THE EFFECTS OF GENDER INTERACTIONS IN THE LAB AND IN THE FIELD

Kate Antonovics, Peter Arcidiacono, and Randall Walsh*

Abstract—An important issue with conducting economic analysis in the lab is whether the results generalize to real-world environments where the stakes and subject pool are considerably different. We examine data from the game show *The Weakest Link* to determine whether the gender of one's opponent affects performance. We then attempt to replicate the competitive structure of the game show in the lab with an undergraduate subject pool. The results in the lab only match when we both employ high stakes in the lab ($\geq$ \$50) and limit our analysis to young contestants in the game show (age $<$ 33).

I. Introduction

THE labor market outcomes of men and women differ along a variety of dimensions. Explanations for these disparities abound. One explanation that has received recent attention is that men and women respond differently to competitive environments. The economic importance of behavior in competitive settings has led to a growing literature on how individual performance is affected by the gender mix of one's opponents. Because adequate performance measures are rarely available in standard data, this literature has focused primarily on experimental techniques, which have provided evidence that men and women respond differently to competition against members of the opposite sex.

Gneezy, Niederle, and Rustichini (2003) examine differences in the rate at which men and women solve computerized mazes. They find that while men always solve the mazes more quickly in a competitive environment than in a piece-rate pay scheme, competition only increases the performance of women when they compete against other women.[1] Similarly, Gneezy and Rustichini (2004) examine gender differences in the response to competition for young children (ages 9 and 10). They find that competition increases the speed at which children run, but that the effect is stronger for boys than for girls. These gender differences in responses to competitive pressures also translate into men and women making different choices about how they would like to be compensated. Niederle and Vesterlund (2007), for example, find that men are significantly more likely to chose tournament-style compensation schemes than women even when there is no gender difference in ability.

Because these laboratory experiments typically involve undergraduate college students in low-stakes competitive environments, a key question in this literature is whether these laboratory outcomes generalize to real-world settings. Our paper addresses this issue by comparing outcomes on the television game show *The Weakest Link* to outcomes in laboratory experiments designed to replicate the competitive structure of the game show. While *The Weakest Link* is itself an artificial environment, it provides an excellent starting point for examining the extent to which results from laboratory experiments can be extrapolated to other settings because the competitive environment on *The Weakest Link* is well defined and, as a result, the key features of the game show can be easily replicated in the lab. In addition, the demographic background of contestants on *The Weakest Link* and the stresses associated with competing for large sums of money and performing in front of a television audience may more closely replicate the realities of the labor market than the typical low-stakes laboratory experiment that uses undergraduate subjects.

Within the field of experimental economics, there exists a large body of work that examines how stakes affect performance. Camerer and Hogarth (1999) provide an extensive analysis of this literature, reviewing 74 experiments with varying levels of financial incentives. While Camerer and Hogarth find that financial incentives have little impact in experiments that involve strategic play such as ultimatum games and sequential bargaining, they find that financial incentives have an ambiguous impact in experiments that involve tasks such as problem solving, recall, and prediction. In particular, in experiments where the tasks are the most similar to the task we consider in this paper (correctly answering trivia questions), financial incentives sometimes hurt and sometimes help performance. One interpretation is that while high stakes give subjects an incentive to focus, they can also lead subjects to perform poorly under pressure. Thus, it is not clear a priori how increasing the stakes in our laboratory experiment will affect overall performance. To our knowledge there exist no studies that examine how the level of incentives impacts performance in competitions against the opposite sex. Thus, one contribution of this paper is to address this gap in the literature.

Our paper also furthers the work of a handful of recent papers that highlight the complementary role of field and laboratory experiments in enhancing our understanding of

Received for publication July 20, 2004. Revision accepted for publication October 29, 2007.

* Department of Economics, University of California at San Diego; Department of Economics, Duke University; and Department of Economics, University of Colorado at Boulder, respectively.

We thank Vince Crawford, Nora Gordon, Terra McKinnish, seminar participants at Northwestern University, the University of California at San Diego, and the University of Colorado at Boulder for their helpful comments. We also thank Patrick Dickinson and Pamela Maine for their research assistance.

[1] That men and women respond differently to competitive settings has a rich history in psychology. Much of the literature focuses on differences in emotional and psychological responses to competition (for example Madden & Kirkby, 1995, and Mazur, Susman, & Edelbrock, 1997), while a smaller number directly examine performance (see Brown, et al., 1997, and Conti, Collins, and Picariello, 2001). Explanations for the disparate responses generally appeal to stereotype-threat, evolutionary psychology, women's fear of success, and higher levels of confidence among men (see Corbin, 1981; Beyer, 1990; Lundberg, Mardberg, and Frankenhaeuser, 1994; Beyer & Bowden, 1997; and Ehrlinger & Dunning, 2003).

individual behavior. These papers emphasize the need to "build a bridge" between the field and the lab (see, for example, Harrison & List, 2004, and Levitt & List, 2007a, 2007b). Indeed, the importance of being able to link field experiments with lab experiments has long been recognized in other disciplines such as the natural sciences. Obviously, one of the challenges in building a bridge between the field and the lab is that field experiments and lab experiments differ along many dimensions. Harrison and List (2004), for example, identify six dimensions along which the lab and the field may differ: the nature of the subject pool, the nature of information that subjects bring to the task, the nature of the commodity, the nature of the task or trading rules applied, the nature of the stakes, and the environment in which the subjects operate.[2] Given these many differences, it is not surprising that field and laboratory experiments do not always yield identical results. List (2006), for example, finds that the behavior of experienced sportscard dealers differs depending on whether they are interacting in a laboratory or field setting.

In this vein, we view our analysis of *The Weakest Link* as a field experiment that we then attempt to replicate in the lab. By linking these two experiments, we hope to both gain a broader insight into the factors that influence how well men and women compete against one another and learn about how well lab experiments in this area can be bridged to the field.

