase of the task. This time period was determined by conducting several time trials to assess how many sums would be feasible in a given amount of time. The two-digit numbers are randomly generated by the experimenter and are uniform for each participant within the phase, but are different across phases themselves.

After the completion of each phase, participants are informed of their raw score of correct questions, but are not notified of their position relative to the rest of the participants. The three phases of the experiment are piece-rate, tournament, and choice, which employ the same real-effort task, but vary in payment scheme.

In the piece-rate scheme participants are compensated per each correct sum at a rate of $0.25 per correct sum. This phase is not a competitive experiment round and subjects are not aware they will be competing in later rounds. In the second phase, the tournament participants are asked to compete with their group mates for payment. A group is defined as four participants seated in a row, and the participants are made aware of who is in their group before the start of phase two. The winner of the tournament, the individual that sums the most correct sets, receives $1.00 per correct response while all others receive nothing.

The phase three tournament allows participants to choose what payment scheme they will apply to the round: piece-rate or tournament. If a participant chooses piece-rate she is paid $0.25 per correct question, just as in phase one. If she chooses tournament, however, her performance is evaluated relative to the performance of her three other group members in the phase two tournament. If a participant solves more problems in phase three than her three group members did in phase two, she is paid $1.00 per correct response. She is paid nothing if she does not sum more correct problems in phase three than her group members did in phase two. This phase three design is particularly important because it compares performance to the previous round. Multiple

women can choose “tournament” in phase three, but their phase three scores are never directly compared. This scoring method eliminates the potential for a participant to select piece-rate out of fear of not wanting to cause another group member to receive no payout for that phase.

Payment is calibrated such that a participant with a 25 percent chance of winning the tournament will receive the same payout as the piece-rate round. This payment scheme disincentives a lack of effort in one scheme over another, assuming that participants will attempt to maximize payout. Final payout for each participant is the sum of the show-up fee, completion fee, and any payment from the randomly selected phase. At the end of the experiment one phase is randomly selected for payment by random number generator.

The experiment concludes with an exit survey that asks a participant to self-assess her performance in phases one and two. Participants are asked to select a number between one and four, corresponding to what they believe their rank is in phases one and two. This facet of the experiment is used to assess levels of overconfidence as suggested by the literature. Overconfidence is defined as a participant thinking she has a higher rank than she actually does, and under-confidence is thinking her rank is lower than it actually is. The survey is not incentivized due to budget restrictions but is useful in delineating sources of competitive preferences.

IV. Data and Results

A. Data

Data are collected on each participant for the number of problems correctly solved in each phase, as well as her choice for payment scheme in phase three and answers to the exit survey questions. The experiment samples forty participants total, broken up into ten tournament groups, all of which are women from various majors and class-years. No economics majors are

allowed to participate, although undeclared potential majors are not prevented from participating in the experiment. Although it would be possible to collect data on the number of attempted sums rather than correct sums, it is more relevant to look at the correct sums. The number of correct sums is relevant in determining tournament winners, and the task would not be a real-effort task if any response were coded the same as a correct response. This variable for correct responses is the raw score value for the number of correctly summed sets for each participant. The mean numbers for correctly summed sets for each phase are expressed in Table 1 below.

	Phase 1	Phase 2	Phase 3
	M (SD)	*M (SD)*	*M (SD)*
Correctly Summed Sets	7.65 *(1.99)*	7.875 *(2.70)*	9.025 *(2.57)*

Table 1: Correctly Summed Sets Across All Phases

B. Formal Hypotheses

The original hypotheses for this experiment are that women's performance does not change under competition, and that women who self-select into competitive environments perform no differently than women who do not select competition. The first null hypothesis states that the mean amount of sets summed in phase one (μ_{P1}) is equal to the mean amount of sets summed in phase two (μ_{P2}). The second null hypothesis states that the mean number of sets summed by the phase three piece-rate choosers (μ_{P3P}) is equal to the mean number of sets summed by the phase three tournament choosers (μ_{P3T}) These hypotheses are expressed as such;

$$H_{0a}: \mu_{P1} = \mu_{P2}, \quad H_{1a}: \mu_{P1} \neq \mu_{P2}$$

$$H_{0b}: \mu_{P3P} = \mu_{P3T}, \quad H_{1b}: \mu_{P3P} \neq \mu_{P3T}$$

We expect to see no difference in the mean number of questions correctly summed between phases one and two, and also no difference in the amount of correctly summed sets between participants who select piece-rate versus tournament in phase three. These hypotheses are both selected as two-tailed tests because the existing literature suggests no concrete directionality for the difference in means for women competing. The literature is clear that one should expect a man's performance to increase with competition, but no such claim can be asserted for women. The two-tailed test is used for both hypotheses as well as additional tests of statistical significance. A 95 percent significance level is also used for the results of this experiment. Because this data may not be normally distributed, both a t-test and Mann-Whitney U test are used to determine statistical significance.

