Evaluation of Early-Stage Ventures: Bias across Different Evaluation Regimes

Daniel C. Fehder

University of Southern California

Fiona Murray
MIT Sloan and NBER

Abstract: This paper explores the selection mechanisms inside a startup accelerator program, measuring how variation in institutional arrangements impacts how judges evaluate businesses opportunities. Specifically, we focus on the role of gender and status cues in the evaluation of high uncertainty projects by explicitly comparing the evaluation of a fixed set of projects across two "evaluation regimes," using detailed data from a high-quality entrepreneurship program. In the "individual paper regime," early-stage startups are evaluated on an individual basis by judges using a written application to a startup program. In the "committee interview regime", evaluators are grouped together in a committee where they are able to observe a short pitch by the co-founders of the venture, can ask questions and can confer with one another. We find strong differences across our two evaluation regimes. Judges in the committee regime are more likely to use the characteristics of the project team (as well as those of the project) in making their score determination, relative to the paper-based regime. We also find evidence that the decision-making process of judges is different in the committee regime: judges seem to achieve greater converge in their score determination in committees, yet disagree substantially in the paper-based regime. Finally, we find that the saliency of co-founder characteristics is higher in the in-person committee regime, and unlike prior studies, we find that that female-founded firms are judged to be higher quality by both male and female evaluators, but only under the committee regime. Our results point to the central role of the type of evaluation regime on the types of ideas selected and the use of different types of information by evaluators.

Every year venture capitalist, angel investors and the leaders of programs such startup accelerators and business plan competitions must evaluate thousands of early-stage ideas and early-stage teams. Many of these are high quality ideas created by collections of intelligent and motivated individuals, and yet only a few will wind-up becoming enterprises with a significant economic impact and the levels of financial rewards that allow investors and programs to continue their activities. As such, the act of selecting high quality early-stage projects is both critically important and yet much like panning for gold or at least choosing a diamond in the rough.

More than just being faced with enormous uncertainty, evaluators of early-stage firms must conduct their search under enormous time pressure. The introductory quote evokes how quickly early-stage evaluators must assess (and dismiss) the viability of candidate projects. As such, they develop screening mechanisms which quickly down-select startups to a much smaller pool for whom they will provide significant attention. Indeed, a growing literature has pointed to time as the constraint driving a substantial share of early-stage investor behavior and even the performance of their portfolio firms (Bernstein, Giroud, and Townsend 2015).

Within such resource-constrained contexts, research has begun to uncover the attributes of early-stage projects that investors attend as they make investment decisions (Bernstein, Korteweg, and Laws 2015; Scott, Shu, and Lubynsky 2015). Some of the key findings of this emerging literature support our second introductory quote above: the quality of the founding team of individuals (as well as their ideas) provide key inputs into evaluators' assessments of early-stage ventures. And indeed it is this focus on attributes of ideas and founding teams that has prompted some to explore the tradeoff between the "horse" and the "jockey" in the venture capital literature (Kaplan, Sensoy, and Strömberg 2009).

If individual characteristics are among the key inputs into the evaluation of early-stage ventures, then there is the potential for biases based on status characteristics (e.g. gender or elite educational background) to arise in the evaluation of early-stage firms. Numerous studies have shown that status characteristics such as gender are used across numerous settings as a proxy for the expected performance of a candidate (Berger et al. 1977; Correll and Ridgeway 2003). Recent work has explored the impact of gender in the evaluation of projects in innovation and entrepreneurship settings where *ex ante* uncertainty around project quality is highest (and thus potential demand for performance proxies is also high) (Brooks et al. 2014; Botelho and Abraham 2016). The results of such studies illustrate the strength of gender in the evaluation of ideas (even holding the idea constant) as well as other observables such as attractiveness. However, while rarely the focus of attention, these results are likely to be strongly shaped by the institutional details of the decision-making processes that are at work – in other words by the evaluation regime that is selected.

In settings such as pitch-competitions where evaluators are required to make fast judgments about the quality of a project and a team, the evaluation regime is often one of rapid *individual* decision making, where less than two to three minutes of information input is followed by instantaneous evaluation. Under such regimes, the attributes of the founders – in particular gender - have a large impact on the decision processes of the judges (Brooks et al. 2014). As evaluators have both more time and more access to objective measures of an idea or candidate's quality, the biasing effects of gender seem to become diminished substantially (Botelho and Abraham 2016) drawing attention to the varying role of status characteristics across differing types of evaluation regime.

The question of how personal characteristics impact evaluation of projects at a stage of high uncertainty is a larger concern in the strategy and innovation literature. A substantial literature in the economic sociology literature has connected characteristics including gender and educational background to the behavior of evaluators in a range of economic situations (Azoulay, Stuart, and Wang 2013; Stuart, Hoang, and Hybels 1999). Recent literature has used the natural or experimental variation in the salience of personal characteristics of idea or project originators to compare the impact of gender and other status biases across different evaluation regimes (Simcoe and Waguespack 2011; Bernstein, Korteweg, and Laws 2015; Botelho and Abraham 2016). These studies show a dramatic reduction in the role of gender or other status cues in the evaluation ideas when evaluators are given richer or more objective information through which to evaluate the quality of an idea or project. At the same time, a growing body of research in the innovation literature is also exploring the performance of consensus-making committees (as opposed to other evaluation regimes) in the evaluation of new ideas whether through the exploration of committee composition (Li and Agha 2015) or in comparison to other forms of project selection (Mollick and Nanda 2015). While a rich literature has developed documenting the role of status characteristics in shaping intergroup behavior (Strodtbeck, James, and Hawkins 1957; Ridgeway 2001), this literature has surprisingly little to say about the role of these forces in committee-based project selection. This ommission is particularly surprising considering the fact that committees play a large role in the selection of innovative ideas and projects across a broad slice of our economy and society including academia, corporations, and innovation ecosystems.

We seek to address this gap in our understanding of the role of gender and status cues in the evaluation of high uncertainty projects by explicitly comparing the evaluation of a fixed set of projects across two "evaluation regimes" using detailed data from a high-quality entrepreneurship program. In the first regime (that we refer to as the "individual paper regime", early-stage startups are evaluated on an individual basis by judges using a written application to a startup program. In the second (which we refer to as the "committee interview regime"), evaluators are grouped together in a committee where they are able to observe a short pitch by the co-founders of the venture, can ask questions and can confer with one another. In addition to use of two evaluation schemes, this paper uses two additional features of the entrepreneurship program that make it particular interesting from an econometric perspective. First, we take advantage of the fact that a large number of early-stage ventures are subjected to these two approaches to screening (within one month) using the same scoring sheet, providing us with variation in the evaluation context of an idea while holding that idea fixed. Second, we are able to exploit random variation in the committee membership to explore the impact of the demographic make-up of committee members on the evaluation outcomes.

Overall, we demonstrate strong differences in the way in which projects are evaluated across our two evaluation regimes. Judges in the committee regime are more likely to use the characteristics of the project team (as well as those of the project) in making their score determination than they are in the paper-based regime. We also find evidence that the decision-making process of judges is different in the committee regime: judges seem to converge quickly in their score determination in committees yet disagree substantially in paper-based regime. These two facts together suggest that the status characteristics of startup teams might represent an easy source of agreement amongst committee members.

Interestingly, the committee's focus on startup team characteristics seems to improve performance. Judges in the committee regimes are more likely to predict the future performance of startup teams than when they judge individually.

Our results would be closely in accord with a highly rationale view of group decision making processes if the only traits that the judges attended to were markers of human capital and achievement (like graduate degrees or STEM degrees) which might reasonably be signals of the underlying capacity and quality of the individual. What is more surprising is that we see similar attention in the committee round to the ascriptive characteristics of the teams: specifically gender. In regards to the judge's use of team gender composition, we find results that are in accord with some of the predictions of theories of gender bias, but are somewhat counter-intuitive and undermine the pat interpretations of judging in committees as an improvement over individual decision processes. First, we find that judges will use information about a project team’s gender composition more in the committee evaluation regime than the individual regime. We measure this at the level of average judge score, individual judge scoring, and difference between scoring regimes showing that the impact of gender on the evaluation of early-stage projects is greater in our committee setting across all of these levels.