Since there are obviously features of the game show that are nearly impossible to replicate in the lab (for example, the fact that the game show is televised in front of a national audience), our goal is not to examine how every difference between the game show and the lab affect results regarding gender and competition. Rather, we focus on the differences over which we have some control. In particular, we examine how the size of financial incentives and the age of subject pool affect our results. Should the qualitative findings be the same for at least a subset of this state space, we will have evidence that the other differences across the lab and the field have little impact on the qualitative implications of competing against the opposite sex.

The financial incentives on *The Weakest Link* are substantial. Contestants on the weekly version of the show, for example, can win in excess of $125,000. In order to build a bridge between our laboratory experiment and the substantial rewards offered on the game show, we run three different versions of our experiment, in which subjects compete for either $20, $50, or $100. We can then compare not only whether our laboratory results are sensitive to the stakes, but also whether the results from the lab move closer to the results from the game show as the stakes in the lab increase. In addition, in order to examine whether and how the age of the subject pool affects our results, we divide the contestants on the game show into two groups: young (age < 33) and old (age ≥ 33). We can then ask whether the results from our laboratory experiment, where the subjects are college-aged, more closely match the behavior of young or old contestants on the game show.

In our analysis, we find that the results from the game show and from the laboratory experiment differ from one another. In addition, neither the difference between the stakes nor the age of the subject pool alone can account for this discrepancy. However, when controlling for both the stakes and the age of the subject pool, we reach similar conclusions about how men and women respond to competition against members of the opposite sex. Thus, it appears that both stakes and age are important determinants of how men and women respond to competitive pressures. This suggests that laboratory experiments that examine the interaction between gender and competition cannot necessarily be generalized to other real-world settings unless efforts are made to account for differences in the stakes and the age profile of subjects.

The remainder of this paper proceeds as follows. Section II develops a simple theoretical model aimed at building a link between our lab experiments and the game show results. Section III describes the rules of the game show and our game show data. Section IV presents our empirical methodology and our game show results. Sections V and VI describe our experimental design and present our experimental results. Section VII compares our game show results to our experimental results. Section VIII concludes.

II. Theory

In the spirit of Levitt and List (2007a, 2007b), we construct a theoretical model that formalizes the relationship between outcomes in the field and those in the lab. In doing so, we identify the restrictions that need to be placed upon our model in order to link the lab and the field.

While easily extended to other contexts, here we focus on three key factors in determining the effect of a given treatment. The factors we are concerned with are the age of the subjects $x \in X$, the level of stakes $s \in \Re^+$, and the context, $c \in \{L, F\}$, which is meant to capture all other differences between the lab and the field. The treatment we are interested in evaluating is the effect of competing against the opposite sex ($T = 1$) versus competing against an opponent of the same gender ($T = 0$).

In the most general case, the expected outcome of a given "experiment" in either the lab or the field is given by equation (1).

$$E[Y] = g(x, s, c|T), \tag{1}$$

where, $g(\cdot)$ maps age, stakes, context, and treatment into the expected outcome. To account for the fact that, in terms of

[2] One advantage of using data from *The Weakest Link* is that we are able to closely replicate the key features of the competitive environment on *The Weakest Link* in the lab. As a result, we have to worry less about differences in the nature of the commodity and the nature of the task or trading rules.

age, lab subjects are drawn from a restricted subset of the field population, we assume that the support of x for field subjects is equal to X while the support of x in the laboratory is a restricted subset of X, $X_L \subset X$.

Because there is no overlap in the support of the stakes available in our field and lab experiment, additional structure on the role of stakes is required to bridge the two settings. To this end, we formalize the effect of changing the stakes as movement between two regimes. In regime 1, no payments are made, and thus, performance is not affected by financial incentives. In regime 2, the stakes are so high that further increases in the stakes have no additional impact on outcomes. Under both regimes, the treatment effect may vary as a function of individual characteristics and context. Formally, let $\lambda(s)$ be the weight placed on the second regime, mapping from $\mathcal{R}^+ \rightarrow [0, 1]$. This function has the following properties:[3]

1. $\frac{\partial \lambda}{\partial s} > 0, \qquad \forall s \in [0, \underline{s}]$,
2. $\lambda = 1, \qquad \forall s > \underline{s}$.

These properties imply that as the stakes increase regime 2 takes on more importance and, above some threshold, $\underline{s}$, regime 2 becomes the only relevant regime. Under these assumptions, the expected treatment effect for individual i facing stakes s_j in context c is given by equation (2).

$$\begin{aligned}\overline{TE}_i = &[1 - \lambda(s_j)] \\ &\times [g_1(x_i, c|T = 1) - g_1(x_i, c|T = 0)] \\ &+ \lambda(s_j)[g_2(x_i, c|T = 1) - g_2(x_i, c|T = 0)],\end{aligned} \tag{2}$$

where, $g_1(\cdot)$ and $g_2(\cdot)$ represent regimes 1 and 2 respectively. A key implication of the specification in equation (2) is that, even if there is no overlap in the support of stakes between the lab and field, we should see the behavior in the lab and field approach one another as the stakes move closer—holding age constant and assuming context does not affect outcomes.

Given the model specification, we now turn to the link between lab and field. Consider what is required for the average treatment effect, across all subjects, in the lab to replicate the average treatment effect in the field. This replication will occur if the following three conditions are met:

1. $\underline{s}$ is low enough that the stakes offered in the lab cross the threshold where regime 2 holds.
2. There is no effect of x on $g_1(\cdot)$ and $g_2(\cdot)$.
3. There is no effect of c on $g_1(\cdot)$ and $g_2(\cdot)$.

In cases were these assumptions do not hold, quantitative links can still be built. For example, given functional forms for $\lambda(\cdot)$, $g_1(\cdot)$, and $g_2(\cdot)$ and variation in s and x in the lab, it is still possible to calculate the expected treatment effect off of the support of s and x in the lab as long as assumption 3 still holds.