	Phase One *M(SD)*	Phase Two *M(SD)*	*t-Value*	*Prob (t-Test)*	*U-Value*	*Prob (M-W U)*
Correctly Solved Sets	7.65(1.99)	7.88(2.7)	-0.42	0.67	-0.44	0.66

Table 2: Difference in Means Between Phase One and Phase Two

The average difference between the whole group means between phase one and phase two is 0.225 correct sums. This is an extremely small difference and both the t-test and Mann-Whitney U test confirm that this difference is not significant at the 95 percent level. This is shown in Table 2. With this insignificant result, one cannot reject the null hypothesis that there is no difference in means between phase one and phase two. This result is somewhat consistent with the literature, suggesting that forcing women to compete does not increase their performance. Although this hypothesis test does not yield statistically significant results, finding that a woman's performance does not spike when she is forced to compete is useful to assess

how current hiring systems may be ill-suited to gender diversity in the workplace. If hiring is framed as a forced competition it seems less likely that a woman will see an increase in performance as a man does. Agency of choice in entry into competition will be further explored in the next section.

	Piece-Rate Choosers *M(SD)*	Tournament Choosers *M (SD)*	*t-Value*	*Prob (t-Test)*	*U-Value*	*Prob (M-W U)*
Correctly Solved Sets in Phase Three	8.36(2.58)	10.58(1.78)	-2.71	0.01	-2.56	0.01

Table 3: Difference in Phase Three Means by Payment Scheme Choice Type

The second formal hypothesis looks only at phase three. In phase three participants choose between piece-rate or tournament payment scheme for how they will be evaluated in their task. This hypothesis compares the difference in mean sets solved correctly between the piece-rate and tournament choosers in phase three. Both a t-test and Mann-Whitney U test are used because the distribution is unknown. Twelve participants select the tournament option, and twenty-eight select piece-rate. These small sample sizes make the Mann-Whitney U test a more appropriate choice for the difference in means, but both tests are conducted for the sake of comparison. The results of these tests can be seen in Table 3. The two-tailed Mann-Whitney U test for the difference in means within phase three (the means of piece-rate participants versus the tournament participants' means) yields a p-value of 0.01. The t-test also indicates statistical significance with a p-value of 0.01. This result is statistically significant at the 95 percent confidence level; therefore, we can reject the null in favor of the alternative hypothesis, that there is a difference in performance between piece-rate and tournament choosers in phase three.

The mean amount of sets summed by the piece-rate group is not the same as the mean amount of sets summed by the tournament group; in fact, the tournament group outperforms the piece-rate group. This is not surprising because one could assume that participants who are better at the task will compete in the tournament. To further explore this result it is important to analyze the performance of the piece-rate choosers in comparison to the tournament choosers to see if the participants that choose tournament are indeed the higher-performing participants. Although the primary focus of the study is the aforementioned hypotheses, it is also useful to investigate other findings in the data. Existing literature does not explicitly address the role of agency in tournament entry, nor does it assess instances of gross under-confidence in women. It is necessary to assess the impact of self-selection on competitive performance, as well as the results of the exit survey, to achieve a more thorough exploration of the implications and conclusions of the experiment. Participant performance in the self-selection component of the experiment indicates the importance of choice in tournament entry.

C. Performance with Self-Selection

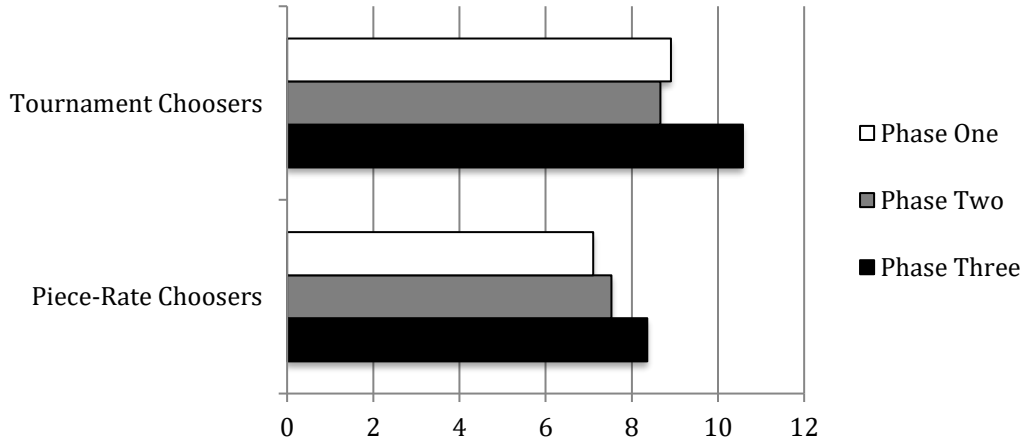

Figure 1: Mean Correct Sets By Phase Three Payment Choice

Given the lack of a significant difference in means between phases one and two, and the significant difference in performance between the two groups of phase three, it would be prudent to examine the phase three data more closely. Surprisingly, the performance of all participants together increases from phase two to phase three; that is, regardless of their payment scheme choice, all participants perform better by being allowed to choose. This improvement is displayed visually in Figure 1. Performance increases on the whole by 1.15 correct sums from phase two to phase three, a difference in means that is significant at the 95 percent level (p-value 0.05). The null hypothesis is that the mean number of sets solved in phase two is equal to the mean number of sets solved in phase three. This significant difference in means from phase two to phase three suggests a fundamental difference between phases two and three that is not present between phases one and two. Furthermore, performance also increases between phase one and phase three. Phase three is different because it gives participants agency for how they are paid. To further examine the performance spike in phase three, one must look at the piece-rate choosers (those who select piece-rate in phase three) compared to the tournament choosers (those who select tournament in phase three) to see the differences between the groups.

