Disagreement Predicts Startup Success: Evidence from Venture Competitions

Luca Gius
MIT Sloan

August 16, 2024

Abstract

I present a new, compelling finding: the more venture competition judges disagree about the quality of a startup, the more likely the startup is to succeed. To explain why, I build on the notion that (i) entrepreneurs pursue opportunities based on their subjective beliefs (ii) common opinion cannot be a source of competitive advantage. Therefore, value is disproportionately created and captured by founders with atypical ideas that spark disagreement, and potential investors should harness disagreement as a predictor of success. I leverage data from 67 venture competitions to show that the empirical implications of this theoretical framework are supported by the data, whereas alternative explanations (e.g., that judges disagree more about risky ventures) are not. Additionally, I provide insights into what evaluators tend to disagree more often (e.g., former entrepreneurs) and which aspects of a startup (e.g., business model) are most polarizing. This work has broad implications for investors and institutions that strive to evaluate the potential of startup ideas.

Correspondence to lgius@mit.edu. I am especially grateful to Pierre Azoulay, Fiona Murray and Scott Stern for their invaluable feedback and advice. I thank Wes Greenblat, Soomi Kim, Danielle Li, Christina Nguyen, Matthew Rhodes-Kropf, Erin Scott, Ezra Zuckerman and various seminar participants for helpful comments and suggestions. I thank Jorge Guzman for sharing code and data to calculate founding strategic differentiation. Finally, I thank Adam Rentschler of Valid Evaluation for providing the data and critical insights on the setting. All errors remain my own.

I. Introduction

The history of entrepreneurship is littered with successful startups that initially baffled investors with their atypical value propositions. Airbnb is a notable example, initially dismissed by many as outlandish (Gallagher 2017). Fred Wilson, a venture capitalist who declined to invest, famously remarked, “we couldn’t wrap our heads around air mattresses on the living room floors as the next hotel room.” This phenomenon is known as the “uniqueness paradox” (Litov, Moreton, and Zenger 2012; Benner and Zenger 2016): unique strategies are crucial for sustained competitive advantage (J. Barney 1991), yet they often appear unfamiliar and challenging for investors to evaluate. As such, atypical ideas might be rejected or face an "illegitimacy discount" (Zuckerman 1999) compared to more conventional, but less valuable, strategies.

How can investors recognize the potential of unique ideas like Airbnb? An email exchange between Wilson and Paul Graham, founder of Y Combinator, offers a clue. While Wilson and his senior colleagues were skeptical, Graham and Wilson’s junior team members were enthusiastic. In other words, Airbnb did not elicit universally lukewarm reactions; instead, the idea was contentious. This paper argues and provides evidence that the Airbnb case is generalizable: disagreement among evaluators predicts a startup’s future success and can help investors navigate the uniqueness paradox.

Why does disagreement predict future startup success? To explain this, I build on the notion that entrepreneurs pursue opportunities they are relatively optimistic about (Agrawal et al. 2024) based on subjective theories of value (Felin and Zenger 2009; 2017). As such, some startup ideas will be uncontroversial; others, particularly the most distinctive, will spark disagreement among evaluators – some championing the idea (e.g., Paul Graham) and others dismissing it (e.g., Wilson). Crucially, however, common or readily accepted theories cannot be a source of competitive advantage (Felin and Zenger 2009; 2017): had there been no skepticism around Airbnb, the hotel industry’s competitive response would likely have been swift. As such, disagreement can be leveraged as a signal of future potential, particularly for unique ideas that are usually both more polarizing and harder to evaluate.

To quantify disagreement, I leverage data from 67 startup competitions and 2650 startups. Each venture is scored by at least three judges using a standardized scoring rubric. Because each judge grades individually, the discrepancy in the judges' scores (i.e., the standard deviation of the scores) provides a measure of how polarizing a startup's value proposition is. This allows me to investigate when the judges are more likely to disagree, what startups are more polarizing, and whether polarizing startups are ultimately more successful.

First, I show that the Airbnb case holds more generally: disagreement among judges predicts a startup's future success. I show that polarizing startups raise significantly more future funding, generate higher revenues, and are more likely to achieve a successful exit than uncontroversial firms. This finding holds even when controlling for the average grade, indicating that the judges do not fully recognize the potential of polarizing startups, since disagreement predicts success even among startups considered to be equally valuable. Consistent with the theoretical framework, I find that startups with distinctive descriptions are (i) more controversial (ii) those for which disagreement is a particularly strong predictor of success. Importantly, polarizing startups are not more likely to fail; thus, disagreement does not reflect the riskiness of the business.

Next, I investigate the sources of disagreement. First, I show that discrepancies in scores in part reflect lack of clarity around a startup's value proposition: the scores converge in later competition rounds after the startups update their first-round application materials based on the judges' feedback. Second, disagreement depends on the composition of the judging panel. Most notably, former founders are significantly more likely to "stand out" and disagree with the other experts. Third, I show that disagreement is also disproportionately higher around certain aspects of a startup idea. Leveraging the fine-grained rubric that the judges score on, I show that score dispersion is highest when evaluating a startup's business model (e.g., its scalability, the potential to create downstream value, or its pricing strategy), whereas the judges are more likely to agree on the quality of the team.

The first contribution of this paper is methodological. The notion that different people can agree to disagree on an entrepreneurial opportunity is central to both the theory-based view of the firm (Felin and Zenger 2009; 2017) and an emerging literature on Bayesian entrepreneurship (Agrawal et al. 2024), but to my knowledge this is the first paper to concretely quantify how much disagreement is sparked by a business idea. I also show that disagreement can serve as a marker of uniqueness. As such, startups that are more distinct in specific dimensions (e.g., the airbed concept in the Airbnb case) will generate more disagreement in those areas. This approach can help measure and investigate (optimal) distinctiveness across a broader range of strategic dimensions than those typically studied by strategy scholars (Zhao et al. 2017).