These assumptions can be relaxed if, as is the case with most experiments, we are interested in drawing qualitative inferences to the field using results from the lab. For example, suppose that in the lab the treatment effect for women of competing against a member of the opposite sex is negative. If this treatment effect varies in magnitude (but not sign) in response to stakes, age, and context, then we can still draw qualitative conclusions. The key assumption is then that the sign of the treatment effect does not vary with age, stakes, or context:

$$\begin{aligned}\text{sign}([1 - \lambda(s_j)][g_1(x_i, c|T = 1) - g_1(x_i, c|T = 0)] + \lambda(s_j) \\ \times [g_2(x_i, c|T = 1) - g_2(x_i, c|T = 0)]) < 0, \qquad \forall s, x, c.\end{aligned}$$

We explicitly test for the sign of the treatment effect being the same across the lab and the field. In cases where the sign of the treatment effect varies with the stakes then, given our assumptions about how stakes affect outcomes, there will still exists a level of stakes $\underline{s}^* < \underline{s}$ such that for all experiments with stakes greater than $\underline{s}^*$ results will be qualitatively the same. Of course, if responses also vary by age and context, then it will be necessary to account for these differences as well. As we mentioned above, empirically, we are able to control for differences in age between the lab and the field, but we are able to do little to account for other discrepancies between the two contexts. Thus, our goal is to see how far we can go in bridging the gap between the field and the lab using variation in only the stakes and age.

III. Game Show Data

We use data collected from recordings of the nationally televised game show *The Weakest Link.* There are two versions of the show, an hour-long weekly show and a half-hour-long daily show, with both versions following the same general structure. After excluding celebrity episodes where the contestants play for charity, our data consist of 28 weekly shows and 73 daily shows.[4]

Each show is divided into a series of timed rounds, with the number of rounds corresponding to the number of players: eight rounds in the weekly show and six rounds in the daily show. Within each round, players are sequentially asked to answer general trivia questions where correct answers translate to an increase in the prize money. Regardless of whether a question is answered correctly, the next contestant is asked a new question. As a result, players do not "compete" to answer individual questions. The first correct answer is worth $1,000 in the weekly show and $250

[3] While the role of stakes in switching between regimes may also vary with x and c, for tractability, we assume that $\lambda(\cdot)$ is independent of x and c.

[4] Our original sample included 75 daily shows. Two episodes had to be dropped due to incomplete data resulting from broadcast interruptions.

in the daily show. After a correct answer, a player can choose to "bank" the money for the team. If the player banks, the next correct answer is again worth $1,000 in the weekly show and $250 in the daily show. Should the player decide not to bank, the amount of money added to the pot following a correct answer increases. For example, in the weekly show, a successive chain of eight correct answers, with no intervening banks, leads to a $125,000 increase in the pot. Failure to answer a question correctly, however, leads to the loss of any unbanked money for that round. Money banked from each round is accumulated into a team bank (which ultimately is awarded to the player who wins the entire game).

After each round, each player votes independently as to which player he or she would like to remove from the show. The player who receives the most votes leaves the game. In the event of a tie, the "strongest link" chooses which player to remove from the subset of players who received the most votes. The strongest link is the player who answers the highest percentage of his or her questions correctly.[5] Once the field of players is reduced to two (this occurs in round 7 of the weekly show and round 5 of the daily show), an additional round is played in which prize money is accumulated in the same fashion as earlier rounds. Then, the two players compete against each other in a head-to-head competition in the final round.

In this final round, five (three) questions are asked of each contestant in the weekly (daily) show, with the winner being determined by who answers the most questions correctly. Once it is clear that a winner has been established, no further questions are asked. For example, in the daily show if the first player answers the first three questions correctly and the second player answers their first three questions incorrectly, it is impossible for the second player to catch up and no further questions are asked. In the event of a tie after five (three) questions, each contestant is asked an additional question until one answers correctly and the other incorrectly. The winner then takes all of the accumulated money, and the other contestants leave with nothing.

In the early rounds of the game, there are strong incentives to correctly answer questions since players who answer questions incorrectly prevent the pot of prize money from growing and run the risk of being voted off. However, as the game progresses, the prospect of the head-to-head competition in the final round may give players an incentive to vote off strong players. As a result, in later rounds, players may deliberately answer questions incorrectly. Interestingly, our analysis reveals that in every round of the game, the weakest player is always more likely to be voted off than the strongest player. For example, in round 6 of the weekly show (in which the incentives to vote off the strong players are the highest), the weakest link is still over 14% more likely to be voted off than the strongest link. This suggests that even at the end of the game, there are strong penalties associated with deliberately answering questions incorrectly. Nonetheless, in our analysis, we explicitly consider the possibility that players may intentionally miss questions.

In terms of drawing general conclusions about linking the results from lab experiments on gender and competition to the field, use of the *The Weakest Link* as a field experiment has both strengths and weaknesses. On the plus side, when compared with the competitive labor market settings that motivate much of the experimental work on gender and competition, the simple structure of the game facilitates rather direct replication of the key elements of the game show in the lab. Further, the pressure to perform well in front of the studio audience as well as on national television may help to replicate the stresses associated with the workplace. Finally, contestants on *The Weakest Link* are drawn from a wide range of occupations and ages, resulting in a demographic profile not unlike that of the workforce.

The biggest potential drawback is likely the fact that contestants on *The Weakest Link* may not be representative of the U.S. population. We have investigated how contestants are selected to be on the *The Weakest Link.* Initially, contestants are screened on their ability to correctly answer trivia questions. Additionally, as is the case with most television game shows, the show's producers select individuals who they believe television audiences will enjoy watching. Thus, contestants on *The Weakest Link* tend to be relatively attractive and charismatic. In addition, the producers wish to ensure that contestants are drawn from a variety of geographic regions within the country and that there are equal numbers of men and women on each show. We cannot rule out the possibility that these selection criteria are correlated with the contestants' ability to compete against members of the opposite sex, our variable of interest. However, the ability to compete against members of the opposite sex is not an explicit selection criterion.

Much of our analysis of the game show data focuses on those individuals who survive to the final two rounds. We focus on the final two rounds because there are only two contestants left, there is no possibility of being voted off and there are unambiguous incentives to answer questions correctly. In addition, for those who make it to the final two rounds, we have better measures of ability because they answer more questions than those voted off in earlier rounds. As will be shown in the next section, the basic results do not change when we use broader sample definitions.