D. Change in Performance Over Rounds Based on Payment Choice

	Phase 1	Phase 2	Phase 3
	M (SD)	*M (SD)*	*M (SD)*
Piece-Rate Choosers	7.10 *(1.61)*	7.53 *(2.80)*	8.36 *(2.58)*
Tournament Choosers	8.91 *(2.27)*	8.66 *(2.38)*	10.58 *(1.78)*

Table 4: Mean Correctly Summed Sets by Piece-Rate and Tournament Choosers Over All Phases

Although it is true that no statistically significant difference in means occurs between phases one and two for the whole group, mean performance does increase by roughly 0.4 correct sums from phase one to phase two. For the tournament choosers, however, performance declines from phase one to phase two by 0.3 sums while performance for the piece-rate choosers increases from phase one to phase two. These changes in performance can be seen in Table 4, where the mean number of correctly summed sets is reported for each phase by payment choice group. Perhaps more interesting still is the change in performance between phase two and phase three. All participants perform the best in phase three, and see improvement between phases one and three as well as between phases two and three. On the group level, only the improvement by the tournament choosers is statistically significant. Although both groups (piece-rate and tournament choosers) increase in mean raw score from phase two to phase three, only the increase by the tournament choosers is statistically significant. Mean score for the tournament choosers increases by 2.12 responses, a difference which is highly significant (p-value 0.023). This result means that the tournament choosers are significantly outperforming the piece-rate choosers in the third phase.

E. Exit Survey

The exit survey is included primarily to reaffirm claims put forth by the literature that people are overconfident. The existing literature suggests that all people are overconfident in their abilities, but men are more overconfident than women. One expects the results of the experiment to reflect the same. This experiment finds the opposite: the participants on the whole are overwhelmingly under-confident in their abilities. The exit survey asks participants to assess their performance and estimate what their rank is in phase one and phase two. If participants think their rank in the phases is higher than their actual rank, they are deemed to be

overconfident. Question one of the survey refers to the rank of phase one and question two referred to the rank in phase two. The percentages for overconfident, correct, and under-confident survey responses are expressed in Figure 2.

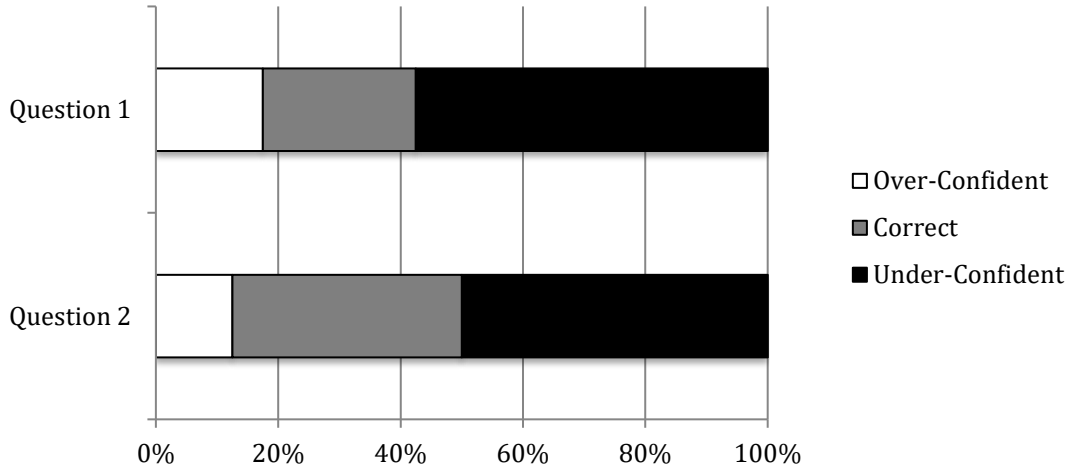

Figure 2: Confidence in Exit Survey Responses

In question one, no participants indicate that they believe they rank first. Fifteen participants indicate that they believe they are second, twenty believe they are third, and five believe they are last. Even participants who answer twelve correct questions (of a total max of thirteen) respond that they do not think they perform the best or even second best in their group. In question one only 17.5 percent of participants are overconfident, that is, overestimate their rank in the tournament, and 12.5 percent of participants are overconfident in their abilities in the phase two exit question. One would perhaps expect the tournament choosers to be more overconfident in their abilities than the piece-rate choosers, but this is not always the case. Despite overconfidence suggested by the literature, these results clearly contradict the norms in confidence for experiment participants. It important to note, however, that this survey is not incentivized, but still provides an avenue for further discussion as to why this study finds such unexpected results. A more complete explanation of the drawbacks to a non-incentivized survey is further discussed

in the limitations section of the discussion. As previously addressed in the review of literature, social conditioning is a major factor in how women behave under competitive constraints, but the forms of conditioning at play in this experiment appear to manifest differently from previous experiments.