Most importantly, this work offers a potential remedy to the "uniqueness paradox," the notion that unique ideas are discounted because they are difficult to evaluate. While this literature has mostly concentrated on publicly traded firms, I focus on the evaluation of entrepreneurial opportunities. I theorize and test the notion that valuable entrepreneurial theories must be polarizing, meaning that they are espoused by the entrepreneur and few others, but not by everyone else. As such, investors should pay particular attention to unique ideas that spark disagreement. More broadly, the lesson that disagreement is a necessary and measurable byproduct of unique and valuable entrepreneurial theories has important implications for designing decision-making processes in start-up competitions and accelerators (Gonzalez-Uribe and Leatherbee 2018; Fehder and Murray 2018; S. Cohen et al. 2019), venture capital firms (Malenko et al. 2023) and other institutions where judges, analysts or investors jointly evaluate start-ups (Csaszar and Eggers 2013; Scott, Shu, and Lubynsky 2020). We know from previous studies that scores assigned by venture competition judges have important signaling and feedback value for startups (Howell 2020; 2021). The fact that judges often vocally disagree with each other does not diminish their importance; on the contrary, I show that such disagreements are useful predictors of future startup success. Entrepreneurs and organizations should leverage this insight to improve the idea selection process, which is inherently challenging (Gompers and Lerner 2001; Kerr, Nanda, and Rhodes-Kropf 2014; Nanda and Rhodes-Kropf

2016; Ewens, Nanda, and Rhodes-Kropf 2018). To conclude, the results should further motivate organizations to understand and carefully select the right aggregation mechanisms while picking ideas (Sah and Stiglitz 1988; Knudsen and Levinthal 2007; Christensen and Knudsen 2010; Csaszar and Eggers 2013), because mechanisms that require consensus will miss good opportunities.

The paper is structured as follows: in Section 2, I present a new, puzzling fact – the more venture competition judges disagree on a startup, the more likely it is to succeed. I explain this nexus between disagreement and success through a theoretical framework based on the theory-based view of the firm and the Bayesian entrepreneurship literature, proposing tests to distinguish it from alternative explanations. In Section 3, I explain the data and setting. I also provide more context around the main variables and lay down the empirical framework. In Section 4, I showcase the results and conclude with the strategic implications in Section 5.

II. Theoretical framework

Entrepreneurs and managers face a key trade-off when choosing a strategy. On the one hand, familiar strategies that everyone understands cannot lead to a sustainable advantage because competitors would quickly imitate them. On the other hand, investors might discount valuable, but unfamiliar strategies because they are hard to evaluate. This tension, known as the "uniqueness paradox" (Litov, Moreton, and Zenger 2012; Benner and Zenger 2016), tempts managers to choose less valuable but more familiar and easier-to-evaluate strategies.

At the heart of the uniqueness paradox is a prediction problem: investors lack the information needed to accurately assess unfamiliar ideas. This arises because the information is confidential, the idea is costly to evaluate[1], or the entrepreneur believes but cannot prove that the idea is valuable. Consequently, investors struggle to distinguish between unfamiliar ideas that are valuable and those that are mediocre, a classic case

[1] For example, (Litov, Moreton, and Zenger 2012) show that public firms with unique corporate strategies are less likely to be covered by analysts – this lack of coverage exacerbates information issues faced by investors and depresses the stock price relative to its true potential.

of adverse selection (Akerlof 1970). To assuage this problem, (Benner and Zenger 2016) suggest that publicly traded firms should clearly communicate their strategy to potential backers (e.g., through conference calls with analysts) and seek sophisticated investors who can evaluate long-term value (e.g., long-term investors and private equity). Instead, I focus on the evaluation of entrepreneurial opportunities, and propose a new solution based on the following puzzling fact: the future success of a startup idea correlates with the disagreement it causes among venture competition judges. As such, investors should use disagreement as an indicator of future success, particularly for unique and unfamiliar startup ideas for which few reliable indicators are available.

Motivating fact: Disagreement among venture competition judges predicts future startup success.

What explains this surprising result? A lack of consensus around an investment opportunity should, in principle, raise concerns. A "hang jury" suggests a potentially flawed evaluation: some jurors might have idiosyncratic preferences or biases, less expertise, or put less effort in their evaluation. But if disagreement connotes nothing more than "background noise" and an ineffective evaluation, it should be less likely – not more – to predict success.

However, not all disagreement is simple noise. A longstanding literature in strategy and entrepreneurship has recognized that fundamental differences in opinion about the value of an opportunity are not only possible – they fuel entrepreneurship (Hayek 1945; J. B. Barney 1986; Van den Steen 2004). The theory-based view of the firm and the Bayesian entrepreneurship literature propose that entrepreneurs pursue opportunities based on their subjective beliefs (or "theories") on how to create and capture value (Felin and Zenger 2009; 2017; Agrawal et al. 2024). Subjective entrepreneurial beliefs need not be universally accepted: entrepreneurs are more confident about the opportunity compared to other people. Outsiders whose prior beliefs align with the entrepreneur will support the theory, while those preferring alternative theories will reject it. This "strategic disagreement" (Klepper and Thompson 2010) – or agreement to disagree – can occur even when evaluators share the same information, ability, knowledge, or attention

(Aumann 1976; Morris 1995). It arises due to discrepancies in their prior beliefs, especially when the judgment cannot be based on objective data.

Controversy might dissuade potential investors but has an important silver lining: a path to a sustainable competitive advantage. Commonly held views cannot yield new insights on arbitraging, recombining, or redeploying resources more effectively (Felin and Zenger 2017; Wuebker, Zenger, and Felin 2023; Agrawal et al. 2024). Similarly, novel strategies that are quickly accepted may be valuable short-term but are easily replicated and unlikely to be valuable long-term. Thus, a view must have both champions and detractors to generate entrepreneurial rents[2]: it needs to spark strategic disagreement. As such, this theoretical framework can explain why disagreement predicts success.