Table 1 presents descriptive statistics for individuals who make it to the final two rounds. As can be seen, roughly equal numbers of male and female contestants make it to the final two rounds (almost every show starts out with an equal

[5] Should there be a tie for the "strongest link," the amount of money banked is used as the tiebreaker.

TABLE 1.—DESCRIPTIVE STATISTICS FOR CONTESTANTS WHO MAKE IT TO THE FINAL TWO ROUNDS

Variable	Females	Males
Daily show		
Number of contestants	71	75
Average number of questions answered	16.1	16.3
Overall percentage correct	62.4%	66.8%
Percentage correct in the final two rounds	54.7%	59.9%
Percentage correct in round 1	72.3%	75.6%
Weekly show		
Number of contestants	27	29
Average number of questions answered	29.1	28.9
Overall percentage correct	60.8%	62.3%
Percentage correct in the final two rounds	56.5%	60.7%
Percentage correct in round 1	67.3%	68.4%

TABLE 2.—PROBABILITY OF ANSWERING A QUESTION CORRECTLY BY GENDER COMPOSITION FOR FINAL ROUND CONTESTANTS

Gender	Opponent's Gender	
	Female	Male
Daily show		
Female		
Final two rounds	0.573	0.542
Previous rounds	0.698	0.655
Contestants	36	39
Male		
Final two rounds	0.641	0.569
Previous rounds	0.704	0.714
Contestants	39	32
Weekly show		
Female		
Final two rounds	0.572	0.561
Previous rounds	0.663	0.613
Contestants	12	17
Male		
Final two rounds	0.650	0.560
Previous rounds	0.684	0.560
Contestants	17	10

number of male and female contestants).[6] For both the weekly and the daily shows, men and women who make it to the final two rounds answer close to the same number of questions per game. While men answer a higher percentage of questions correctly than women, this difference is small.

IV. Game Show Results

A key component of our analysis is measuring the underlying ability of the game show's contestants. Since we do not actually observe the contestants' true underlying abilities, we proxy for ability in a number of different ways. First, we use the cumulative percentage correct in all of the previous rounds. That is, suppose contestant i answered N_i questions before the final round. The ability measure used is given by

$$A_i = \frac{\sum_{n=1}^{N_i} c_{in}}{N_i}, \tag{3}$$

where c_{in} equals 1 if contestant i answered the nth question correctly and equals 0 otherwise.

Table 2 shows the probability of answering a question correctly for men and women both in rounds previous to the final two rounds and in the final two rounds. These probabilities are broken out by whether the contestant faced a man or a woman in the final round. The most striking feature of table 2 is the result for men in the daily show. While men in the daily show have essentially the same cumulative percentage correct in previous rounds whether they faced a woman or man in the final two rounds, performance in the final two rounds is very different for those who face a woman and those who face a man. Men who face women give correct answers for over 64% of their final round questions, while those who face men answer less than 57% of their final round questions correctly. Note that if men perform better when they compete against women, then these results are likely to understate the effect of opponent's gender on male performance since men who face women in the final two rounds are likely to have faced women in previous rounds, thereby biasing upward the cumulative percentage correct in previous rounds. The relationship is less clear for the other groups since the cumulative percentage correct differs depending upon the gender of the opponent in the final two rounds. The only clear pattern from the other cells is that groups who perform well in previous rounds also perform well in the final two rounds.

In order to separate out the effects of gender and ability, we estimate logit models of the probability of answering a question correctly in the final two rounds.[7] We assume that the probability of answering the kth question correctly, $c_k = 1$, depends upon individual i's ability, A_i, whether the contestant is a woman, S_i, whether the contestant's opponent is a woman, S_j, and whether the observation comes from the weekly show, W. The probability of a correct answer then follows:

$$P(c_k = 1) = \frac{\exp(\alpha A_i + \beta S_i + \gamma_1(1 - S_i)S_j + \gamma_2 S_i S_j + \delta W)}{1 + \exp(\alpha A_i + \beta S_i + \gamma_1(1 - S_i)S_j + \gamma_2 S_i S_j + \delta W)}. \tag{4}$$

Because our primary goal is to explore the potential to replicate results from *The Weakest Link* in the laboratory, a key concern is the difference in the age profiles between the undergraduate students who compose our laboratory sample and the contestants on the game show. To help us bridge the

[6] In previous studies of data from *The Weakest Link,* Antonovics, Arcidiacono, and Randall (2005) and Levitt (2004) find no discrimination in voting patterns against females while the former find that women discriminate against men in their voting behavior. However, this discrimination should not affect our results—particularly performance in the final round—once we condition on ability.

[7] Recall that in the last two rounds there are only two contestants left, so the sex ratio remains unchanged and there are no incentives to answer incorrectly.

TABLE 3.—LOGIT ESTIMATES OF THE PROBABILITY OF ANSWERING A QUESTION CORRECTLY IN FINAL TWO ROUNDS†

Ability Measure:	Cumulative (1)	Cumulative†† (2)	Round 1 (3)	Round 1 (4)
Female	−0.016	−0.023	−0.040	−0.046
	(0.151)	(0.151)	(0.150)	(0.150)
Male × female opponent	0.270*		0.294*	
	(0.154)		(0.153)	
Female × female opponent	−0.005		0.026	
	(0.155)		(0.155)	
(Age ≥ 33) × male × female opponent		0.412*		0.474**
		(0.212)		(0.210)
(Age ≥ 33) × female × female opponent		0.083		0.118
		(0.211)		(0.210)
(Age < 33) × male × female opponent		0.155		0.147
		(0.193)		(0.193)
(Age < 33) × female × female opponent		−0.078		−0.051
		(0.208)		(0.207)
Age < 33		0.090		0.085
		(0.154)		(0.154)
Percentage correct	1.466**	1.431**	0.409**	0.393**
	(0.367)	(0.367)	(0.188)	(0.189)
Log likelihood	−972.4	−971.9	−978.3	−977.4

†Results include round dummies and round dummies interacted with type of show. Observations are at the question level; 1,472 observations. Columns 2 and 4 include a dummy for age less than 33 separately for men and women. Estimated standard errors are given in parentheses. ††See text for discussion. **Statistically significant at 95% level. *Statistically significant at 90% level.