V. Discussion

A. No Increased Performance With Forced Competition

Like much of the existing literature on the topic, the results of this study do not definitively answer any questions about how women may or may not compete, nor do they fully support or refute the hypotheses of previous experiments. The least surprising finding is the difference in means between phase one and phase two. This difference in means is found to be statistically insignificant which is expected given historical evidence. Past research suggests that women's performance does not improve with competition if the women are from an environment that has conditioned them to be more docile and unwilling to compete. It is no surprise that these women behave similarly to other women in American experiments instead of like the women in Swedish studies. Even more compelling is that neither the tournament choosers nor the piece-rate choosers perform significantly better between phases one and two. Although these findings are not statistically significant, they are highly practically significant for institutions like **<Removed for MEA Competition>** . For example, making **<Removed for MEA Competition>** "standardized test optional" would remove an aspect of forced competition from the admissions process and could increase the number of applications from women.[1]

[1] I recognize that **<Removed for MEA Competition>** has over 50 percent women, but increased applications from women could also mean an increase in applications from women of color, women international students, women LGBTQ students, and other traditionally underrepresented groups.

Scaling past <**Removed for MEA Competition**>, this finding implies that women do not benefit from competition in the same way that men do. This difference means that during competitive hiring processes men see a spike in performance in response to competition, whereas women do not. This is especially concerning in this group of participants who will soon enter the job market and will be forced to compete for limited positions. It is important to note, however, performance does not decline with competition, but rather does not improve. This is only concerning when compared with men that do see an improvement, where a woman's ability can be misrepresented because of her lack of improvement with competition. It is not that women are unable to perform, but as compared to men they can be perceived as less able. Another iteration of this experiment will be necessary to determine if these women do not see a spike in performance when also competing against men. The findings from this experiment alone cannot make that claim for the women of <**Removed for MEA Competition**>.

B. Differences in Performance: Piece-Rate vs. Tournament

The second formal hypothesis of the experiment assesses the difference in performance between the two payment-scheme-choice groups, the piece-rate and tournament choosers. The tournament choosers outperform the piece-rate choosers. This result is expected because one assumes that the more qualified participants will choose the tournament in efforts to maximize payout. This is somewhat the case. In phase one the tournament choosers sum an average 8.72 sets, where the piece-rate choosers sum 7.24 sets. This is a statistically significant difference (p-value 0.006) but no such significant difference is found for the difference in means in phase two (p-value 0.23). This supports the hypothesis that competition does not increase performance in women, and furthermore suggests that the tournament choosers are not inherently "better" at the

task than the piece-rate choosers. Differences in performance arise in the phase three round. The difference in performance between the tournament and piece-rate choosers in phase three is statistically significant (p-value 0.05) but is difficult to explain given the fact that the tournament choosers are not inherently better at the task than the piece-rate choosers.

A compelling explanation for this change in performance is the introduction of agency on the part of the participant. Until the final phase, participants are given a set of constraints by which they are to be evaluated and are not allowed to choose until phase three. Both groups, piece-rate and tournament choosers, improve in performance in phase three. The piece-rate choosers see an increase in performance that is not statistically significant, unlike the tournament choosers, whose improvement is statistically significant. This is particularly interesting because the tournament choosers do not improve when forced to compete, but do improve when they elect to compete, indicating that agency has an impact on a woman's performance. On the whole, performance increases from phase two to phase three, indicating that it almost doesn't matter which payment scheme is chosen, just that a choice is made. The presence of choice and agency result in increased performance.

What is more interesting still is why some women choose to compete while others don't. To some degree, this is a question of preferences and cannot be answered by this experiment, but it is also a question of risk preferences and confidence. It is likely that the more risk-loving, confident participants will select tournament in phase three because they are unafraid of a no-payment outcome. It is also possible that some of the risk of no-payment is removed by the presence of a show-up and completion fee; participants know that they can take a risk on the tournament and will still be paid. Rather than frame the question as to why women will enter the tournament, in this case it is more relevant to consider why the majority of women do not select

the tournament. In this experiment it is revealed that under-confidence may inhibit tournament entry, but it is not suggested that over-confidence promotes tournament entry.

C. Under-Confidence and Social Conditioning

The results of the exit survey suggest that women do not enter the tournament because they are under-confident in their abilities. Risk preferences are also a cause of the lack of tournament entry, but the explanation of under-confidence is both unusual for an experiment of this type and more quantifiable than the risk-aversion argument that relies heavily on preferences not directly measured in this experiment. The under-confidence exhibited in the exit survey is highly unusual especially when compared to the results of the Niederle and Vesterlund (2006) study, which finds its participants to be overconfident in their abilities. It is surprising to find that 57.5 percent of participants are under-confident in their abilities in phase one, and 50 percent of participants are under-confident in their abilities in phase two. It is possible that participants are bad at guessing their rank but, even if that is true, it is unusual that they would mostly underestimate. Because this under-confidence is so unusual, it is likely a result of environmental factors.