My measure of disagreement – i.e., dispersion in venture competition scores – will comprise both "background noise" (e.g., lazy or uninformed dissent) and strategic disagreement, making it hard to separate one from the other. However, while all startup propositions can occasionally spark dissent due to "background noise," only the most unique and distinct propositions can spark strategic disagreement. As such, the framework offers two other testable implications that, as I will show, are supported by the data. First, disagreement should be higher for more unique startups. Second disagreement should predict success only for unique startups, since background noise (the only source of dissent for familiar ideas) cannot predict success.

Supporting fact # 1: Unique startups experience higher levels of disagreement.

Supporting fact # 2: Disagreement predicts success only for startups with unique value propositions.

Alternative theories could also, in principle, explain why disagreement predicts venture success. One possibility is that disagreement serves as a proxy for venture risk (Howell 2021). According to this view,

[2] Familiar strategies are not the only example of ideas that spark little disagreement and fail to create a sustainable advantage. Extremely contrarian ideas that everyone believes to be wrong also fall into this category. Entrepreneurs need at least some champions to attract the necessary resources, and without them, even the most groundbreaking ideas can struggle to gain traction.

more polarizing startups should be both more likely to succeed and more likely to fail, and less likely to simply survive as viable businesses. However, I will demonstrate that this hypothesis is not supported by the data.

Supporting fact # 3: Disagreement does not predict startup failure.

Second, (Malenko et al. 2023) show that early-stage VCs often use a ‘champions’ voting rule: all it takes is one supporting partner to make an early-stage investment. They suggest that the purpose of the rule is to catch ‘asymmetric’ outliers, i.e., startups that are especially strong along some dimension and weak on others; under some assumptions, they argue that outliers are on average both more likely to succeed and unlikely to be selected by consensus-based voting rules. I will show that disagreement predicts success independently on whether the polarizing startup is an ‘asymmetric outlier’ or not.

Supporting fact # 4: Disagreement predicts success even after controlling for ‘asymmetric outliers.’

The link between disagreement and the potential of unique theories to generate entrepreneurial rents has two practical implications for investors, entrepreneurs, and strategy researchers. First, as discussed, it can help assuage the uniqueness paradox. I will show that disagreement predicts success even for startup ideas with the same average grade, indicating that judges, on average, discount valuable but polarizing ideas by not considering the disagreement signal. Second, strategy scholars have been historically interested in understanding the role and importance of strategic differentiation (J. Barney 1991; Porter 1996; Zhao et al. 2017). However, measurement issues have forced most researchers to focus on the distinctiveness of a few selected aspects of otherwise multifaceted strategies: what metric, for example, could possibly capture the uniqueness of Airbnb’s scaling strategy relatively to HomeAway or the hotel industry? When reasonable metrics of uniqueness are unavailable, my theoretical framework suggests instead to use disagreement as a proxy. For example, in my data, I observe how much the judges disagree around specific aspects of a startup’s strategy (including, e.g., scaling and pricing) whose uniqueness is otherwise hard to pin down.

III. Data and Methods

3.1 Setting: Venture Competitions

This article examines venture competitions, where entrepreneurs and startup founders pitch their products and business plans to a panel of judges. Participants usually compete through multiple rounds to win monetary prizes, funding, or a spot in an accelerator. These competitions also offer founders networking opportunities, visibility, and valuable feedback on their strategies. The businesses pitched at these competitions are high-technology, high-growth startups. Judges include former founders, angel and VC investors, and industry experts. Each judge assigns a quantitative score between 1 and 7 to each startup, using a consistent, detailed grading rubric.

Through Valid Evaluation, a private company assisting competition organizers, I gained confidential access to data on 67 competitions and 118 rounds. Part of this data was previously used in (Howell 2020; 2021) to demonstrate the importance of feedback and how winning a challenge increases the likelihood of securing VC funding. The complete list of competitions is in the Online Appendix (Table 1A). Most of the competitions are based in Arizona, with approximately 50% of the sample coming from the Arizona Innovation Challenge (2012-2019), which awarded winners up to $150,000 and access to an accelerator. Thus, most startups in my database were judged using a highly consistent rubric.

Overall, the data includes information on how 1054 judges scored 2650 firms on 24 different dimensions on average. I observe 619 of the businesses across multiple competitions, and match 1060 firms to Crunchbase and 742 to Pitchbook. Additionally, I successfully matched 875 judges to their LinkedIn profiles. The descriptive statistics are in Table 1.

Most competitions involve multiple rounds. Typically, the first round consists of a written application that is graded by multiple judges individually. Selected startups advance to later rounds, and most of the final rounds are 5-15 minute pitches followed by questions and then deliberation. After excluding all rounds (typically with live pitches) where the judges do not grade individually, the calculated median competition has two rounds.

I exclude all observations where a startup is scored by fewer than three judges. In the remaining data, the median startup is scored by five judges per round. All judges share the same information and application materials about the start-ups they evaluate. The judges involved are typically entrepreneurs, angels, VC investors and experts from the business community. Most judges are male (75%). 34% are former founders, and 35% are startup investors. About a third have an MBA and 11% have a PhD. They are volunteers without conflicts of interest in the competing startups: anecdotally, they are primarily motivated by the opportunity to contribute to the local startup ecosystem, and almost never invest in the startups they judge (Howell 2021).

The 2650 startups in the sample are highly technological, young, but not fledgling: the median startup is three years old at competition time, and 17% held a patent before the competition. The Arizona Innovation Challenge, for example, explicitly requires applicants to be “moving towards commercialization” of an innovative technology. Moreover, it requires companies to fit into specific sectors (Advanced Materials & Manufacturing, Aerospace / Defense, Cleantech/Renewable Energy, IT Hardware, and IT Software) as reflected in the data (Table 2A, Online Appendix). Still, these startups’ industries broadly represent the U.S. ecosystem with one exception: due to the AIC requirements and the presence of a few cleantech-focused competitions, the sample skews towards cleantech and renewable energy (18% of the startups). I collect outcome data for the 1060 firms matched with Crunchbase and the 742 matched with Pitchbook. As of 2023, 20% of the matched startups are out of business, 10% have been acquired, and the remaining are still operating. Pitchbook forecasts that the average business in the sample has a 40% chance of a successful exit (either through acquisition or IPO). Additionally, of those with funding and revenue data, 62% have more than $1 million in yearly revenues, and they have raised an average of $8.4 million in funding.