two samples we also add to the model interactions between S_j, the indicator for whether the contestant's opponent is a woman, and indicators for whether the contestant is less than or greater than the median age on the game show, 33.[8]

Results are presented in table 3. Ability is measured either as the cumulative percentage correct in all previous rounds or as the percentage correct in round 1. Results using the cumulative percentage correct in all rounds except the final two are presented in the first two columns. In column 1, we present results aggregated across both age categories. We find that a man has a much higher probability of answering a question correctly when he faces a woman than when he faces a man. Using the results from column 1, at the mean male ability level, a man's probability of answering a question correctly in the final round increases from 48.7% when facing a man to 55.5% when facing a woman. No such gender effect exists for female contestants. Because one might be concerned that the incentive to strategically answer questions incorrectly increases in later rounds, columns 3 and 4 report results using only the percentage correct in the first round as our ability measure. Focusing again on the results aggregated across age groups (column 3), while increased noise in the ability measure leads to a lower coefficient on ability, the effect of opponent's gender remains unchanged.

Finally, to facilitate comparisons to laboratory experiments that use undergraduate subjects, columns 2 and 4 interact the indicator for whether the contestant's opponent is a woman with indicators for the contestant's age. The results show that the positive effect that men receive when competing against women is driven by men aged 33 and above. For younger male contestants, this effect is much smaller in magnitude and not statistically significant. All other results remain unchanged.

[8] Ideally, we would focus just on college-age contestants. Unfortunately, small sample sizes preclude such analysis.

In order to test whether our results are robust to both the sample of questions asked and the composition of individuals, we extend our analysis to include all rounds of the game. Since each contestant faces multiple opponents in earlier rounds, the relevant variable for the effects of gender on competition is the sex ratio, defined as the fraction of the contestant's opponents who are women. We use the logit specification above and include separate round dummies for both the weekly and the daily shows as well as interactions between the round dummies and the sex of the contestant in order to account for the possibility that the difficulty of the questions may vary by show and by round.[9]

As a proxy for ability, we use the cumulative percentage correct from past and future rounds of the show. That is, we use information on all questions answered by the contestant apart from those answered in the current round. Given that an individual answers N_i questions, the ability measure used in modeling the probability of correctly answering a question in the rth round is given by

$$A_{ir} = \frac{\sum_{n=1}^{N_i} c_{in} - \sum_{n=1}^{N_{ir}} c_{ir}}{N_i - N_{ir}}, \tag{5}$$

where N_{ir} is the number of questions answered in round r.

Results are presented in table 4.[10] In the first column, we use the cumulative percentage correct to proxy for ability and restrict the sample to those who made it to the final

[9] Separately estimating these specifications for the weekly show and daily show leaves the qualitative results unchanged but decreases the precision of the estimates.

[10] The logit estimates in table 4 require the parametric assumption that the regressors are orthogonal to all error terms—contemporaneous, future, and past. Given the inclusion of round, show, and gender fixed effects as well as interactions of these fixed effects, this assumption will hold as long as the included ability measure effectively controls for all remaining unobserved heterogeneity across individuals in the probability of correctly answering a given question.

TABLE 4.—PROBABILITY OF ANSWERING A QUESTION CORRECTLY IN ALL ROUNDS†

Questions Analyzed: Ability Measure: Contestant Sample:	All Rounds Cumulative Final Round (1)	All Rounds Cumulative Final Round (2)	Round > 1 Round 1 Round > 1 (3)	Round > 1 Round 1 Round > 1 (4)	Rounds > 2 Rounds 1 and 2 Rounds > 2 (5)	Rounds > 2 Rounds 1 and 2 Rounds > 2 (6)
Percentage correct	1.690**	1.643**	0.462**	0.456**	0.628**	0.620**
	(0.248)	(0.250)	(0.094)	(0.094)	(0.146)	(0.147)
Female	0.024	0.027	−0.028	−0.025	−0.069	−0.067
	(0.116)	(0.116)	(0.102)	(0.102)	(0.108)	(0.108)
Male × sex ratio	0.230*		0.308**		0.342**	
	(0.134)		(0.122)		(0.125)	
Female × sex ratio	0.060		0.103		0.237*	
	(0.139)		(0.127)		(0.130)	
(Age ≥ 33) × male × sex ratio		0.332*		0.412**		0.501**
		(0.177)		(0.158)		(0.165)
(Age ≥ 33) × female × sex ratio		0.106		0.205		0.209
		(0.187)		(0.167)		(0.171)
(Age < 33) × male × sex ratio		0.154		0.223		0.161
		(0.162)		(0.147)		(0.152)
(Age < 33) × female × sex ratio		0.008		−0.001		0.028
		(0.183)		(0.169)		(0.173)
Age < 33		−0.021		0.071		0.063
		(0.117)		(0.102)		(0.108)
Observations	3,987	3,987	5,834	5,834	4,439	4,439

†Results include round dummies by sex and round dummies interacted with type of show by sex. Observations are at the question level. Sex ratio is defined as the fraction of the contestant's opponents that are women. Estimated standard errors are given in parentheses.
**Statistically significant at 95% level. *Statistically significant at 90% level.

round, aggregating across age groups. The effect of an opponent's gender on performance in all rounds is similar to the effect identified when examining only the final rounds: men perform better in the presence of women, with women unaffected by the gender composition. The third column includes results for all individuals who make it to at least round 2 using performance in round 1 as the ability proxy. Again, aggregating across ages, we see strong positive effects when men face women. Finally, this result also holds in column 5 where we use percentage correct in rounds 1 and 2 as our ability measure and focus on those who make it to at least round 3.

Once again, in order to bridge to the laboratory results presented below, columns 2, 4, and 6 report results when the sex ratio variable is interacted with indicators for the contestant's age. As with the final round results analyzed in table 3, the gender competition effect for men is largely associated with older contestants and becomes much smaller and statistically insignificant when we focus on contestants under the age of 33.