Social conditioning is often given as an explanation for why women prefer not to compete. In this experiment, under-confidence is clearly a relevant factor in dictating tournament entry. Unexpectedly, the individuals that are overconfident in their abilities are not all tournament choosers. In fact, no tournament choosers are overconfident in their abilities in phase one, and only two tournament choosers are overconfident in their phase two ability, indicating that for these women, their risk-loving preferences are likely the cause of their tournament entry, overpowering their under-confidence.

It has been suggested that men and women have differing views on what it means to be competent in a task or field, especially when that task is typically gendered as a man's task (Cech et al 2011). This under-confidence as displayed by the participants of this study is likely a result of social conditioning. What is unclear, however, is if the environment of **<Removed for MEA Competition>** is a major factor in the shaping of these young women's perceptions of themselves or if the conditioning to be under-confident is occurring prior to entry into the institution.

D. Limitations and Considerations

As previously mentioned, the idea of a gendered task can have an effect on the performance of participants. This effect, known as stereotype threat, often comes into play when comparing performance between men and women. Because this study samples only women, stereotype threat should affect all of the women in the study, but can still be a source of bias in the results. Because computational math can be considered a task that favors men, the women in the study may experience a sense of under-confidence in their abilities. This could help explain why the women in the study exhibit an unexpected level of under-confidence. Despite the potential for stereotype threat in this experiment, computational math is the most prudent choice for a real-effort task. Solving mazes is a common real-effort task, but is more difficult to swiftly correct by hand and can take much longer than computational math problems. In future iterations of the experiment it could be necessary to include more than 13 sets of sums so that participants can complete even more sets. More sets are not included in this experiment because of budgetary restrictions, but adding several more sets will not likely alter results significantly. Only two women complete all thirteen sets, and do so with few seconds remaining. It is possible that these

two participants could sum an additional set or two, but this increase in performance will not likely change the results.

In addition to the presence of stereotype threat in the real-effort task, there are other considerations about the experiment that are worth addressing. In this experiment every effort is made to ensure reproducibility of results. Unfortunately, there are some variables that cannot be controlled. For example the experiment room was cold several nights for the experiment. It has been suggested that temperature can have an effect on productivity and could make the experimental environment less reliably uniform across all sessions (Kaufmann et al 1999). The temperature of the room was not considered in the experiment's initial design and was unavoidable once testing began.

Because the experiment can only be conducted in multiples of four (with each group having four members) a major consideration for this experiment is ensuring that enough people show up for each session. Because a lack of participants can result in not being able to run the experiment at all, the experimenter needs to ensure that backup participants are available in the event of no-shows. In the event that too many people come to the experiment the extra participants still need to be paid the show-up fee, adding to the total cost of the experiment. It is generally accepted in experiments that some participants will not show up for their specified time, but in this experiment very few participants did not show up, making the show-up fees an inefficiency that could have been better used to plan for additional sessions of the experiment. In one experiment all participants arrived on time, with no extras, but one of the women was a minor, and had to be let go as well as her entire group. The IRB did not give explicit approval to run the experiment with minors and the participant's guardians could not have been contacted for signature at the time of the experiment, so an entire group of participants was lost. In this case, it

would have perhaps been easy to pull a woman from the hallway outside the experiment room, but this would be impossible due to risks of misgendering subjects. This issue of misgendering subjects will be further addressed in the appendix.

Monetary constraints were a concern for this experiment. There is a balance between getting enough subjects to have valid results and a large enough subject pool, but also an upper bound on the amount of money that can be spent on the experiment. It is difficult to estimate average payout for each group because payment varies wildly between phases. For example, phase one is generally low paying, because each participant receives $0.25 for each correct response, but if phase three is chosen, and many participants choose tournament, there can be several winners of the phase, with a maximum payout per participant of $18.00. The wage reflected by the work done in the study also needs to reasonably fit within the confines of both the minimum wage of the institution as well as norms in the discipline of experimental economics. The average payout is $8, with the experiment time ranging from 30 minutes to 45 minutes. Because of these considerations of wage and budget, the decision was made not to incentivize the exit survey. Because the survey is not incentivized, the results cannot be analyzed as rigorous evidence of behavior. Despite the lack of incentives, however, the exit survey provides insights that wouldn't have been addressed within the scope of the three phases and was ultimately extremely important for the analysis of results.

Furthermore, interpersonal relationships between subjects can influence the validity of the findings. For example, teammates may be very willing to compete against one another, but a resident may be unwilling to compete against her RA. For this reason, participants who arrive and appear to know another participant (chatting in the hallway, for example) are placed in separate groups so they will not be competing against one another. Because the participants are

drawn from **<Removed for MEA Competition>** , it is highly likely that competing participants know each other in some regard. Similarly, it is very difficult to sample participants unknown to the experimenter. Efforts were made to sample strangers from the student body, but this is somewhat difficult due to the limited subject pool.