3.2 Main variables

All variables used in the analysis are described in Table 1, along with their descriptive statistics. In this section, I delve deeper into the three primary measures at the core of this study: disagreement, uniqueness, and startup success.

3.2.1 Measuring disagreement

One way to measure polarization would be to compare how much value the entrepreneur sees in the startup idea relative to potential investors, competitors, and clients. Unfortunately, I do not observe the founder's self-assessment in my setting. The Airbnb example, however, suggests a way forward: we can recognize Airbnb as polarizing not only because Brian Chesky's view contrasted sharply with that of the hotel industry incumbents. Rather, a stark disagreement can also arise between external evaluators, some of whom detest the idea, and others who champion it – in the same way as Paul Graham and Fred Wilson disagreed on the viability of Airbnb. This approach allows me to bypass the issue of measuring an entrepreneur's optimism relative to others, focusing instead on the extent of disagreement among external evaluators about the business's value proposition.

An advantage of this approach is that the measure can be constructed by any group of potential investors, making my findings particularly actionable. Additionally, it allows me to compare more and less polarizing startups that have been assigned the same average grade, thereby controlling for perceived potential even if I don't observe all pitches or startup characteristics that the judges do. A potential downside is that the method fails to flag extremely contrarian ideas rejected by all judges as polarizing. However, extremely contrarian ideas and flawed ideas are indistinguishable to potential investors by definition: no reasonable polarization measure could help tease them apart.

In practice, for every startup I calculate each judge's average score (across dimensions and rounds). Then, for each startup, I measure the standard deviation of the judges' average scores, provided that at least three judges score the startup.

3.2.2 Measuring uniqueness

Previous work has focused on proxies of strategy uniqueness that either highlight unusual combinations of industries (Litov, Moreton, and Zenger 2012), or distinctive descriptions as found in patents (Verhoeven, Bakker, and Veugelers 2016) and websites (Guzman and Li 2023). Measures of industry and patent uniqueness are not applicable in my setting. Many of the competitions I analyze cater to specific industries (e.g, cleantech or software), making competition and industry fixed effects collinear with a variable capturing industry combinations. Additionally, only a minority of the startups in my sample have applied for or have been issued a patent.

The measure that I rely on is most closely related to (Guzman and Li 2023), who define a startup to have a distinct strategy based on their website description at founding time. A great upside of their measure is that it is based on public information, i.e. the "About Us" section of a startup's website scraped through the wayback machine (https://web.archive.org/). Unfortunately – as acknowledged by the authors – this also makes the measure somewhat noisy, since the "About Us" section often includes non-strategic information (e.g., mailing address or team background), invalid text or error messages. In their original paper, the authors have data on 12,000 startups, a sample size big enough to establish a positive relationship between success and uniqueness significant at the 5% to 10% value. However, only 117 startups from their sample are also in my dataset; the measure remains too noisy even after reconstructing the measure for 965 additional startups for which I have website information.

Instead, I use the textual descriptions submitted by 1476 startups as part of their competition application package. Unlike the website descriptions, these snippets are written specifically to describe to the judges, in a few sentences, what the startup does, who the customers are, and the overall strategy. Consequently, the resulting (confidential) textual measure of distinctness is much more precise than the (public) measure based on (Guzman and Li 2023), although the two are strongly correlated.

The methodology is virtually identical. First, I obtain the word embeddings of each description. Instead of relying on doc2vec embeddings as in (Guzman and Li 2023), I use state-of-the-art transformer embeddings obtained through OpenAI's API. Then, I compute the maximum similarity $maxsim_i$ of each description relative to every other startup description in my sample. My uniqueness measure is $1 - maxsim_i$.

3.2.3 Measuring success

To conclude, I take a broad view of what constitutes success for a startup. I rely on four different indicators, all based on outcomes[3] as of 2023. First, I measure how much funding the startup raised, indicating how successful the venture was in attracting resources (sources: Pitchbook and Crunchbase). Second, I look at whether the startup has at least $1 million in yearly revenues, a proxy for market traction (source: Crunchbase). Third, I consider the rank assigned by Crunchbase to gauge the venture's prominence relative to its peers[4]. Fourth, I look at the likelihood of exit (either via IPO or acquisition) as forecasted by Pitchbook[5].

3.3 Empirical design

Most results in this study are based on a series of regressions at the startup-competition level.

First, I show that higher disagreement predicts success $Y_{s,c}$ (e.g., amount of funding raised): $\hat{\beta}_1 > 0$. $Disagreement_{s,c}$ is the standard deviation of scores assigned to startup *s* in competition *c*. The controls $Z_{s,c}$ include competition fixed effects and startup characteristics (including the average score received).

$$Y_{s,c} = \beta_o + \beta_1 Disagreement_{s,c} + \beta_2 Z_{s,c} + \varepsilon_{s,c}$$

[3] To minimize truncation concerns, in all specifications where I use success indicators, I include founding year fixed effects and exclude startups judged later than 2019.
[4] https://about.crunchbase.com/blog/crunchbase-rank-trend-score/. I normalize the rank in percentiles for easier interpretation.
[5] https://pitchbook.com/media/press-releases/pitchbook-predicts-vc-backed-exits. The score goes from 0 to 100. Additionally, startups that successfully IPO'd or were acquired as of 2023 are assigned a score of 100.

I use the same empirical design to show that $Disagreement_{s,c}$ predicts success especially well for unique ideas (supporting fact # 2), that $Disagreement_{s,c}$ does not predict failure (supporting fact # 3) and that estimates of β_1 are robust to controlling for ‘asymmetric outliers’ (supporting fact # 4).