V. Weakest Link Laboratory Experiment

The results from *The Weakest Link* indicate that men perform better when they compete against women than when they compete against men, whereas the performance of women is unaffected by the gender of their opponent. Further, once we control for the age of the subject, we find that this positive effect from competing against women is driven by older men and that for men under the age of 33 there is no significant effect from competing against women. Because all of the game show results are associated with high stakes, we must rely on the lab to explore the role of stakes in this setting.

In order to provide the laboratory counterpart to our field results, we conducted two sets of experiments, one in May 2005 and the other in April 2007. Both sets of experiments were conducted at the University of California, San Diego (UCSD). Students were recruited through flyers either posted on campus or distributed to large undergraduate classes in political science and economics. The flyers gave students an email address to contact if they were interested in participating. Students who sent an email were scheduled into an available time slot for the experiment. Overall, we held 26 experimental sessions (thirteen in 2005 and thirteen in 2007) with a total of 202 participants.

The student body at UCSD is ethnically diverse. For example, in 2004, only 67% of freshman at UCSD reported English as their native language. This is a concern since the questions asked on *The Weakest Link* are culturally specific to the United States, and individuals who have had less exposure to U.S. culture are less likely to answer questions correctly. In the 2005 experiment, we control for this by asking students to tell us whether they are native English speakers. In the 2007 experiment, we required that subjects be native English speakers to participate.

Each experimental session was divided into two rounds and involved four men and four women.[11] In the first round, all eight participants were seated together in a room and told that they would be asked to answer twenty questions. Each

[11] Due to scheduling difficulties, a small number of sessions involved a slightly different gender balance or a slightly different number of participants.

participant was given an answer booklet in which to record his or her answers. Each question was read out loud and repeated once. Participants had ten seconds in which to write down their answers to each question and were not permitted to go back and change the answer to previous questions. There were no penalties for incorrect answers or spelling mistakes. Participants earned $1 for each correct answer but were not told how many questions they correctly answered until the entire experiment was over. The goal of this initial round was to develop a baseline ability measure for each subject.

In the second round, the eight participants were randomly divided into four pairs, two mixed-gender pairs and two same-sex pairs. Each pair was escorted to a separate room. Here participants were told that they would be competing against one another, and that they would take turns answering questions. One of the participants was randomly selected to go first. On each subject's turn, a question was read out loud and the participant was given ten seconds to give a verbal response. After a response was given, the subject was told whether his or her answer was correct, and their response was recorded. Then, at the start of the other subject's turn, a new question was asked. The experiment continued in this alternating fashion until each player was asked ten questions. In the 2005 experiments, the individual who correctly answered the most questions was rewarded $20, and in 2007 the individual who correctly answered the most questions was awarded either $50, or $100. The amount of the award in 2007 was randomly assigned. In addition, as in 2005, subjects were only made aware of the stakes immediately prior to the start of the head-to-head competition.

As is the case on *The Weakest Link,* in the event of a tie, each participant was asked an additional question. If a participant answered his or her additional question correctly, but their opponent did not, then the participant who gave the correct answer was declared the winner. On the other hand, if both gave correct answers or if both gave incorrect answers, then they were asked additional questions until the tie was broken.

Finally, if at some point in the competition one participant was so far behind that it was no longer possible to catch up, the competition would come to an end. For example, if one participant answered the first six questions correctly and the other missed the first five questions, then it would be impossible for the losing player to catch up. At this point, a winner was announced. Subjects were fully informed of this rule in advance. *The Weakest Link* also follows this convention.

Each experimental session lasted less than 45 minutes. Subjects were given $5 for showing up, and an additional $3 if they arrived more than five minutes before the start of the experiment. On average, in the piece-rate portion of the experiment, subjects answered approximately seven questions correctly.[12] Thus, the total expected payment for the piece-rate and head-to-head rounds combined was $17.07, $32.07, and $57.07, depending on whether the stakes in the head-to-head competition were $20, $50, or $100.

To facilitate the comparison between our game show and laboratory data, the questions we asked were randomly selected from those used on the *The Weakest Link.*[13] In order to ensure that the questions used in the piece-rate section of the experiment did not systematically differ from those used in the head-to-head competition and to allow for the use of question-level fixed effects, we used the same set of questions in adjacent experimental sessions, once in the piece-rate stage and once in the experimental stage. In addition, the questions and their order were identical in the 2005 and 2007 experiments. No subject participated in both the 2005 and 2007 experiment, and there is no indication that subjects in 2007 had any information about the questions asked in 2005.

VI. Experiment Results

We first focus on how well the subjects performed in both the piece-rate and in the head-to-head competition. Since the head-to-head competition ends as soon as it is no longer possible for the losing player to catch up, many games end before twenty questions have been asked (ten questions for both player 1 and player 2). Empirically, the shortest game we observe lasts for thirteen questions. To avoid sample selection issues, we thus focus our attention on the first thirteen questions asked in both the piece-rate and head-to-head competition.

Table 5 presents the probability of answering a question correctly in the piece-rate and in the head-to-head competition broken out by the gender of the subject's opponent and the stakes. Subjects do not have an opponent in the piece-rate portion of the experiment, but we can still categorize subjects according to whether they ultimately face a man or a woman in the head-to-head competition and ask how they perform in the piece-rate section.[14] Because we are concerned about the impact of being a nonnative English speaker on ability to answer trivia question from *The Weakest Link,* the analysis focuses on the performance of native English speakers.[15] There were not enough observations to

[12] In 2005, the average number answered correctly was 6.97, and in 2007 it was 7.17.

[13] Questions were taken from Lewis, Ballheimer, and Larter (2001), *Weakest Link Quiz Book.* A small number of questions were discarded because they had become out-of-date.

[14] Since subjects are randomly matched with their opponent in the head-to-head competition and since they do not know with whom they will be matched until the piece-rate portion of the experiment is over, on average, in the piece-rate section, there should be no performance gap between those who will eventually compete against men versus those who will eventually compete against women. Empirically, however, small differences can (and do) arise.

[15] Note that native English speakers playing against nonnative English speakers are included. Removing those individuals has no impact on the qualitative results but does result in a loss of precision.