Our results could come from two sources: first, the saliency of co-founder characteristics is higher in the in-person committee round, making them more useful signals to judges in the committee round. This interpretation is in accord with previous findings on the importance of saliency on gender bias. Surprisingly, however, the impact of gender on the evaluation of these projects goes in the opposite direction than theory would predict - being positive and statistically significant, suggesting that female-founded firms are judged to be higher quality, on average, by the committee members selecting firms to participate in the program. Second, we find no evidence of differences between male or female evaluators in their average score given to female entrepreneurs. Our measured effect of the differential impact of gender on committee-based evaluation seems to be driven by the decisions of both male and female evaluators.

4.2. THEORETICAL FRAMEWORK AND LITERATURE REVIEW

Substantial attention is increasingly being paid by scholars of organizations and innovation to the institutions that shape the evaluation of early stage ideas. Recognizing that evaluation is a two-sided process, research has explored both the evaluators and the evaluated projects.

Recent scholarship has focused on the impact of evaluator background on the outcomes of the evaluation process (Boudreau et al. 2016; Li and Agha 2015), providing evidence that the composition of a committee matters for the eventual evaluation of a new idea or venture. In addition, recent research has attempted to compare evaluation taking place on very different platforms and with quite distinctive regimes. For example, (Mollick and Nanda 2015), showing substantial similarity between evaluation outcomes for both expert-based and crowd-based judgments.

This paper hopes to bridge these two literatures by asking not only how compositional differences across different committees impacts the eventual judgment of that committee, but also how different evaluation regimes change the judgment processes of individual judges. In addition, we provide the first evaluation of gender bias in a committee setting. While there is a growing awareness of differences in the magnitude of gender bias across different evaluation settings, we know of no other study that examines how bias is moderated by an individual's presence in a committee. This is despite the fact that committees are an important setting for evaluations where gender bias has been implicated as diverse as venture capital investment and academic tenure decisions.

4.2.1 Theories of Bias in Individual Evaluation

In many of the settings in which startups are evaluated, these evaluations are made by individuals without reference to or input from others. There are many ways in which investors make contact with and evaluate early stage entrepreneurs including referrals (Shane and Stuart 2002; Hallen 2008) and business plan submission (Kirsch, Goldfarb, and Gera 2009; Honig and Karlsson 2004). In each of these channels, Early-stage ventures with growth expectations must convince an individual of the value for their proposed product or service and the wisdom of their approach. Thus, it is critical to have a clear model of the mechanisms that shift the evaluation of early stage ventures by individuals.

How do investors and entrepreneurship professionals make investment decisions? Recent work has attempted to elucidate the features of entrepreneurial ventures, which are most important to evaluators. In a paper using experimental variation in startup profiles, studies have shown that information on co-founder background is most salient (Bernstein, Korteweg, and Laws 2015) and that attributes such as gender and attractiveness have a strong impact on the evaluation idea quality (Brooks et al. 2014). These data are broadly consistent with a view of social identity theories that have been developed in both the sociology (Tajfel 1982; Correll and Ridgeway 2003) and economics literature (Akerlof and Kranton 2000).

A broad range of research in social psychology has demonstrated that male actors are believed to be more capable in accomplishing tasks than female actors (Fiske et al. 2002). Drawing from these empirical regularities in the experimental data and observational data, sociologists and social psychologists have developed a set of theoretical predictions about how beliefs about status characteristics inform the choices and behaviors of actors in a range of situations (Berger et al. 1977; Correll and Ridgeway 2003). Because specific information about individuals is often non-existent or difficult to obtain, evaluators use general societal beliefs about differences in the capabilities and capacities of representative individuals from different genders (as well as race, class, etc.) to inform decisions - instead of information about the individual at hand. There is a remarkable similarity between theoretical mechanisms believed to be driving the data in the status beliefs literature and the economic investigation of statistical discrimination (Phelps 1972). In both theories, the costless-ness of using socially relevant information is contrasted with the difficulties and cost of observing information about specific individuals.

While information-based theories might make sense in many settings, such as hiring, where *ex-ante* quality in a role is very difficult to observe, there are other theories of gender bias which hold in conditions where objective information is available about the performance capabilities of the person or, importantly, of the project being evaluated. In the case of startup evaluation, there are aspects of new venture that are observable to the judges and directly inform the potential performance of the venture: for instance the quality of the thought informing the choice of market and technology. Even in cases where objective performance evaluation is available, it is possible that men and women are judged using different yard-sticks. Through field and experimental evidence, scholars have demonstrated more stringent evaluation criteria being applied to women relative to their male counterparts, a phenomenon called "double standards" (Foschi, Lai, and Sigerson 1994; Foschi 1996).

> ***Hypothesis 1:*** *Female Co-Founders will have a negative impact on the level of evaluation of early-stage firms*

The saliency of gender is one of the key boundary conditions for generating gender bias in both status characteristics and double standards theory. Saliency can be interpreted as the magnitude of gender's weight in the decision processes of evaluators, and it can be moderated by demand-side, supply-side, task-specificity and contextual factors. Demand-side factors include aspects of the decision-making environment that increase an evaluator's need or desire to use gender as an input into decision making. Research has shown that evaluators under time-constraints, for instance, are more likely to use status characteristics such as gender to evaluate performance of different candidates (Biernat, Kobrynowicz, and Weber 2003). On the supply side, there is often variation, whether through choice or circumstance, about the degree to which information about gender or race can be inferred through the stimuli supplied to the evaluator, shifting the ability of the evaluator to use these ascriptive characteristics to form judgments about individuals (Goldin and Rouse 2000; Bertrand and Mullainathan 2003). At the same time, the advantage or disadvantage of maleness is entirely contingent on the match between the gendered stereotypes of a role and the individual's performance. For instance, women might be evaluated more leniently in certain types of leadership behavior, leading to an paradoxical advantage because they were expected to underperform their male peers (Biernat and Fuegen 2001). Lastly, prevailing rates of female participation in certain roles, like entrepreneurship, can vary across countries or regions, leading to varying expectations about female performance depending upon the background and location of the evaluator (Thébaud 2015).

In different evaluation regimes, it is possible to shift the supply-side of gender saliency across regimes. For example, when evaluators are asked to individually judge the quality of an applicant using a paper or electronic application form, the only gender information available to the judge is provided via the names of the co-founders. In contrast, in person interviews (either with individuals or by committee) makes the gender of the individuals being evaluated more salient. For example, if a team of co-founders is asked to present a short pitch and answer questions for the judges, then the gender cues are much stronger than for a paper-based presentation. Comparing the in-person (or in-skype) presence of women on an early-stage company's founding team with that of a paper-based evaluation represents a significant difference in terms of the richness of the gender cue across the two conditions.

Hypothesis 2: *Co-founder gender will have a larger impact on the evaluation of early-stage firms when the salience of this and other status characteristics is increased*

Lastly, the importance of gender and other ascriptive characteristics can vary depending upon the identity of the evaluator. While individuals might understand the prevailing expectations about a particular race or gender, they might also have individual preferences for that race or gender, due to their membership in that ascriptive group, that outweighs their use of these more diffuse general social expectations. Across different sociological and social psychological traditions, the mechanism generating a preference for similar others has been termed homophily (McPherson, Smith-Lovin, and Cook 2001) or in-group bias (Brewer 1979). Regardless of the exact social-psychological basis for the preference, it is reasonable to expect there to be differences between judge-startup dyads where gender is shared or different.