A second set of regressions investigates what startup and competition covariates $X_{s,c}$ explain disagreement. This allows me to show that unique ideas are more likely to spark disagreement (supporting fact # 1).

$$Disagreement_{s,c} = \alpha_o + \alpha_1 X_{s,c} + \alpha_2 Z_{s,c} + \varepsilon_{s,c}$$

The latter specification also sheds light on other, descriptive facts about the nature of disagreement, e.g., how it depends on a startup’s industry or whether the venture has obtained a patent. In some specifications I will study disagreement at a more refined unit of analysis, for example to explore what specific dimension the judges are more likely to disagree about, or whether disagreement decreases in later stages of the competition. In these cases, I will focus on $Disagreement_{s,c,i}$ where i denotes that disagreement is calculated as a std. deviation of scores within a specific round (e.g., first vs. final round) or for a specific dimension (e.g., product, team or business model). In all these regressions I cluster the standard errors by startup since scores assigned to a same venture in different rounds and in different dimensions are strongly correlated.

Finally, I will focus on what judges are more likely to express dissent. To do so, I take the absolute difference of the score given by a judge from the average score given by the other judges to the same startup in the same round: a higher absolute difference Abs. $\text{difference}_{s,r,j}$ denotes higher disagreement with the other judges. Then I analyze the relationship between Abs. $\text{difference}_{s,r,j}$ and judge characteristics X_j while controlling for round and startup fixed effects δ_r, δ_s:

$$\text{Abs. difference}_{s,r,j} = \delta_o + \delta_1 X_j + \delta_r + \delta_s + \varepsilon_{s,r,j}$$

IV. Results

4.1 The extent of disagreement

A key premise of this work is that venture competition judges do not always agree on the value of a startup opportunity: in other words, I expect the dispersion of scores assigned to each startup to be "high." By what standard? One way to benchmark the within-startup dispersion in scores is to compare it to the overall variance in scores: greater disagreement results in a higher "within sum of squares" as a proportion of the "total sum of squares."

$$\underbrace{\sum_{i=1}^{N}(Grade_i - \overline{Grade})^2}_{Total\ sum\ of\ squares\ (100\%)} = \underbrace{\sum_{j=1}^{K} n_j\,(\overline{Grade}_j - \overline{Grade})^2}_{Between\ sum\ of\ squares\ (38\%)} + \underbrace{\sum_{j=1}^{K}\sum_{s=1}^{n_j}(Grade_{sj} - \overline{Grade}_j)^2}_{Within\ sum\ of\ squares\ (62\%)}$$

I perform this variance decomposition for the $K = 2{,}650$ startups graded by at least 3 distinct judges ($n_j \geq 3$), totaling $N = 21{,}850$ judge-startup observations. The "within sum of squares" represents the deviation of each judge's score from the venture's average grade. Without disagreement, it would be 0. Conversely, the "between sum of squares" represents the deviation of each startup's score from the overall mean. Higher dispersion suggests that judges clearly differentiate between high-quality and low-quality startups. If judges evaluated all startups as having the same quality, it would be 0. I find that the "within sum of squares" accounts for 62% of all variation, indicating that most score variation reflects disagreement among judges about the same startup, rather than differences between startups.

To further illustrate, I calculate Cohen's kappa, a traditional metric for assessing inter-rater agreement (J. Cohen 1960)[6]. On a scale going from 0 to 1, I calculate a coefficient of 0.13 (95% C.I. [0.114, 0.146]) which is considered a very high level of disagreement. For example, in the medical literature "any kappa below 0.60 indicates inadequate agreement among the raters" (McHugh 2012).

[6] This metric is usually assessed for categorical variables: as such, I round up each judge's average rating to the closest integer as reflected in the grading rubric.

4.2 Motivating fact: disagreement correlates with future success

In this section, I present the key motivating fact of this paper: disagreement among venture competition judges predicts a startup's future success. I regress various proxies for future startup success on the standard deviation of the judges' scores: the purpose is to illustrate the predictive power of disagreement, not to identify a causal relationship. In table 2, I show that a 20% increase in disagreement relative to its average is associated with one-third more future funding, a 0.8 percentile higher Crunchbase ranking, a 2% higher exit probability (as forecasted by Pitchbook) and a 2.5% higher likelihood to have at least $1M yearly revenues by 2023. These results are robust to controls for competition, industry, founding year fixed effects, and whether the startup was granted a patent before the competition.

All regressions include the startup's average grade as a control. As such, I am effectively comparing more and less polarizing startups that are otherwise judged to be equally valuable and likely to succeed. There are two important reasons for this. First, as reflected by the fact that average grades strongly predict success[7], this allows me to control for important information observed by the judges (e.g., pitching quality) but not by the econometrician. More subtly, the fact that disagreement predicts success even when comparing startups with the same average grade shows that judges effectively discount polarizing startups and predict startup success less accurately than they could.

4.3 Supporting evidence

In this section I present four facts supporting the notion that the predictive power of disagreement stems from the fact that distinct ideas generating dissent are more likely to succeed.

[7] As highlighted in (Howell 2021) the positive correlation between grades and future outcomes reflects in part good judgement, but is also a self-fulfilling prophecy: the startups with the highest grades win the competition and receive awards and recognition, while losers can infer that they ought abandon the venture.

4.3.1 Unique, polarizing startups are especially likely to succeed

First, I show that distinct ideas play a central role in the disagreement-success nexus.

In table 3 I show that more distinct startups are more polarizing: an increase of a standard deviation in distinctiveness (i.e., textual dissimilarity relatively to other startups' descriptions) is associated with a 4-7% increase in the standard deviation of disagreement. This result is robust to controls for competition, industry, founding year and patent grant fixed effect. In the last column, I also include a measure of business uniqueness based on website information (Guzman and Li 2023): while the coefficient of differentiation remains positive and strongly significant, the business uniqueness' proxy is only noisily correlated with disagreement.

In Table 4, I show that the predictive power of disagreement for the likelihood of exit, yearly revenues, and Crunchbase rank disappears among the 50% least distinct ideas: the coefficient of "Std. dev. scores" is non-significant, while the interaction "Std. dev. scores # Above Median Diff." is strongly significant. This supports the notion that the standard deviation of scores includes both background noise (e.g., disagreement driven by inattention, mistakes, or misunderstandings by some judges) – unlikely to correlate with success – and strategic agreement to disagree, which most strongly predicts success. But only unique ideas spark agreement to disagree, whereas for the less distinct ideas, disagreement is likely mostly noise.