As discussed in the review of literature, it is difficult to make sweeping claims about an entire gender's behavior. These data are not indicative of every woman, but are an indication of how women behave at **<Removed for MEA Competition>**. The backgrounds of each of these women are unknown, but it would be interesting to see if differences arise depending on upbringing – for example, if a woman attended an all-girls school, is an international student, or holds leadership positions on campus. These considerations provide several different possibilities for future research, or if the experiment is to run again.

E. Opportunities for Future Research

Because of budgetary constraints this experiment was not able to sample men as well as women. Running this experiment with groups of all men, all women, and mixed gender groups will provide more information as to the causes of the results of this experiment. For example, if all participants across all genders are to under-assess their abilities, one will be able to conclude that this under-confidence is not a gendered issue. Sampling a larger group will require more budget than was feasibly available for this experiment but would be a necessary next step in determining causality for tournament performance and choice outcomes. It would also be interesting to run the experiment at different colleges, perhaps a comparison between **<Removed for MEA Competition>** and **<Removed for MEA Competition>** to determine if **<Removed for MEA Competition>** is unique in this under-confidence. An additional budget will also allow

for an incentivized exit survey that could help determine the root cause of tournament entry or lack thereof.

If the experiment were to run again with only women it will also be prudent to reconsider the kinds of hypothesis tests used. Two-tailed tests are used in this experiment as suggested by literature on how competition will affect women's performance. Using the results of this experiment as a starting point for continuing research, it would be more accurate to use a one-tailed test for a difference in means between the competitive and non-competitive phases. What is unclear, however, is what direction the tests should be if the experiment is to run again with men. Although existing experiments (Niederle and Vesterlund, 2006) have used one-sided hypothesis tests, the under-confidence exhibited by the women in this study renders those findings unhelpful in dictating the directionality of potential future hypothesis tests.

Another potential improvement on this experimental design would be the inclusion of a step for validation of ability. If, as in this experiment, women are generally under-confident in their ability it may be useful to determine if their performance increases with validation in the form of affirmation of good performance. By adding a treatment where participants are told their rank in each phase before completing the next one it may be possible to better calibrate participants' perceptions of their ability. This may alleviate some of the potential stereotype threat effects, which could also be remedied by changing the real-effort task to solving mazes.

A final avenue for additional research is the exploration of competitive preferences among non-binary individuals. As economics as a discipline begins to recognize gender identities other than the gender binary, research on competitive tendencies for non-binary individuals will be important to create more inclusive and diverse workspaces. The findings of

this study, although somewhat unexpected, provide a jumping off point for future research and raise more questions about how women compete.

VI. Conclusions and Implications

Although the results of this study are inconsistent with the existing literature, especially in regards to confidence, these results provide a new set of questions to answer for how women compete with other women. If one considers the role of agency in competition it could become relevant to alter how hiring and promotion processes are framed. If entry into the job market is framed as a choice to enter into a competition, rather than a forced tournament, it could improve a woman's performance and allow more qualified women to enter fields dominated by men. Mood and information framing can have an effect on competitive performance (Kuvaas and Kaufman, 2003), but this has not yet been studied in the context of gendered self-selected tournaments. Changing hiring policy could be as simple as reframing workplace competition to focus less on potential negative outcomes (not getting a job) and more on potential positives (experience interviewing). Studies on the reframing of stress indicate that it is possible to view a traditionally negative emotion, such as stress, in a more positive light by framing (Liu et al. 2017). If this same process could be applied large-scale to women in regards to competition, it could decrease inhibitions of tournament entry.

Furthermore, the issue of under-confidence in young women is a concerning finding especially when one considers that these self-reported feelings of inadequacy are not founded in one's ability but rather come as result of social conditioning. If we are to strive for gender diversity in upper management tracks we must address this issue of under-confidence in young women, and must begin to ask questions about our institutions. An institution such as **<Removed**

for MEA Competition> cannot hope to undo two decades of social conditioning in just four years of undergraduate study. Despite active efforts by the institution's administration to increase gender diversity among faculty, support organizations for women students, and increase resources for women on campus, women are still under-confident. By the time a woman reaches college it is likely too late to change her attitudes about competition. To start work on fixing the problem of gaps in competitive performance, economists need to backtrack to determine what we are teaching our women to make them under-confident in their abilities and unwilling to compete. Instead of asking why women do not perform well under competition we need to answer why young women feel as if they are unqualified to try.

VII. Appendix

A. Experimental Considerations on Gender

For the context of this experiment it is important to note that few, if any, economics experiments consider the implications of potentially misgendering subjects. Published papers often do not outline how women and men are selected for experiments, which can be an unintended risk to participants. Papers suggest that participants are gendered and selected based on their physical appearance at the discretion of the experimenter. Not only is this method inaccurate for data and reporting of results, but also it is risky. Misgendering a subject, or assuming subjects to be a gender that they are not, can be both triggering and negatively impactful. For these reasons, in this experiment, an initial survey is used to allow potential participants to indicate their gender (by fill-in-the-blank question), to remove the risk of misgendering a subject. It is possible that this process causes demand effects by allowing participants to realize that they are being selected based on gender, but several other questions are included in the survey to prevent participants from guessing the purpose of the survey. It is also determined that despite the potential for demand effects in the experiment it is far more important to ensure the safety of participants in regards to the sensitive topic of gender identity.