Hypothesis 3: *Male Judges will be additionally critical of female-founded startups*

4.2.2. Theories of Group Performance of Evaluation Tasks

Nearly all of the experimental evidence for the importance of status characteristics exists in settings where individuals make evaluations of ideas or others without reference to other evaluators. While there are sociological studies of group decision-making (Ridgeway 2001), most of the literature focuses on close qualitative accounts of group processes that help enrich the theoretical understanding of status mechanisms that generate and perpetuate inequality. In short, there is relatively little sociological or social psychological literature that explores the role of potential gender or status biases that may arise in evaluation settings where groups are involved. This is despite a large and voluminous literature across psychology and experimental economics which shows that groups that groups do as well or outperform individual judgment in a range of tasks (Blinder and Morgan 2005; Charness and Sutter 2012). This is particularly troubling in the light of the fact that committees are a standard evaluation regime across many settings where gender bias has been recorded: venture capital investment (Brooks et al. 2014) and academic tenure decisions (Bailyn 2003). There are number of theories that have been proposed for the observed performance advantage of teams. We review three below: increased cognitive capacity, increased solution finding, groups as checks against our biases. Lastly, we will consider how the composition of evaluation committees may shift the evaluations of startup projects, especially those that are female-led, using theories of how status characteristics of group members shape group processes.

One of the ways in which "two heads are better than one" is that two heads may be able to hold more easily hold complex points of view that can be synthesized through conversation than may be undertaken by individuals. A broad range of studies in experimental economics have found that groups more quickly converge to the optimal solution of a game than do individuals, especially in games that are particularly complex or counterintuitive (Charness and Levin 2005; Cooper and Kagel 2005). In these complex games where alternative points of view of opposing players are required, the authors of the studies suggest that teams are more easily able to model the game through the simulation of each side of the game or through teammates holding opposing points of view on potential choices. The clear articulation of points of view with the down-selection of these views through group discussion is a cognitive architecture easily available to groups but is more difficultly mimicked by individuals.

A related but distinct advantage of "two heads" is the increased probability that individuals will find a solution to a game or task through parallel search. Both psychologists and economists have studied individual and group performance on tasks like the "Wason selection task" which satisfy a simple "truth wins" criterion: the tasks are easy to solve once a rule/insight is gained about the task and this insight/solution is hard to find but easily transmitted once discovered. If *p* is the probability that any individual finds the solution concept shared across the tasks, then the probability of the group finding the solution is 1-(1-p)^n even if the individuals in the group do not influence each other's efficacy. This means that groups will always perform better than an individual in "truth wins" type tasks simply by aggregating the possibility that a solution is found. Studies exploring these types of tasks have found performance of groups somewhere between the p and 1-(1-p)^n boundaries for the predicted performance of groups (Michaelsen, Watson, and Black 1989; Maciejovsky et al. 2013).

Lastly, "two heads are better than one" because individuals find it easier to adhere to aspirational behavior when it is viewed as a group norm or lack of adherence to the ideal behavior is viewed as defection from the group. Research has shown that group interaction increase the likelihood of beneficial behaviors such as studying and weight loss amongst students (Babcock et al. 2015) as well as loan-repayment amongst microloan recipients (Feigenberg, Field, and Pande 2010) and fair enforcement of rules towards minority voters in India (Neggers 2015). Across these diverse settings, the creation of groups or the composition of the group had a substantial impact on the behavior of individuals in the group relative to a baseline behavior outside of a group. Taken together, the three channels above lead us to the following hypothesis:

> ***Hypothesis 4:*** *Group-based deliberations are more likely to outperform individual deliberation in their ability to recognize high quality startups*

While there is substantial evidence that groups may be able to complete tasks at a higher level and make better decisions than individuals, groups do not always perform well. The potential for 'groupthink' is a large danger in groups that has led to famously tragic results (Janis 1972; Bénabou 2012). The problem of groupthink begins with the basic psychological force for individuals to conform with group processes and decisions even in the face of clear personal information that the group decision/behavior is incorrect (Asch 1951; Goeree and Yariv 2007). During group deliberations, An individual's preference for conformity leads to their overweighting shared information and beliefs during discussions over differing information (e.g. voting or scoring) (Stasser and Titus 1985). This preference for emphasizing common ground is posited to lead to convergence that is too quick relative to optimal length of group deliberation.

Of course, conformity with group opinion is not always an irrational behavior. A substantial literature in the economics of information has stressed the potential importance of evaluatory signals from other individuals in situations of high uncertainty. In a baseline models of herding (Banerjee 1992; Bikhchandani, Hirshleifer, and Welch 1992), individuals are rationally more likely to imitate the behavior of others who have faced a decision in advance of them when access to information about the performance implications of various choices is either restricted from them or too costly to obtain. The herding dynamics in these analytical models are echoed by the results of behavioral simulations of agents with similar information restrictions on the performance implications of different choices (Strang and Macy 2001; David and Strang 2006). While presented in a stark manner in these models, the tradeoffs between imitation and costly search for performance information that underlie the social influence of choices ranging from personal investment strategy (Shiller 1995) to choice of music (Salganik, Dodds, and Watts 2006; Salganik and Watts 2008).

Hypothesis 5: *Individuals in a committee setting will exhibit evidence of peer-effects in their voting*

The theories we have considered above have largely ignored how the composition of the committee might impact the deliberative process, especially with regards to the status characteristics of the startup teams. There is a long tradition in social psychology that has explored how the status characteristics of group members shape group dynamics. This literature has shown that differences in the starting status of each group member predicts the degree of contributions given by that individual in group deliberations (Bales 1950) and also the deference paid to that individual and their opinions within the group (Strodtbeck, James, and Hawkins 1957). Thus, we might expect any potential biases held by higher status group members, for instance male judges, to predominate within group decision-making. On the other hand, the use of status characteristics in groups emerges from fast-forming group norms (Ridgeway 2001) and the presence of female judges on the committee might shift the capacity of groups to use gender cues as a rationale for devaluing the capacity of female-led teams (Berger et al. 1992). Thus, social psychological theories of the role of status characteristics within group processes provide two contradicting views of the impact of committee's gender composition on their evaluation of the capacity of early-stage startup teams.

4.3. EMPIRICAL APPROACH

Our research design explores the variation in evaluation outcomes for a fixed set of judges and a fixed set of early-stage ventures that are evaluated through two distinct evaluation regimes: individual-based and committee-based evaluation. In an ideal setting to explore the ways in which gender and other characteristics are evaluated under different regimes, and by judges of different genders, an observer would like to independently vary both the gender saliency of the project team, the evaluation regime, and the gender of the evaluators. We are fortunate in having a setting that closely conforms to such a situation and to these design criteria.

In our setting, as described below, each early-stage startup is randomly allocated to two sets of judges and evaluated in each of these two regimes. Moreover, each judge evaluates multiple startups in both the individual and committee evaluation regimes. Most importantly, in the setting of our study, the matching between judge, startup, and evaluation regime is randomized explicitly.[1] Exploiting this random allocation of judges, startups, and evaluation regime, we can separately identify the impact of founder status characteristics on the evaluation of these firms across the two evaluation regimes.

By exploring differences across these two evaluation regimes, we show how judges use information differently depending upon the context in which they are evaluating new firms. Recent research has focused on the impact of evaluator background on the outcomes of the evaluation process (Boudreau et al. 2016; Li and Agha 2015), providing evidence that the composition of a committee matters for the eventual evaluation of a new idea or venture. In addition, recent research has attempted to link evaluations across very different platforms (Mollick and Nanda 2015), showing substantial similarity between expert and crowd-based judgment. This paper hopes to bridge these two literatures by asking not only how compositional differences across different committees impacts the eventual judgment of that committee, but also how different evaluation regimes change the judgment processes of individual judges.

[1] Allocation of judges was random in the MassChallenge program but stratified on industry-expertise attempted gender balance.

After establishing whether there are differences across these two contexts, we characterize the performance implications of these two evaluation regimes using methodologies from previous studies (Li and Agha 2015).