Interestingly, all polarizing startups – including the less differentiated ones – raise more funding on average. One possibility is that the 'champions voting rule' adopted by many early-stage VCs (Malenko et al. 2023) mechanically favors all polarizing startups, including those that spark disagreement because of background noise. This blanket 'champions voting rule' thus encourages subpar investments in undifferentiated but polarizing startups that are less likely to result in an IPO, acquisition, or high-revenue business.

4.3.2 Rejecting alternative explanations

In Table 5, I show that the predictive power of disagreement does not change even when controlling for asymmetric outliers. “Venture Grade Asymmetry” measures the standard deviation of average scores across dimensions, so that it is higher if the startup performed well on some dimensions (e.g., team) but not others (e.g., product).

To conclude, in column 5 of Table 5, I show that polarizing startups are not more likely to fail and close down. This contradicts the notion that disagreement captures venture risk (i.e., startups both more likely to be highly successful and to fail).

4.4 Descriptive evidence on the determinants of disagreement

In this section, I present descriptive evidence on the sources of disagreement, including (i) lack of clarity around the underlying proposition, (ii) judge characteristics, and (iii) startup characteristics.

4.4.1 Disagreement decreases after first-round feedback

In Table 6, I show that disagreement around the same startup decreases by around 20% in later rounds by exploiting the fact that some startups are judged across consecutive rounds in the same competition and with the same rubric.

Increasing agreement is unlikely driven by structural changes in the startups’ value propositions, given that rounds happen typically within one to two months of each other. Instead, one potential reason is enhanced clarity around the value proposition: between rounds, startups can submit new application materials to clarify aspects of their proposition based on judges’ feedback. Second, the judges might learn from other experts’ judgements and update their own accordingly. Since the effect remains the same when focusing on competitions where all judges vary from one round to another (column 3), the first explanation seems more plausible.

4.4.2 The role of judges

Next, I examine which experts "stick out" with consistently discrepant beliefs from the rest. In the Online Appendix (Figure 1A), I calculate each judge's "propensity to dissent," defined as the proportion of times a judge gave a "dissenting" grade (i.e., at least 2 grades out of 7 away from the other judges' average). The distribution of the "propensity to dissent" is characterized by a concentrated left tail, indicating a mass of judges unlikely to dissent with others, and a long right tail of "contrarians" who systematically disagree with the others.

Who are these "contrarians"? To answer this, I analyze which judge characteristics predict higher absolute deviation in scores relative to other experts (Table 7). Strikingly, former founders are the most likely to disagree with the rest of the panel. Their scores are 7% more divergent from the average scores of the other judges compared to non-founders. This aligns with the idea that individuals pursuing entrepreneurship are more likely to disagree with others on whether an opportunity is valuable. Education, on the other hand, does not appear to be a major factor in determining who disagrees the most.

A potential takeaway from these results is that "oversampling" advice from former entrepreneurs, which is less aligned with common wisdom, could foster diversity of perspectives.

4.4.3 Polarizing aspects of a startup idea

What aspects of a startup idea are most polarizing? To answer this question, I leverage the fact that judges assign individual grades to predetermined dimensions of a detailed grading rubric. I focus on the 90% of startups in my sample judged using an identical rubric consisting of nine dimensions and 24 subdimensions[8].

In table 8, I explore whether disagreement revolves primarily around team (e.g., is there any gap in personnel?), a startup's product (e.g., technology validation and intellectual property), industry

---

[8] The 2024 grading rubric for the Arizona Innovation Challenge, comprising most of the results for this section, is available at https://www.azcommerce.com/media/4hgptrxn/aic_rubric.pdf.

attractiveness (e.g., is there a big, growing market open to new entrants?), business model (e.g. how is value created and captured? Is the business scalable?), or market validation and analysis (e.g., customer engagement, partnership plans, the startup's overall market analysis). The dependent variable of the regression is the standard deviation of judges' scores assigned to a specific dimension (e.g., team) for each team in each competition round. I find that judges disagree the least about the quality of a startup's team but the most about the business model. These results align with investors disproportionately reacting to information about a firm's human capital (Bernstein, Korteweg, and Laws 2017), as different investors are more likely to agree on what constitutes a high-quality team. Conversely, a business plan represents the firm's subjective – hence more polarizing – view on how to organize its resources to create and capture value (Amit and Zott 2001). The magnitudes are economically significant: for example, disagreement about the business model is 6% higher than about the startup team. Specific areas of disagreement include business scalability, the potential to create downstream value, pricing, and the power of incumbents (see the Online Appendix, Figure 2A).

Table 9 shows that disagreement increases by 4-10% after a startup obtains its first patent, even after controlling for founding year, average grade, and industry. This suggests that judges disagree more when evaluating technical novelty. This result aligns with a long-standing literature on how entrepreneurial opportunities arising from technological innovations in R&D intensive sectors are hard to assess and surrounded by commercial uncertainty (Rosenberg 1982; Kline and Rosenberg 2009)[9]. It also further corroborates the notion that disagreement does not capture venture risk – if it did, a patent grant, which arguably derisks the startup, would decrease disagreement rather than increase it.

[9] Indeed, Table 3A and Figure 3A in the online appendix show that disagreement is highest in R&D intensive sectors including Advanced Manufacturing, Advanced Materials, and Aerospace/Defense, and is lower in Software.

V. Discussion and Conclusions

This paper presents and explores the consequences of a new, puzzling fact: disagreement among venture competition judges predicts a startup's success.

Using data from 62 venture competitions, I show that a 20% increase in the standard deviation of the judges' scores is associated with – among other outcomes – one-third more funding and a 2% higher exit probability. This result holds even after conditioning on the judges' scores, indicating that the judges – including seasoned investors – do not take this signal into account when evaluating a startup.