TABLE 5.—EXPERIMENT PERCENTAGE CORRECT BY PIECE RATE AND HEAD TO HEAD†

		Opponent	
	Overall	Female	Male
Female			
Piece rate			
Full sample	0.354	0.363	0.345
	1,079	520	559
Stakes = $20	0.378	0.365	0.389
	429	208	221
Stakes = $50+	0.338	0.362	0.317
	650	312	338
Head to head			
Full sample	0.290	0.291	0.289
	538	261	277
Stakes = $20	0.296	0.314	0.278
	167	78	89
Stakes = $50+	0.286	0.276	0.300
	325	156	169
Male			
Piece rate			
Full sample	0.388	0.366	0.409
	1,170	598	572
Stakes = $20	0.397	0.400	0.393
	494	260	234
Stakes = $50+	0.382	0.343	0.420
	676	338	338
Head to head			
Full sample	0.349	0.311	0.389
	587	299	288
Stakes = $20	0.345	0.315	0.378
	249	130	119
Stakes = $50+	0.352	0.308	0.396
	338	169	169

†Total number of observed responses are given underneath each percentage.

see any meaningful differences between stakes of $50 and $100. There is, however, a clear difference in the reaction to the stakes occurring between $20 and $50. Therefore, we present results for stakes of $20 and for stakes of $50 or more.

The first two sets of rows in table 5 report data for women. Stakes clearly affect how women respond to the gender of their opponent. When the stakes are $20, women in the head-to-head competition perform better when they face women than when they face men. Note that this result is not explained by ability differences. In fact, the sample of women chosen to face women in the head-to-head competition fare worse in the piece-rate portion of the experiment than do those chosen to face men. This result is consistent with Gneezy et al. (2003), who find that women perform worse when they compete against men than when they compete against women. Raising the stakes, however, leads to an entirely different result; when the stakes are $50 or more, women perform better in the head-to-head competition when they face men than when they face women, despite the fact that women assigned to face men in the head-to-head competition perform worse in the piece-rate section of the experiment than do those assigned to face women.[16] Thus, stakes appear to influence how women respond to the gender of their opponent.

The second two sets of rows show the results for men. A similar opposite-sex effect is found for men when the stakes are $20; in the head-to-head competition, men who face women perform significantly worse than those who face men even though there is no performance gap in the piece-rate portion of the experiment between those who are chosen to face women and those who are chosen to face men. The results, however, are less clear when the stakes are $50 or more since the men who are chosen to face men in the head-to-head competition perform better in both the piece-rate and the head-to-head competition than those who are chosen to face women.

With the descriptive statistics pointing to poor performance when facing members of the opposite sex in low-stakes environments and ambiguous effects for high-stakes environments, we estimate a logit model of the probability of answering a question correctly controlling for both individual fixed effects and question-specific fixed effects. As before, we avoid the selection issues in the head-to-head questions by focusing only on the first thirteen questions. However, for piece-rate questions we use the full sample so that we can more precisely identify the individual fixed effects. These results are presented in table 6.[17]

The first column of table 6 presents the estimates from a logit model of the probability of answering a question correctly that includes individual and question-specific fixed effects and that controls for whether the question was asked in the piece-rate or head-to-head portion of the experiment. As column 1 shows, women's performance is negatively affected by moving from the piece-rate to the head-to-head competition, while men appear to perform equally well in both. In the second column, we include interactions for whether the subject's opponent was of the opposite sex. Here we see that although women are less likely to answer questions correctly in the head-to-head competition than in the piece-rate portion of the experiment, their performance is not significantly affected by the gender of their opponent. However, for men, their performance falls when they face women in the head-to-head competition relative to when they face men.

The third column presents our preferred specification and points to the importance of stakes in estimating the effects of gender on competition. Consistent with the descriptive statistics, we find that, when the stakes are low, women perform significantly worse when they face men than when

[16] The finding that women perform better when they compete against men in the high-stakes environment disappears once additional controls are added.

[17] The lab experiment can also be used to further test whether a possible explanation for the game show results is that women perform worse on more difficult questions relative to men. To measure question difficulty, we calculated by question the mean probability of a correct answer. We then regressed question answers on female, question difficulty, and question difficulty interacted with female. The interaction term was small and insignificant.

TABLE 6.—LOGIT ESTIMATES OF THE PROBABILITY OF ANSWERING A QUESTION CORRECTLY IN THE LAB†

	(1)	(2)	(3)
Female HTH	−0.467**	−0.415*	
	(0.154)	(0.219)	
Female × male opponent HTH		−0.102	
		(0.301)	
Female HTH × (stakes = $20)			−0.120
			(0.340)
Female HTH × (stakes = $50+)			−0.614**
			(0.283)
Female × male opponent HTH × (stakes = $20)			−0.964**
			(0.476)
Female × male opponent HTH × (stakes = $50+)			0.474
			(0.387)
Male HTH	−0.196	0.050	
	(0.145)	(0.202)	
Male × female opponent HTH		−0.492*	
		(0.284)	
Male HTH × (stakes = $20)			0.145
			(0.308)
Male HTH × (stakes = $50+)			−0.016
			(0.260)
Male × female opponent HTH × (stakes = $20)			−0.838**
			(0.428)
Male × female opponent HTH × (stakes = $50+)			−0.210
			(0.379)
Log likelihood	−1,691	−1,689	−1,685

†Observations are at the question level; 4,324 observations. All results include individual and question fixed effects. Estimated standard errors are given in parentheses.
**Statistically significant at 95% level. *Statistically significant at 90% level.

they face women. When the stakes are $50 or higher, however, women's performance in the head-to-head competition appears to be unaffected by the gender of their opponent. Indeed, even though the results are insignificant, the sign of the point estimate suggests that women may even perform better against men when the stakes are high.[18] Male performance also responds to the level of the stakes. Men in low-stakes environments perform worse in the head-to-head competition when they face women than when they face men. When the stakes are high, however, this effect is small and insignificant.