4.3.1. Institutional Setting: MassChallenge

In this study, we utilize observational data on two types of judging in one high quality entrepreneurship program, MassChallenge, across multiple years. MassChallenge is a startup accelerator founded in Boston. (While it has now expanded to other regions, our data only include start-up ventures that applied to MassChallenge Boston). MassChallenge receives yearly thousands of applications from around the world from early-stage ventures that wish to enter their four-month residential program in Boston (and now elsewhere). During the program, startups receive substantial mentoring and education[2], but this paper focuses specifically on the evaluation of these firms prior to their admission (or decline of admission) from the program.

The details of the MassChallenge evaluation process provide a natural setting in which to explore the impact of differing evaluation regimes. Of the thousands of firms that apply to the program, a subset of these firms receives two types of evaluation within one month of each other by virtue of having made it past an initial first screening round. For this paper, we will focus on the subset of firms that are evaluated by both modalities of evaluation – individual and committee - in order to be able to compare a fixed set of firms across two different types of evaluation regime.

When a judge agrees to participate in selecting candidates for the MassChallenge program, they agree to review a number of startup applications through two rounds. In the first round (individual-based regime), each judge individually receives a number of applications through MassChallenge's online platform. Each judge is asked to score the startup based on this written application. They do not know which other judges are evaluating this each startup in the first round nor do they have information on the evaluations of that startup by other judges. MassChallenge then aggregates the scores of each individual judge through a simple average. Startups in the top 250-300 in this round are then passed on to the

[2] For more information on the impact of the MassChallenge program itself, see Fehder (2015).

second round. In the second round (committee-based regime) the startups are given five minutes to give an in-person (or Skype-based) pitch to the judges and then have a five-minute Q&A period. Through this in-person presence, the status characteristics of the founders are made more salient. After these ten minutes with the startup's founding team, the committee are given five minutes to commune and then each judge scores the startup on their own. Critically, the same score sheet is used in both round 1 and round 2.

We believe that the first round of judging in our setting (individual-based regime), closely approximates crowd-based mechanisms (Surowiecki 2005) in the sense that each individual gives their evaluation of a startup without any social context. In contrast, the second round of judging (committee-based regime) approximates the selection mechanisms seen in many other settings where ideas are selected including NIH panels (Li and Agha 2015), Angel meetings (Kerr, Lerner, and Schoar 2010), and Venture Capitalist partner meetings (Kerr, Nanda, and Rhodes-Kropf 2014).

3.2 Analytical Approach

By following the evaluation behavior of a fixed set of judges across a fixed set of business opportunities but across two different evaluation regimes, we hope to capture differences across these regimes in terms of individual judging behavior (i.e. the use of co-founder status characteristics in evaluating an early-stage startup) and efficacy (which regime is better) and potential sources of performance differences.

Our main empirical specifications evaluate the individual evaluations of judges across both evaluation regimes There is a long tradition of examining funded and unfunded business plans to extrapolate the features of successful business ventures with respect to one of their earliest incarnations (MacMillan, Siegel, and Narasimha 1986; Roberts 1991; Kirsch, Goldfarb, and Gera 2009). We build upon the use of choice-based analysis of evaluation across specific institutional arrangements that has become well-used in economics (Kirsch, et al. 2009; Ackerberg 2007; Li 2012), sociology (Castilla 2011), and management (Boudreau et al. 2016).

In our first set of regressions, we will look at how the characteristics of the founders influence the average score received in the first and second round of judging as well as role of evaluator characteristics on the average score received. The main regressions of interest will examine each round separately and measure the impact of measures of founder ascriptive identity and human capital. Specifically, we will measure:

$$\mathrm{S_i} = \alpha + \beta_1\ \mathrm{FEMALE\_FOUNDER_i} + \beta_2\mathrm{ELITE\_EDU}_{\mathrm{i}} + \beta_3\ \mathrm{STEM\_EDU_i} + \beta_4\mathrm{PRIOR\_ENTR_i} + \varepsilon_{\mathrm{i}} \quad (1)$$

Here, $\mathrm{S_i}$ is the average score received by each of the startups in a particular round of judging. The purposes of these regressions is to evaluate how much founder characteristics impact the average score received by each startup in both the of the rounds of judging.

Next, we move to the analysis of the choices of individual judges and how they vary across the two evaluation regimes. To do so, we will evaluate a series of regressions that observe individual judge behavior across different startups and committees. The general framework for these regressions is as follows:

$$SCORE_{s,j,r} = f(\varepsilon_{s,j};\ X_{s,r}, X_{j,r}, X_{j-s}) \qquad (3)$$

We are interested in predicting the score given to a startup, indexed by s, by a particular judge, indexed by the subscript j, in a given round, indexed by the subscript r. The Vector $X_{s,r}$ describes a set of round-specific attributes of the startup. We are especially concerned with status attributes of the startup's founders and their impact on the score received by the startup across round. $X_{j,r}$ describes a vector of attributes describing the judge. Throughout much of the analysis, this vector will be absorbed by a judge-specific fixed effect. Lastly, X_{j-s} measures interactions between the characteristics of judges and startups in terms of background. Overall, we are interested in assessing potential differences in the impact of these founder attributes on the decision processes of individual judges across the rounds.

In our evaluation of the behavior of individual judges, we were not able to adequately control for the variation in quality across the different startups. If startup founders with different backgrounds are associated with startups of different quality, then our estimates of the impact of gender and the other status and human capital traits we measure might be biased by their correlation with underlying quality differences. To address this concern, we will examine whether these founder traits drive changes in the difference between the scoring between the first two rounds. Building off of prior work examining differences between evaluations in other economic settings (Castilla 2011), we estimate the following models:

$$\mathrm{D_i} = \alpha + \beta_1\,\mathrm{FEMALE\_FOUNDER_i} + \beta_2\mathrm{ELITE\_EDU_i} + \beta_3\,\mathrm{STEM\_EDU_i} + \beta_4\mathrm{PRIOR\_ENTR_i} + \varepsilon_i \quad (2)$$

Here, $\mathrm{D_i}$ is the difference between the scores received by the same startup in the first and second round. By focusing on the differences across the rounds of evaluation, we control systematically for variation in the quality of the different startups in our sample in a way similar to other differencing methods in econometrics (Wooldridge 2010). We are concerned with the influence of certain status characteristics, like gender and elite education, on the differences across the two evaluation regimes. Specifically, we are interested in seeing if the impact of gender increases as gender becomes more salient in the evaluation process.

4.5. RESULTS

Our regression results begin with Table 3, which contains a series of linear probability models the average score level. Specifically, we predict the Average Score received in each round using a series of OLS regressions. Model 3-1 examines the impact of various earned and ascriptive characteristics on the average score received by each startup in round 1. The results show that startups whose founders have elite degrees are on average scored 2 points higher on average (the maximum score is 100). In addition, startups whose founders have an MBA degree receive 2.41 points higher evaluations in round 1. In Model 3-2, we move from round 1 to round 2. Similar to round 1, startups with founders holding degrees from elite institutions receive 8.6 more points on average. In Round 2, however, gender becomes a significant predictor of round 2 score, predicting an increase of 6.332 higher score on average for startups with at least one female founder. This accords with the predicted importance of increased saliency in the use of gender for evaluation, but the direction of the magnitude (positive) is different from the majority of the gender evaluation literature.

Next, we ask if characteristics of the judges makes a difference for the average scores received by each team in each round. Model 3-3 begins this analysis by adding two features of the committee, number of reviewers and number of female reviewers that might impact the score received by the company. We find no significant impact of these factors on the score received in the paper round of judging. Next, Model 3-4 examines the impact of these judge-side factors on evaluation in round 2. Similar to our findings in the previous model, we find no significant evidence for the impact of judge-side factors for committee round scores. Next, we ask whether the count of female judges matters less than the mere presence of at least one female judge in the evaluation process. In Model 3-5, we assess the impact of at least one female judge on the average score in the paper round, finding no effect. Next, we conduct the same analysis on committee rounds in Model 3-6, finding no significant effect of having at least one female judge.