To explain these findings, I build on the notion that a valuable startup value proposition needs to spark some disagreement: common opinion cannot be a source of opportunity (Felin and Zenger 2017; Agrawal et al. 2024). I find this explanation to fit the data better than alternative possibilities, such as disagreement capturing risk.

Unique startups play a central role. Disagreement is driven by the distinctiveness (i.e., textual dissimilarity) of a startup's value proposition description relative to other competition participants. Moreover, the disagreement-success nexus disappears among the 50% least distinct startups.

Next, I focus on the determinants of disagreement. I find that disagreement is pervasive: levels of agreement five times higher would still be considered unacceptably low in the health research community (McHugh 2012). Differences in opinion partly stem from a lack of clarity around the startup's value proposition, which can be resolved through rounds of feedback and iteration. Disagreement also reflects differences in the judges' backgrounds, with former entrepreneurs disproportionately more likely to disagree with the other judges. Additionally, disagreement depends on specific aspects of the startup value proposition: it is higher around a startup's business model than its team.

5.1 Managerial Implications

This work has practical implications for potential investors and budding entrepreneurs seeking to understand the origins of valuable strategies. It validates (Thiel and Masters 2014)'s informal approach to assessing entrepreneurial theories ("what do you believe that no one else believes?"). I propose a slight twist: founders and investors should ask themselves "is the value proposition polarizing?" or "can I find two reasonable experts that agree to disagree on my theory?" An affirmative answer suggests that the startup idea is potentially valuable.

That being said, disagreement predicts success only for the most distinct ideas, which are especially hard to judge and potentially discounted (Litov, Moreton, and Zenger 2012; Benner and Zenger 2016). Experts can also disagree about familiar ideas due to biases, inattention, or errors of judgement: not all disagreement is a useful signal. Recent empirical evidence shows that many venture capital funds adopt an indiscriminate "champion-based" approach that greenlights an early-stage investment if anyone agrees with it (Malenko et al. 2023). This investment rule favors all polarizing ideas by design. While the results of this paper rationalize why consensus-based mechanisms should be avoided, they caution against a blanket 'champion-based' approach. Indeed, some of the results suggest that VCs might already be indiscriminately favoring polarizing ideas irrespective of whether they are distinct. This is ineffective, because disagreement is a useful marker for future success only for unique and unfamiliar startup ideas.

5.2 Limitations and Future Research

The primary limitation of this paper is that it focuses solely on measuring disagreement among experts. While this makes the measure particularly actionable for investors, it overlooks the interesting outlier case where a contrarian entrepreneur disagrees with all experts on the viability of a business idea. The nature of the data also prevents making causal claims about the correlation between disagreement, uniqueness, and success. Future work will need to tackle this challenge, perhaps in a controlled laboratory setting.

The new measure of startup polarization has several potential applications. First, it could contribute to a better understanding of the origins of spinouts. Past work (Klepper and Sleeper 2005; Klepper 2007; Klepper and Thompson 2010) has documented how spinouts are often rooted in strategic disagreement between parent companies and key employees, perhaps prompted by a change in ownership (Kim 2022). However, measurements of strategic disagreement have been limited to "mini-case studies" involving painstaking historical analysis. My proposed measure of polarization allows studying this phenomenon – as well as the thematically related literature on contrarian investing in startups (Wu 2016) – at a larger scale.

Other work could leverage the polarization measure to explore how disagreement influences a startup's strategic options. As suggested by (Felin and Zenger 2017), a polarizing startup and an incumbent likely disagree on what resources are valuable and necessary to control (Barney 1986; Van den Steen 2010). Discrepant beliefs between a startup and an incumbent might prevent cooperation (Marx and Hsu 2015) and acquisitions, or could result in transactions that are disproportionately beneficial for one party and costly for the other (Bryan, Ryall, and Schipper 2022).

In conclusion, empirical research on entrepreneurial theorizing and contrarian entrepreneurship is in its early stages. Many questions remain about how heterogeneity and value originate from beliefs, the role of contrarians in driving economic progress, and the best strategies in an environment where reasonable people agree to disagree. My measure of startup polarization represents one way to capture "strategic disagreement," offering a potential path forward.

References

Agrawal, Ajay, Arnaldo Camuffo, Alfonso Gambardella, Joshua S. Gans, Erin L. Scott, and Scott Stern. 2024. "Bayesian Entrepreneurship." *Working Paper*.

Akerlof, George A. 1970. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *The Quarterly Journal of Economics* 84 (3): 488. https://doi.org/10.2307/1879431.

Amit, Raphael, and Christoph Zott. 2001. "Value Creation in E-business." *Strategic Management Journal* 22 (6–7): 493–520. https://doi.org/10.1002/smj.187.

Aumann, Robert J. 1976. "Agreeing to Disagree." *The Annals of Statistics* 4 (6). https://doi.org/10.1214/aos/1176343654.

Barney, Jay. 1991. "Firm Resources and Sustained Competitive Advantage." *Journal of Management* 17 (1): 99–120. https://doi.org/10.1177/014920639101700108.

Barney, Jay B. 1986. "Strategic Factor Markets: Expectations, Luck, and Business Strategy." *Management Science* 32 (10): 1231–41. https://doi.org/10.1287/mnsc.32.10.1231.

Benner, Mary J., and Todd Zenger. 2016. "The Lemons Problem in Markets for Strategy." *Strategy Science* 1 (2): 71–89. https://doi.org/10.1287/stsc.2015.0010.

Bernstein, Shai, Arthur Korteweg, and Kevin Laws. 2017. "Attracting Early-Stage Investors: Evidence from a Randomized Field Experiment: Attracting Early-Stage Investors." *The Journal of Finance* 72 (2): 509–38. https://doi.org/10.1111/jofi.12470.

Bryan, Kevin A., Michael D. Ryall, and Burkhard C. Schipper. 2022. "Value Capture in the Face of Known and Unknown Unknowns." *Strategy Science* 7 (3): 157–89. https://doi.org/10.1287/stsc.2021.0126.