It is also important to note that existing literature in economics does not distinguish between gender and sex, and papers will often equate the terms “woman” and “female”. For this reason I have chosen to explicitly use the word “woman” to refer to subjects in existing experiments that explicitly declare to assess differences in gender, not sex. As gender and sex are not synonymous it is inaccurate to refer to all women as female, so I have consciously chosen to refer to all subjects in prior experiments as “women.” Although this may seem awkward at times for the sake of phrasing and flow, it is an important step in making the discipline more inclusive and socially relevant to historically marginalized groups.

B. Instructions for Participants

Instructions:
Welcome to this experiment in decision-making. Please read these instructions carefully as they detail how you will be paid based on your performance in today's session. Please do not talk during today's session. If you have a question, please raise your hand and an experimenter will answer your question in private.

In the experiment today you will be asked to complete a task in three different phases and an exit survey. None of these phases will take more than four minutes. There will be breaks between tasks while experimenters collect your papers. You will receive $3 as a show up fee. At the end of the experiment you will receive $2 for completing the phases, regardless of your performance. In addition, one phase will be randomly selected at the end of the experiment, and you will be compensated based on your performance in that phase. The phase will be selected by randomly drawing a number between 1 and 3, using a random number generator. The payment method varies across phases and will be explained in detail before beginning each phase.

Your total earnings from the experiment are the sum of the $3 show up fee, $2 completion fee, and any payment from the randomly selected phase.

Phase 1 – Piece-rate
For Phase 1 you will be asked to sum sets of 5 random two-digit numbers on a sheet of paper provided. You will be given 4 minutes to solve as many of these sets as possible. You may use pen and paper to solve these sums but you will not be given a calculator. Answers must be written in the space provided to the right of the numbers in the empty space. Below is an example of a set:

35	23	14	87	99	

In this example you would write "258" in the blank box.

If you cross out an answer and write a new one, the previous answer that has been crossed out will not be considered. If you cross out an answer at all, the answer that is crossed out will be considered as a question left blank. There will be no penalty for incorrect answers, so try to answer to the best of your ability where possible. Answers will only be counted if they are in the box.

A warning will be given at 10 seconds remaining, as well as a countdown from 3 seconds to 0. The experimenter will indicate the end of the round by calling "stop", at which point you will be asked to put down your pens and wait for your sheet to be graded. An experimenter will come to your station with an answer key and write your number of correct questions in the upper right-

hand corner of your page. Your score will be given back to you, but your score will be anonymous from your other group members.

If Phase 1 is randomly selected for payment, you will be compensated $0.25 for each problem you solve correctly in 4 minutes. An incorrect answer does not decrease your payout. This payment scheme is referred to as piece-rate.

Please do not talk with one another during the duration of the experiment, and do not attempt to look at the tasks or workspace of other participants. If you have questions, please raise your hand. At this time an experimenter will place your sheet face down in front of you and will also provide you with a pencil to use. Please do not start the task until you have been told to do so.

After the task is over please wait quietly while the experimenters ready the next phase. You will then receive additional instructions.

Phase 2 – Tournament

As in Phase 1 you will have four minutes to sum sets of 5 random two-digit numbers. These numbers will be different than those of the previous task though of similar difficulty. In this phase your payment depends on your performance relative to that of the other participants in your group. Each group consists of four people, seated around your table with you.

If Phase 2 is randomly selected for payment your earning will depend on the number of problems you solve relative to the three other group members. The participant who correctly solves the most questions will receive $1 for each correct response, while all others receive no payment. We refer to this as the tournament payment scheme. You will again be given your own score at the end of the round, but will not be informed of your position in the tournament until the end of the experiment. If there are ties the winner will be randomly decided among those who are tied. Two-way ties will be decided by coin flip, and all other ties will be decided by random number generator.

Please do not talk with one another during the duration of the experiment, and do not attempt to look at the tasks or workspace of other participants. If you have questions, please raise your hand. At this time an experimenter will place your task face down in front of you. Please do not start the task until you have been told to do so.

After the task is over please wait quietly while the experimenters ready the next phase. You will then be provided with instructions.

Phase 3 – Choice

As in the two previous phases you will be given four minutes to sum sets of 5 random two-digit numbers. These numbers will be different than those in task 1 or 2. In this round you will now decide which of the two previous payment schemes you prefer to apply to this round.

If Phase 3 is selected, your earnings are determined differently depending on which scheme you choose, either piece-rate or tournament.

If you select piece-rate, you receive $0.25 for each correct answer.

If you select tournament, your performance will be evaluated relative to the performance of the other three group members in the Phase 2 tournament. The Phase 2 tournament was the round immediately before this one. If you solve more problems this round than your group members did in the previous phase, then you receive $1.00 per correct problem. You will receive no earnings if you do not perform better now than the others in your group did in Phase 2. You will be given your absolute score following this task, but you will not be notified of your performance relative to the group until after the experiment. If there are ties the winner will be randomly determined among those who are tied. Two-way ties will be decided by coin flip, and all other ties will be decided by random number generator.