Overall, Table 3 suggests that the importance of status characteristics such as gender and elite education become more important in the evaluation of early-stage firms in the second round where the aspects of the cofounders become more salient relative to the first round of judging, supporting hypothesis 2. Interestingly, the positive and large increase in the average scores of female co-founded ventures seems to provide evidence against hypothesis 1. While the gender of the cofounders may increase in saliency in round 2, the committee structure of round 2 seems to produce a positive bias in the evaluation of these firms, rather than the negative effect expected from the rest of the literature.

In Table 4, we move from an analysis of individual judge's scores pooled across both rounds to examining the differences across rounds. Overall, we present OLS regression models predicting the score chosen by each judge Model 4-1 shows systematically different impact of the status characteristics of founders across the two funding rounds. In particular, the estimated coefficient for each of the status characteristics in round 2 is larger than the coefficient for the same status characteristic in round 1 for all but prior entrepreneurship and MBAs. We conduct an F-test to evaluate the possibility that all of the coefficients from Round 2 are the same as the Round 1 coefficients and find substantial evidence (F = 23.26) to reject this hypothesis. Next, we ask whether our key founder characteristic is different across the two rounds, and we find statistically significant evidence that female-led firms are judged differently across the two rounds (F = 8.03). Model 5-2 adds the gender identity of the judge by round to the analysis. It finds no evidence that female-led startups are evaluated differently by male judges across either of the evaluation rounds. Overall, Table 4 provides strong evidence that the use of status characteristics amongst judges differs across the two rounds.

In Table 5, we move from the level of individual judges back to an analysis of average score. This time, we focus on the difference in scores received across rounds and present a series of regressions that examines what founder characteristics predict score changes. By focusing on differences across rounds, we are able to hold the quality of the startups constant and focus on how the evaluation of founder characteristics changes. In Model 5-1, we show that having a female cofounder is predictive of showing a large (5.895) and statistically significant score change between the two rounds of judging. In Models 5-2 through 5-4, we run the same regressions with the inclusion of various evaluator-side characteristics and find that our estimates of the impact of various founder characteristics are robust to the inclusion of evaluator characteristics.

Next, we attempt to characterize the potential channels for these effects as well as the performance implications in Table 6. First, we look at the impact of the scoring choices of other judges on the choice of score by each judge across the two rounds in Model 6-1. The predictive power of the leave-out mean on the score choice of a judge can come from two sources: shared perceptions of the quality of the firm and consensus formation amongst evaluators. Obviously, such consensus formation is impossible in the paper round, so our estimate of the impact of the average score given by other judges in this round shows us how much the judges agree on the quality of the candidate without consulting. We find that a 1 point increase in the average score of the other judges predicts a 0.14 point increase (out of 100) by the focal judge. In contrast, in the committee round, the average score given by the other judges is highly predictive of the focal judge's score. A one point increase in the average score of the other judges predicts a 0.8 increase in the score given by the focal judge, suggesting that there is a large amount of consensus-building in the committee round.

In Model 6-2 we attempt to further characterize the difference in the impact of other judges on the scoring decisions of individual judges by considering the impact of more judges. In the paper round, the addition of an additional judge has a large and statistically significant decrease on the average predicted score choice of an individual judge. We interpret this result as saying that there will be a mean reversion process at work in the paper round such that a higher number of judges suggests that the average judge will be more likely to have a lower score as a result of the central tendency in the data generating process. Meanwhile, the interaction term now seems to be playing the role of the leave-out mean in the paper round. Interestingly, these mean reversion processes do not seem to be active in the committee round. Neither the coefficient on the number of judges or the interaction between number of judges and leave-out mean is statistically significant in the committee round. At the same time, the magnitude of the leave-out mean in the committee round is largely the same as in model 6-1.

Lastly, Models 6-3 through 6-6 characterize the performance implications of the different modes of evaluation. By including a measure of ex post realized quality in our models of individual judge score choice, we are able to ask the question of whether an individual judge is more likely to choose better firms individually or as a member of a committee (i.e. are more sensitive to this ex post measure of quality which will be correlated with factors unobserved to the econometrician). In this regression, Judges are sensitive to ex post quality in both modes of evaluation, but they seem to be more sensitive to start-up quality in the committee-based evaluation regime. When we test the hypothesis that the coefficients measuring judge sensitivity are the same across evaluation modes, we find significant evidence to reject the null that they are the same (F = 118.68). Thus, we can say that a given judge on average seems more able to predict the quality of the startups in committee-based regime. One concern about this result is that we are measuring part of the treatment effect of the program in this exercise as scores received predict admission to MassChallenge. To ensure that our results are robust to admission status, we look at the subset of non-admitted firms in Appendix B and find that our findings seem to hold in this subsample that are not affected by the MassChallenge treatment (i.e. committee evaluation seems to make an individual judge on average more attentive to the quality of the startup). In Models 6-4 through 6-6, check that our measure of the judge sensitivity to ex post quality is robust to the inclusion of founder characteristics and find that our estimates do not change with the inclusion of these factors.

4.6. CONCLUSION AND DISCUSSION

Much of the literature on innovation – in the economy and particularly within large corporations – has emphasized the role of selection in winnowing the funnel of innovation projects as they develop and require additional resources (see Tswisch and Ulrich). However relatively little is know about the ways in which the design of evaluation regimes at each stage in the funnel shape the portfolio of projects that are selected. In contributing to the nascent literature exploring the impact of the structure of evaluation on the project selection, we attempt to clarify the degree of different types of bias in different evaluation regimes. Our results point to the central role of the type of evaluation regime on the types of ideas selected and the use of different types of information by evaluators.

Judges seem to use information about the backgrounds and identities of the co-founders of projects at far higher rates in our committee round, yet their use of this information seems to lead to better choices (in the sense that the scores are more related to downstream external investment). As a potential channel through which we might explain these stark differences across evaluation regime, we demonstrated a strong convergence effect in the committee evaluation regime that does not seem to exist in the individual paper regimes.

One of the final elements which is not directly addressed in our current analysis is the strong positive impact of female founder's gender identity in the committee evaluation regime. While we do purge some of the quality concerns through our analysis of score differences across rounds, we are not able to address the possibility that female-founded firms are actually on average of higher quality in our sample. It is a possibility that female-led ventures in our sample might be associated with higher quality ventures and that the committee-based judges are able to pick out these higher quality projects. Experimental economics has demonstrated a general undersupply of women into competitions (Niederle and Vesterlund 2005), suggesting that we might expect that female-led teams require higher quality hurdles before they will engage.

4.7. REFERENCES

Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and Identity." *Quarterly Journal of Economics*, 715–53.

Asch, Solomon E. 1951. "Effects of Group Pressure upon the Modification and Distortion of Judgments." *Groups, Leadership, and Men. S*, 222–36.

Azoulay, Pierre, Toby Stuart, and Yanbo Wang. 2013. "Matthew: Effect or Fable?" *Management Science* 60 (1): 92–109.

Babcock, Philip, Kelly Bedard, Gary Charness, John Hartman, and Heather Royer. 2015. "Letting down the Team? Social Effects of Team Incentives." *Journal of the European Economic Association* 13 (5): 841–70.

Bailyn, Lotte. 2003. "Academic Careers and Gender Equity: Lessons Learned from MIT1." *Gender, Work & Organization* 10 (2): 137–53.

Bales, Robert F. 1950. "Interaction Process Analysis; a Method for the Study of Small Groups." http://doi.apa.org/psycinfo/1950-04553-000.

Banerjee, Abhijit V. 1992. "A Simple Model of Herd Behavior." *The Quarterly Journal of Economics*, 797–817.

Bénabou, Roland. 2012. "Groupthink: Collective Delusions in Organizations and Markets." *The Review of Economic Studies*, September, rds030. doi:10.1093/restud/rds030.