VII. Bridging the Lab and Field Results

The goal of this paper is to see whether we can build a bridge between our game show results and our results in the lab. We do this in order to determine whether and how lab studies of gender interactions can be generalized to other settings. The first step in building this bridge is to control for the difference in the age profiles of the two subject pools. Once we control for age in the game show data, we find that the positive effect that men experience when competing against women is driven by men who are older than 33, and that for men under 33 this effect is largely absent. Next, we turn to the lab results and control for stakes. We find that both men and women experience a significant negative impact from playing against the opposite sex at the lower stakes ($20) that are typical or slightly above the stakes that have been used in recent experiments in this area. Once we raise the stakes to $50 and above, this negative opposite-sex effect disappears and there is no significant effect of opponent's gender on either men or women.

Thus, we find that when we compare our higher-stakes lab experiments to outcomes for the younger contestants on the game show, our results are consistent (there is no effect for either men or women of competing against a member of the opposite sex). The fact that we are able to build a bridge between the lab and the game show by accounting only for differences across the two environments in stakes and age lends credibility to many papers that use game show data. Namely, it suggests that the other contextual differences between game shows and the lab do not change the qualitative impact of the treatment—at least not in the case of gender interactions. However, the fact that the lab and the field only overlap once differences between the stakes and age have been accounted for suggests that generalizations from the lab to the labor market may be limited unless efforts are made to address these discrepancies.

VIII. Conclusion

There is a growing interest in understanding how the link between gender and performance in competitive environments may impact labor market outcomes. In response to this interest several recent papers have used laboratory experiments in an attempt to identify how the performance of men and women is affected by competing against members of the opposite sex. This research is part of a larger, and

[18] We also find that women's performance in the head-to-head competition falls as the stakes increase, suggesting that women perform worse when the pressure is higher. Though not statistically significant, this pattern also appears to hold for men.

growing, focus on the use of lab experiments to understand the behavior of individuals in economic settings. The increased prevalence of lab experiments in economic research has led to a nascent literature on the ability to draw inference regarding behavior in real-world or field settings from experiments performed in the lab.

This paper contributes to the literature on bridging lab and field outcomes by evaluating the potential to replicate behavior in the high-stakes field setting of the television game show *The Weakest Link* in a typical laboratory setting. Our study yields two primary findings. First, in our laboratory experiment, our results regarding gender and competition differ depending on the stakes, suggesting that the results from previous experimental studies in this area may be sensitive to the financial incentives being offered. Second, we find that we are able to replicate field behavior in the lab, but only under specific conditions. When using lower stakes ($20) that are typical of existing research in this area, we find that our field and lab results conflict. At higher stakes ($50 and above) we are able to bridge our lab and field results. However, obtaining these concordant results requires limiting our field sample to younger subjects whose age is closer to that of the undergraduate students who comprise our laboratory sample. While these results are encouraging, the fact that the lab and the field only overlap for this particular case suggests that care must be taken when making generalizations about gender and competition from the lab to the labor market.

REFERENCES

Antonovics, Kate, Peter Arcidiacono, and Randall Walsh, "Games and Discrimination: Lessons from The Weakest Link," *Journal of Human Resources* 40:4 (Fall 2005), 918–947.

Beyer, Sylvia, "Gender Differences in the Accuracy of Self-Evaluations of Performance," *Journal of Personality and Social Psychology* 59:5 (November 1990), 960–970.

Beyer, Sylvia, and Edward Bowden, "Gender Differences in Self-Perceptions: Convergent Evidence from Three Measures of Accuracy and Bias," *Personality and Social Psychology Bulletin* 23:2 (February 1997), 157–172.

Brown, N. L., S. L. Brown, R. M. Brown, L. R. Hall, and R. Holtzer, "Gender and Video Game Performance," *Sex Roles* 36:11–12 (June 1997), 793–812.

Camerer, Colin, and Robin Hogarth, "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk and Uncertainty* 19:1–3 (1999), 7–42.

Conti, Regina, Mary Ann Collins, and Martha Picariello, "The Impact of Competition on Intrinsic Motivation and Creativity: Considering Gender, Gender Segregation and Gender Role Orientation," *Personality and Individual Differences* 31:8 (December 2001), 1273–1289.

Corbin, Charles, "Sex of Subject, Sex of Opponent, and Opponent Ability as Factors Affecting Self-confidence in a Competitive Situation," *Journal of Sport Psychology* 3:4 (1981), 265–270.

Ehrlinger, Joyce, and David Dunning, "How Chronic Self-views Influence (and Potentially Mislead) Estimates of Performance," *Journal of Personality and Social Psychology* 84:1 (January 2003), 5–17.

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

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

Harrison, Glenn, and John List, "Field Experiments," *Journal of Economic Literature* 42:4 (December 2004), 1009–1055.

Levitt, Steven, "Testing Theories of Discrimination: Evidence from the Weakest Link," *Journal of Law and Economics* 41:2 (October 2004), 431–452.

Levitt, Steven, and John List, "What Do Laboratory Experiments Measuring Social Preferences Tell Us About the Real World?" *Journal of Economic Perspectives* 21:2 (Spring 2007a), 153–174.

——— "Viewpoint: On the Generalizability of Lab Behaviour to the Field," *Canadian Journal of Economics,* 40:2 (May 2007b), 347–370.

Lewis, Garry, David Ballheimer, and Sarah Larter, *Weakest Link Quiz Book* (London: Carlton Publishing Group, 2001).

List, John, "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions," *Journal of Political Economy* 114:1 (February 2006), 1–37.

Lundberg, Ulf, Bertil Mardberg, and Marianne Frankenhaeuser, "The Total Workload of Male and Female White Collar Workers as Related to Age, Occupational Level, and Number of Children," *Scandinavian Journal of Psychology* 35:4 (December 1994), 315–327.

Madden, Chris, and Robert Kirkby, "Gender Differences in Competitive Stress," *Perceptual and Motor Skills* 80 (June 1995), 848–850.

Mazur, Allan, Elizabeth Susman, and Sandy Edelbrock, "Sex Difference in Testosterone Response to a Video Game Contest," *Evolution and Human Behavior* 18:5 (September 1997), 317–326.

Niederle, Muriel, and Lise Vesterlund, "Do Women Shy Away from Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics* 122:3 (August 2007), 1067–1102.