You will now be given a piece of paper on which you will choose the piece-rate or the tournament scheme.

Please do not talk with one another during the duration of the experiment, and do not attempt to look at the tasks or workspace of other participants. If you have questions, please raise your hand. At this time an experimenter will place your task face down in front of you. Please do not start the task until you have been told to do so.

Exit survey

This is the final aspect of this experiment. An experimenter will hand you an exit survey to complete. This exercise is not timed. When you complete your survey please sit quietly until an experimenter comes to collect it from you. After your survey is collected the experimenter will use the random number generator to select a round for payment. Once the payment round is selected you will be individually asked to step into another room to privately receive payment. After you receive payment you will be dismissed from the experiment.

Works Cited

Adams, Renee B. and Funk, Patricia, Beyond the Glass Ceiling: Does Gender Matter? (December 1, 2009). *UPF Working Paper Series*; ECGI - Finance Working Paper No. 273/2010

Atkinson, Stanley and Boyce Baird, Samantha and Frye, Melissa. 2003 "Do female Mutual Fund Managers Manage Differently?" *The Journal of Financial Research 26(1): 1-18*

Barber, Brad M. and Odean, Terrence; Boys will be Boys: Gender, Overconfidence, and Common Stock Investment, *The Quarterly Journal of Economics*, Volume 116, Issue 1, 1 February 2001, Pages 261–292

Booth, A. L. and Nolen, P. (2012), "Gender differences in risk behaviour: does nurture matter?*" *The Economic Journal*, 122: F56–F78.

Campbell, Anne, "Evolutionary and Neurohormonal Perspectives on Human Sexuality" *The Journal of Sex Research.* 2004. 41 (1), pp. 16-26

Dreber, A., von Essen, E. & Ranehill, 2011 "Outrunning the gender gap—boys and girls compete equally" *E. Exp Econ* 14: 567.

E. Cech, B. Rubineau, S. Silbey, C. Seron. Professional Role Confidence and Gendered Persistence in Engineering. *American Sociological Review*, 2011; 76 (5): 641

Eriksson, Tor and Teyssier, Sabrina, and Villeval, Marie Claire. 2009 "Self-selection and the Efficiency of Tournaments". *Economic Inquiry*, 47 (3), pp. 530-548.

Gneezy, Uri, and Aldo Rustichini. 2004. "Gender and Competition at a Young Age." *American Economic Review*, 94(2): 377-381.

Gneezy, Uri and Niederle, Muriel, and Rustichini, Aldo. 2003 "Performance in Competitive Environments: Gender Differences", *The Quarterly Journal of Economics*, 118 (3), 1 pp. 1049–1074

Gordon, E. (2015). *Opinion | Why Women Compete With Each Other*. *Nytimes.com*. Retrieved 17 February 2018, from https://www.nytimes.com/2015/11/01/opinion/sunday/why-women-compete-with-each-other.html

Gupta, Nabanita Datta, and Poulsen, Anders, Villeval, Marie Claire. 2013 “Gender matching and competitiveness: experimental evidence”. *Economic Inquiry*, 51 (1), pp. 816-835.

Jianakoplos, N. A. and Bernasek, A. (1998), ARE WOMEN MORE RISK-AVERSE?. *Economic Inquiry*, 36: 620–630. doi:10.1111/j.1465-7295.1998.tb01740.x

Kaufmann, H., Mazur, X., Fussenegger, M., & Bailey, J. (1999). Influence of low temperature on productivity, proteome and protein phosphorylation of CHO cells. *Biotechnology And Bioengineering*, *63*(5), 573-582.

Kuvaas, B., & Kaufmann, G. (2003). Impact of mood, framing, and need for cognition on decision makers' recall and confidence. *Journal Of Behavioral Decision Making*, *17*(1), 59-74. doi:10.1002/bdm.461

Levy, H., Elron, E., Cohen, A. (1999). “Gender differences in risk taking and investment behavior: An experimental analysis”. Unpublished manuscript, The Hebrew University.

Liu, J., Vickers, K., Reed, M., & Hadad, M. (2017). Re-conceptualizing stress: Shifting views on the consequences of stress and its effects on stress reactivity *PLOS ONE*, *12*(3), e0173188. doi:10.1371/journal.pone.0173188

Moss-Racusin, C., Dovidio, J., Brescoll, V., Graham, M., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings Of The National Academy Of Sciences*, *109*(41), 16474-16479.

Niederle, Muriel and Vesterlund, Lise. 2006. “Do Women Shy Away From Competition? Do Men Compete Too Much?, *The Quarterly Journal of Economics*,122 (3), pp. 1067–1101

Pulford, B. D., & Colman, A. M. 1996. “Overconfidence, base rates and outcome positivity/negativity of predicted events”. *British Journal of Psychology, 87*, 431-445.

Spencer S, Steele CM, Quinn DM. 1999. Stereotype threat and women’s math performance. *J. Exp. Soc. Psychol.* 35:4–28