Berger, Joseph, Joseph Berger, M. Hamit Fisek, and Robert Zane Norman. 1977. *Status Characteristics and Social Interaction: An Expectation-States Approach*. Elsevier New York. http://library.wur.nl/WebQuery/clc/223456.

Berger, Joseph, Robert Z. Norman, James W. Balkwell, and Roy F. Smith. 1992. "Status Inconsistency in Task Situations: A Test of Four Status Processing Principles." *American Sociological Review*, 843–55.

Bernstein, Shai, Xavier Giroud, and Richard R. Townsend. 2015. "The Impact of Venture Capital Monitoring." *The Journal of Finance*. http://onlinelibrary.wiley.com/doi/10.1111/jofi.12370/full.

Bernstein, Shai, Arthur G. Korteweg, and Kevin Laws. 2015. "Attracting Early Stage Investors: Evidence from a Randomized Field Experiment." *Journal of Finance, Forthcoming*, 14–17.

Bertrand, Marianne, and Sendhil Mullainathan. 2003. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." National Bureau of Economic Research. http://www.nber.org/papers/w9873.

Biernat, Monica, and Kathleen Fuegen. 2001. "Shifting Standards and the Evaluation of Competence: Complexity in Gender-Based Judgment and Decision Making." *Journal of Social Issues* 57 (4): 707–24.

Biernat, Monica, Diane Kobrynowicz, and Dara L. Weber. 2003. "Stereotypes and Shifting Standards: Some Paradoxical Effects of Cognitive Load." *Journal of Applied Social Psychology* 33 (10): 2060–79.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1992. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *Journal of Political Economy*, 992–1026.

Blinder, Alan S., and John Morgan. 2005. “Are Two Heads Better than One? Monetary Policy by Committee.” *Journal of Money, Credit and Banking*, 789–811.

Botelho, Tristan L., and Mabel Abraham. 2016. “Pursuing Quality: How Uncertainty Magnifies Double Standards in A Multistage Evaluation Process.” *ADMINISTRATIVE SCIENCE QUARTERLY*.

Boudreau, Kevin J., Eva C. Guinan, Karim R. Lakhani, and Christoph Riedl. 2016. “Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science.” *Management Science*. http://pubsonline.informs.org/doi/abs/10.1287/mnsc.2015.2285.

Brewer, Marilynn B. 1979. “In-Group Bias in the Minimal Intergroup Situation: A Cognitive-Motivational Analysis.” *Psychological Bulletin* 86 (2): 307.

Brooks, Alison Wood, Laura Huang, Sarah Wood Kearney, and Fiona E. Murray. 2014. “Investors Prefer Entrepreneurial Ventures Pitched by Attractive Men.” *Proceedings of the National Academy of Sciences* 111 (12): 4427–31.

Castilla, Emilio J. 2011. “Bringing Managers Back In Managerial Influences on Workplace Inequality.” *American Sociological Review* 76 (5): 667–94.

Charness, Gary, and Dan Levin. 2005. “When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect.” *The American Economic Review* 95 (4): 1300–1309.

Charness, Gary, and Matthias Sutter. 2012. “Groups Make Better Self-Interested Decisions.” *Journal of Economic Perspectives* 26 (3): 157–76. doi:10.1257/jep.26.3.157.

Cooper, David J., and John H. Kagel. 2005. “Are Two Heads Better than One? Team versus Individual Play in Signaling Games.” *American Economic Review*, 477–509.

Correll, Shelley J., and Cecilia L. Ridgeway. 2003. “Expectation States Theory.” In *Handbook of Social Psychology*, 29–51. Springer. http://link.springer.com/chapter/10.1007/0-387-36921-X_2.

David, Robert J., and David Strang. 2006. “When Fashion Is Fleeting: Transitory Collective Beliefs and the Dynamics of TQM Consulting.” *Academy of Management Journal* 49 (2): 215–33.

Feigenberg, Benjamin, Erica M. Field, and Rohini Pande. 2010. “Building Social Capital through Microfinance.” National Bureau of Economic Research. http://www.nber.org/papers/w16018.

Fiske, Susan T., Amy JC Cuddy, Peter Glick, and Jun Xu. 2002. “A Model of (often Mixed) Stereotype Content: Competence and Warmth Respectively Follow from Perceived Status and Competition.” *Journal of Personality and Social Psychology* 82 (6): 878.

Foschi, Martha. 1996. “Double Standards in the Evaluation of Men and Women.” *Social Psychology Quarterly* 59 (3): 237–54. doi:10.2307/2787021.

Foschi, Martha, Larissa Lai, and Kirsten Sigerson. 1994. “Gender and Double Standards in the Assessment of Job Applicants.” *Social Psychology Quarterly*, 326–39.

Goeree, Jacob K., and Leeat Yariv. 2007. “Conformity in the Lab.” http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.192.4785.

Goldin, C., and C. Rouse. 2000. “Orchestrating Impartiality: The Impact of Blind Auditions on the Sex Composition of Orchestras.” *American Economic Review* 90 (4): 715–41.

Hallen, Benjamin L. 2008. “The Causes and Consequences of the Initial Network Positions of New Organizations: From Whom Do Entrepreneurs Receive Investments?” *Administrative Science Quarterly* 53 (4): 685–718.

Honig, Benson, and Tomas Karlsson. 2004. “Institutional Forces and the Written Business Plan.” *Journal of Management* 30 (1): 29–48.

Janis, Irving L. 1972. “Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes.” http://doi.apa.org/psycinfo/1975-29417-000.

Kaplan, S.N, Berk A. Sensoy, and P. Strömberg. 2009. “Should Investors Bet on the Jockey or the Horse? Evidence from the Evolution of Firms from Early Business Plans to Public Companies.” *The Journal of Finance* 64 (1): 75–115. doi:10.1111/j.1540-6261.2008.01429.x.

Kerr, W.R., J. Lerner, and A. Schoar. 2010. “The Consequences of Entrepreneurial Finance: A Regression Discontinuity Analysis.” NBER.

Kerr, W.R., Ramana Nanda, and Matthew Rhodes-Kropf. 2014. “Entrepreneurship as Experimentation.” *Journal of Economic Perspectives* 28 (3): 25–48. doi:10.1257/jep.28.3.25.

Kirsch, David, Brent Goldfarb, and Azi Gera. 2009. “Form or Substance: The Role of Business Plans in Venture Capital Decision Making.” *Strategic Management Journal* 30 (5): 487–515.

Li, Danielle, and Leila Agha. 2015. “Big Names or Big Ideas: Do Peer-Review Panels Select the Best Science Proposals?” *Science* 348 (6233): 434–38. doi:10.1126/science.aaa0185.

Maciejovsky, Boris, Matthias Sutter, David V. Budescu, and Patrick Bernau. 2013. “Teams Make You Smarter: How Exposure to Teams Improves Individual Decisions in Probability and Reasoning Tasks.” *Management Science* 59 (6): 1255–70.

MacMillan, Ian C., Robin Siegel, and PN Subba Narasimha. 1986. “Criteria Used by Venture Capitalists to Evaluate New Venture Proposals.” *Journal of Business Venturing* 1 (1): 119–28.

McPherson, M., L. Smith-Lovin, and J.M. Cook. 2001. “Birds of a Feather: Homophily in Social Networks.” *Annual Review of Sociology*, 415–44.

Michaelsen, Larry K., Warren E. Watson, and Robert H. Black. 1989. “A Realistic Test of Individual versus Group Consensus Decision Making.” *Journal of Applied Psychology* 74 (5): 834.

Mollick, Ethan, and Ramana Nanda. 2015. “Wisdom or Madness? Comparing Crowds with Expert Evaluation in Funding the Arts.” *Management Science*. http://pubsonline.informs.org/doi/abs/10.1287/mnsc.2015.2207.

Neggers, Yusuf. 2015. “Enfranchising Your Own? Experimental Evidence on Polling Officer Identity and Electoral Outcomes in India.” *Working Paper*.

Niederle, Muriel, and Lise Vesterlund. 2005. “Do Women Shy Away from Competition? Do Men Compete Too Much?” National Bureau of Economic Research. http://www.nber.org/papers/w11474.

Phelps, Edmund S. 1972. “The Statistical Theory of Racism and Sexism.” *The American Economic Review* 62 (4): 659–61.

Ridgeway, Cecilia L. 2001. “Social Status and Group Structure.” *Blackwell Handbook of Social Psychology: Group Processes*, 352–75.

Roberts, E.B. 1991. *Entrepreneurs in High Technology: Lessons from MIT and beyond*. Oxford University Press, USA.

Salganik, M.J., P.S. Dodds, and D.J. Watts. 2006. “Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market.” *Science* 311 (5762): 854.

Salganik, M.J., and D.J. Watts. 2008. “Leading the Herd Astray: An Experimental Study of Self-Fulfilling Prophecies in an Artificial Cultural Market.” *Social Psychology Quarterly* 71 (4): 338–55.

Scott, Erin L., Pian Shu, and Roman Lubynsky. 2015. “Are’Better’Ideas More Likely to Succeed? An Empirical Analysis of Startup Evaluation.” *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, no. 16-013. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2638367.

Shane, Scott, and Toby Stuart. 2002. “Organizational Endowments and the Performance of University Start-Ups.” *Management Science* 48 (1): 154–70.

Shiller, Robert J. 1995. “Conversation, Information, and Herd Behavior.” *The American Economic Review* 85 (2): 181–85.

Simcoe, Timothy, and D.M. Waguespack. 2011. “Status, Quality, and Attention: What’s in a (missing) Name?” *Management Science* 57 (2): 274.

Stasser, Garold, and William Titus. 1985. “Pooling of Unshared Information in Group Decision Making: Biased Information Sampling during Discussion.” *Journal of Personality and Social Psychology* 48 (6): 1467.

Strang, David, and Michael W. Macy. 2001. “In Search of Excellence: Fads, Success Stories, and Adaptive emulation1.” *American Journal of Sociology* 107 (1): 147–82.

Strodtbeck, Fred L., Rita M. James, and Charles Hawkins. 1957. “Social Status in Jury Deliberations.” *American Sociological Review* 22 (6): 713–19.

Stuart, Toby, Ha Hoang, and Ralph C. Hybels. 1999. “Interorganizational Endorsements and the Performance of Entrepreneurial Ventures.” *Administrative Science Quarterly* 44 (2): 315–49. doi:10.2307/2666998.

Surowiecki, James. 2005. *The Wisdom of Crowds*. Anchor. https://books.google.com/books?hl=en&lr=&id=hHUsHOHqVzEC&oi=fnd&pg=PR11&ots=ZrcDXmOpfi&sig=QcvNr8x8Pb9FFgeXJwwr2sTov9k.

Tajfel, Henri. 1982. “Social Psychology of Intergroup Relations.” *Annual Review of Psychology* 33 (1): 1–39.

Thébaud, Sarah. 2015. “Status Beliefs and the Spirit of Capitalism: Accounting for Gender Biases in Entrepreneurship and Innovation.” *Social Forces* 94 (1): 61–86.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT press. https://books.google.com/books?hl=en&lr=&id=yov6AQAAQBAJ&oi=fnd&pg=PP1&dq=Wooldridge+econometric+analysis+of+cross-sectional&ots=iWi0BHFD0U&sig=22OaDm5lL1u5rHqgra8_AcowRuU.

4.8. TABLES AND FIGURES

Table 1: Variable Definitions

Variable	Definition	Source
Startup Characteristics		
Female Founder	Dummy variable = 1 if at least one cofounder is female	MC, LI
Elite Edu	Dummy = 1 if at least one cofounder attended a	MC, LI
STEM	Dummy = 1 if at least one cofounder received a degree in engineering, science, or math	MC, LI
MBA	Dummy = 1 if at least one cofounder received an MBA	MC, LI
Ln Funding	Logged Dollars of Outside investment received in first two years after potential graduation from MC program	VX, CB
Judge Characteristics		
Female Judge	Dummy = 1 if judge is female	MC
Dependent Variables		
Judge Score	Score given by one judge	MC
Avg. Judge Score, Paper Round	Average Score across all judges in Paper Round	MC
Avg. Judge Score, Committee Round	Average Score across all judges in Commitee Round	MC
Judge Score Difference	Difference of the	MC

MC – MC application; LI – LinkedIn; VX – VentureXpert; CB – Crunchbase

Table 2: Summary Statistics

	Mean	Std. Dev	Min	Max
Startup Characteristics				
Female Founder	0.15	0.36	0	1
Elite Edu	0.21	0.41	0	1
STEM	0.26	0.44	0	1
MBA	0.15	0.36	0	1
Ln Funding	3.26	5.74	0	17.15
Judge Characteristics				
Female Judge	0.16	0.37	0	1
Round Characteristics				
# Judges, Paper Round	4.69	1.77	2	10
# Judges, Committee Round	5.07	0.87	2	9
Dependent Variables				
Judge Score	49.59	27.48	0	100
Avg. Judge Score, Paper Round	65.10	10.13	16	95
Avg. Judge Score, Committee Round	49.66	23.98	5	100
Judge Score Difference	-16.24	24.36	-78	52

Table 3: Predictors of Average Score for Each Round

	(1) Avg Score, Paper Round	(2) Avg Score, Committee	(3) Avg Score, Paper Round	(4) Avg Score, Committee	(5) Avg Score, Paper Round	(6) Avg Score, Committee
Female Founder	0.437 (0.847)	6.332*** (1.941)	0.461 (0.850)	6.431*** (1.947)	0.453 (0.848)	6.306*** (1.943)
Serial Entrepreneur	0.175 (0.930)	3.572* (2.130)	0.186 (0.931)	3.534* (2.131)	0.162 (0.931)	3.545* (2.134)
Elite Degree	1.963** (0.925)	8.605*** (2.119)	1.984** (0.927)	8.689*** (2.124)	2.012** (0.927)	8.590*** (2.124)
MBA Degree	0.939 (0.952)	5.505** (2.181)	0.936 (0.953)	5.488** (2.182)	0.944 (0.952)	5.509** (2.183)
STEM Degree	1.069 (0.825)	8.866*** (1.889)	1.030 (0.827)	8.855*** (1.894)	1.006 (0.827)	8.918*** (1.895)
# Judges			-0.463 (0.562)	1.093 (1.288)	-0.434 (0.553)	0.858 (1.268)
# Female Judges			-0.090 (0.552)	-1.133 (1.263)		
At least one female judge					-0.657 (0.883)	-0.145 (2.024)
Observations	875	875	875	875	875	875
Year Fixed Effects	X	X	X	X	X	X
Industry Fixed Effects	X	X	X	X	X	X

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Fixed Effects Models of Individual Judge Scores

	(1) Judge Score	(2) Judge Score
Female Founder, Committee Round	4.494^{***}	4.343^{**}
	(0.997)	(2.175)
Female Founder, Paper Round	-0.218	-1.388
	(1.003)	(2.282)
STEM Degree, Committee Round	5.611^{***}	5.615^{***}
	(1.052)	(1.053)
STEM Degree, Paper Round	0.310	0.334
	(1.032)	(1.033)
Grad Degree, Committee Round	9.753^{***}	9.754^{***}
	(1.277)	(1.277)
Grad Degree, Paper Round	2.913^{**}	2.891^{**}
	(1.233)	(1.234)
Elite Degree, Committee Round	6.664^{***}	6.662^{***}
	(1.133)	(1.134)
Elite Degree, Paper Round	2.189^{**}	2.175^{**}
	(1.103)	(1.103)
MBA Degree, Committee Round	-0.611	-0.613
	(1.332)	(1.332)
MBA Degree, Paper Round	-0.907	-0.882
	(1.320)	(1.321)
Serial Entrepreneur, Committee Round	0.792 (1.092)	0.790 (1.093)
Serial Entrepreneur, Paper Round	-0.021	-0.038
	(1.060)	(1.060)
Female Founder X Male Judge, Committee Round		0.183 (2.401)
Female Founder X Male Judge, Paper Round		1.425 (2.494)
Observations	8920	8920
Year Fixed Effects	X	X
Industry Fixed Effects	X	X
Judge Fixed Effects	X	X
Round Fixed Effects	X	X

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Models of Score Difference across Evaluation Modes

	(1) Score Difference	(2) Score Difference	(3) Score Difference	(4) Score Difference	(5) Score Difference
Female Founder	5.895^{***}	5.855^{***}	5.979^{***}	5.892^{***}	5.970^{***}
	(2.015)	(2.015)	(2.021)	(2.016)	(2.021)
Serial Entrepreneur	3.396	3.363	3.390	3.424	3.347
	(2.211)	(2.211)	(2.212)	(2.214)	(2.212)
Elite Degree	6.642^{***}	6.606^{***}	6.714^{***}	6.599^{***}	6.704^{***}
	(2.200)	(2.200)	(2.204)	(2.204)	(2.204)
MBA Degree	4.567^{**}	4.570^{**}	4.554^{**}	4.560^{**}	4.553^{**}
	(2.263)	(2.263)	(2.264)	(2.265)	(2.264)
STEM Degree	7.797^{***}	7.889^{***}	7.742^{***}	7.835^{***}	7.825^{***}
	(1.961)	(1.963)	(1.964)	(1.965)	(1.965)
# Judges		1.330			1.556
		(1.305)			(1.336)
# Female Judges			-0.719		-1.043
			(1.281)		(1.311)
At least one female judge				0.755	
				(2.085)	
Observations	875	875	875	875	875
Year Fixed Effects	X	X	X	X	X
Industry Fixed Effects	X	X	X	X	X

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Exploring the Potential Channels and Performance Implications

	(1)	(2)	(3)	(4)	(5)	(6)
	Judge Score	Judge Score	Judge Score	Judge Score	Judge Score	Judge Score
Leave Out Mean, Paper Round	0.140***	-0.159		0.091***	0.118***	
	(0.027)	(0.143)		(0.032)	(0.027)	
Leave Out Mean, Committee Round	0.806***	0.737***		0.787***	0.780***	
	(0.014)	(0.078)		(0.017)	(0.015)	
# Judges, Paper Round		-4.713***				
		(1.881)				
# Judges, Committee Round		-0.390				
		(0.862)				
Leave Out Mean X # Judges, Paper Round		0.062**				
		(0.029)				
Leave Out Mean X # Judges, Committee Round		0.014				
		(0.016)				
Ex Post Funding, Paper Round			0.239***	-0.443		0.195***
			(0.067)	(0.298)		(0.067)
Ex Post Funding, Committee Round			1.271***	0.254		1.006***
			(0.067)	(0.162)		(0.068)
Leave Out Mean X Ex Post Funding, Paper Round				0.010**		
				(0.004)		
Leave Out Mean X Ex Post Funding, Committee Round				0.000		
				(0.003)		
Female Founder, Committee Round					3.486*	4.124*
					(1.883)	(2.143)
Female Founder, Paper Round					-1.393	-1.312
					(1.975)	(2.248)
STEM Degree, Committee Round					0.988	4.413***
					(0.916)	(1.041)
STEM Degree, Paper Round					0.412	0.125
					(0.894)	(1.020)
Grad Degree, Committee Round					1.915*	7.981***
					(1.115)	(1.264)
Grad Degree, Paper Round					2.568**	2.566**
					(1.072)	(1.220)

Elite Degree, Committee Round					1.629*	6.687***
					(0.985)	(1.116)
Elite Degree, Paper Round					1.956**	2.205**
					(0.956)	(1.087)
MBA Degree, Committee Round					-0.115	0.103
					(1.154)	(1.315)
MBA Degree, Paper Round					-1.013	-0.735
					(1.145)	(1.303)
Serial Entrepreneur, Committee Round					0.325	-0.650
					(0.948)	(1.084)
Serial Entrepreneur, Paper Round					-0.139	-0.197
					(0.919)	(1.053)
Female Founder X Male Judge, Committee Round					-2.852	-0.796
					(2.078)	(2.365)
Female Founder X Male Judge, Paper Round					1.668	1.300
					(2.158)	(2.457)
Constant	11.390***	58.340***	47.739***	11.464***	11.166***	42.317***
	(0.975)	(1.917)	(0.824)	(1.038)	(0.982)	(0.869)
Observations	8920	8920	8920	8920	8920	8920
Year Fixed Effects	X	X	X	X	X	X
Industry Fixed Effects	X	X	X	X	X	X
Judge Fixed Effects	X	X	X	X	X	X
Round Fixed Effects	X	X	X	X	X	X

Standard errors in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendices

Appendix A: Robustness of results to Judge X Round Fixed Effects

In Table 4 of the main results, we examine the impact of Founder characteristics on the evaluation behavior of Judges. In those models, We use Judge fixed effects, but it is possible that the results turn on differences of the judges in mean evaluations across the rounds. To address this concern, we run the same regressions below, but with Judge X Round fixed effects instead of separate Judge and Round fixed effects. The table below shows no substantial differences between the parameter estimates in these models and those in Table 4.

Table A-1:

	(1) Judge Score	(2) Judge Score
Female Founder, Committee Round	4.483*** (1.011)	4.951** (2.286)
Female Founder, Paper Round	-0.032 (1.022)	-2.492 (2.408)
STEM Degree, Committee Round	6.140*** (1.065)	6.131*** (1.066)
STEM Degree, Paper Round	0.382 (1.048)	0.433 (1.049)
Grad Degree, Committee Round	9.503*** (1.300)	9.511*** (1.300)
Grad Degree, Paper Round	3.028** (1.253)	2.982** (1.254)
Elite Degree, Committee Round	6.988*** (1.147)	6.987*** (1.147)
Elite Degree, Paper Round	2.122* (1.117)	2.100* (1.118)
MBA Degree, Committee Round	-0.394 (1.351)	-0.396 (1.351)
MBA Degree, Paper Round	-1.083 (1.343)	-1.034 (1.344)
Serial Entrepreneur, Committee Round	0.790 (1.108)	0.802 (1.110)
Serial Entrepreneur, Paper Round	-0.073 (1.077)	-0.105 (1.078)
Female Founder X Male Judge, Committee Round		-0.579 (2.537)
Female Founder X Male Judge, Paper Round		2.983 (2.645)
Observations	8920	8920

Year Fixed Effects	X	X
Industry Fixed Effects	X	X
Judge Fixed Effects	X	X

Appendix B: Robustness of Performance evaluation to admission

In this Appendix, we evaluate the sensitivity of our analysis of the performance implications across the evaluation modes to the inclusion of both admitted and unadmitted firms in our sample. Below, we run the same analysis on the sample of unadmitted firms because these are the firms least likely to suffer from contamination from any Matthew effects that might arise from the evaluation by members of the MassChallenge community. Below, we find measurements of the differences across evaluation regimes that support the qualitative findings of Table 7, if not the magnitude of the result. We find that the differences across the rounds is statistically significant (F = 12.61) supporting our contention that the committee structure is more sensitive to selecting high quality teams, even if those teams are not selected for admission to MassChallenge.

Table B-1: The impact of Ex Post Quality on Judge Score, by Round

	(1)
	Judge Score
Ex Post Funding, Paper Round	-0.053
	(0.112)
Ex Post Funding, Committee Round	0.530***
	(0.120)
Constant	66.288***
	(0.907)
Observations	5097
Year Fixed Effects	X
Industry Fixed Effects	X
Judge Fixed Effects	X
Round Fixed Effects	X