Christensen, Michael, and Thorbjørn Knudsen. 2010. "Design of Decision-Making Organizations." *Management Science* 56 (1): 71–89. https://doi.org/10.1287/mnsc.1090.1096.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20 (1): 37–46. https://doi.org/10.1177/001316446002000104.

Cohen, Susan, Daniel C. Fehder, Yael V. Hochberg, and Fiona Murray. 2019. "The Design of Startup Accelerators." *Research Policy* 48 (7): 1781–97. https://doi.org/10.1016/j.respol.2019.04.003.

Csaszar, Felipe A., and J. P. Eggers. 2013. “Organizational Decision Making: An Information Aggregation View.” *Management Science* 59 (10): 2257–77. https://doi.org/10.1287/mnsc.1120.1698.

Ewens, Michael, Ramana Nanda, and Matthew Rhodes-Kropf. 2018. “Cost of Experimentation and the Evolution of Venture Capital.” *Journal of Financial Economics* 128 (3): 422–42. https://doi.org/10.1016/j.jfineco.2018.03.001.

Fehder, Daniel, and Fiona Murray. 2018. “Evaluation of Early-Stage Ventures: Bias across Different Evaluation Regimes.” *Academy of Management Proceedings* 2018 (1): 17519. https://doi.org/10.5465/AMBPP.2018.17519abstract.

Felin, Teppo, and Todd R. Zenger. 2009. “Entrepreneurs as Theorists: On the Origins of Collective Beliefs and Novel Strategies.” *Strategic Entrepreneurship Journal* 3 (2): 127–46. https://doi.org/10.1002/sej.67.

———. 2017. “The Theory-Based View: Economic Actors as Theorists.” *Strategy Science* 2 (4): 258–71. https://doi.org/10.1287/stsc.2017.0048.

Gallagher, Leigh. 2017. *The Airbnb Story: How Three Ordinary Guys Disrupted an Industry, Made Billions... and Created Plenty of Controversy*. Boston: Houghton Mifflin Harcourt.

Gompers, Paul A., and Joshua Lerner. 2001. *The Money of Invention: How Venture Capital Creates New Wealth*. Boston: Harvard Business School Press.

Gonzalez-Uribe, Juanita, and Michael Leatherbee. 2018. “The Effects of Business Accelerators on Venture Performance: Evidence from Start-Up Chile.” *The Review of Financial Studies* 31 (4): 1566–1603. https://doi.org/10.1093/rfs/hhx103.

Guzman, Jorge, and Aishen Li. 2023. “Measuring Founding Strategy.” *Management Science* 69 (1): 101–18. https://doi.org/10.1287/mnsc.2022.4369.

Hayek, F. A. 1945. “The Use of Knowledge in Society.” *The American Economic Review* 35 (4): 519–30.

Howell, Sabrina T. 2020. “Reducing Information Frictions in Venture Capital: The Role of New Venture Competitions.” *Journal of Financial Economics* 136 (3): 676–94. https://doi.org/10.1016/j.jfineco.2019.10.009.

Howell, Sabrina T. 2021. “Learning from Feedback: Evidence from New Ventures*.” *Review of Finance* 25 (3): 595–627. https://doi.org/10.1093/rof/rfab006.

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

Kim, J. Daniel. 2022. "Startup Acquisitions, Relocation, and Employee Entrepreneurship." *Strategic Management Journal* 43 (11): 2189–2216. https://doi.org/10.1002/smj.3408.

Klepper, Steven. 2007. "Disagreements, Spinoffs, and the Evolution of Detroit as the Capital of the U.S. Automobile Industry." *Management Science* 53 (4): 616–31. https://doi.org/10.1287/mnsc.1060.0683.

Klepper, Steven, and Sally Sleeper. 2005. "Entry by Spinoffs." *Management Science* 51 (8): 1291–1306. https://doi.org/10.1287/mnsc.1050.0411.

Klepper, Steven, and Peter Thompson. 2010. "Disagreements and Intra-Industry Spinoffs." *International Journal of Industrial Organization* 28 (5): 526–38. https://doi.org/10.1016/j.ijindorg.2010.01.002.

Kline, Stephen J., and Nathan Rosenberg. 2009. "An Overview of Innovation." In *Studies on Science and the Innovation Process*, by Nathan Rosenberg, 173–203. WORLD SCIENTIFIC. https://doi.org/10.1142/9789814273596_0009.

Knudsen, Thorbjørn, and Daniel A. Levinthal. 2007. "Two Faces of Search: Alternative Generation and Alternative Evaluation." *Organization Science* 18 (1): 39–54. https://doi.org/10.1287/orsc.1060.0216.

Litov, Lubomir P., Patrick Moreton, and Todd R. Zenger. 2012. "Corporate Strategy, Analyst Coverage, and the Uniqueness Paradox." *Management Science* 58 (10): 1797–1815. https://doi.org/10.1287/mnsc.1120.1530.

Malenko, Andrey, Ramana Nanda, Matthew Rhodes-Kropf, and Savitar Sundaresan. 2023. "Catching Outliers: Committee Voting and the Limits of Consensus When Financing Innovation."

Marx, Matt, and David H. Hsu. 2015. "Strategic Switchbacks: Dynamic Commercialization Strategies for Technology Entrepreneurs." *Research Policy* 44 (10): 1815–26. https://doi.org/10.1016/j.respol.2015.06.016.

McHugh, Mary L. 2012. "Interrater Reliability: The Kappa Statistic." *Biochemia Medica* 22 (3): 276–82.

Morris, Stephen. 1995. "The Common Prior Assumption in Economic Theory." *Economics and Philosophy* 11 (2): 227–53. https://doi.org/10.1017/S0266267100003382.

Nanda, Ramana, and Matthew Rhodes-Kropf. 2016. "Financing Entrepreneurial Experimentation." *Innovation Policy and the Economy* 16 (January):1–23. https://doi.org/10.1086/684983.

Porter, M. E. 1996. "What Is Strategy?" *Harvard Business Review*.

Rosenberg, Nathan. 1982. *Inside the Black Box: Technology and Economics*. Cambridge [Cambridgeshire] ; New York: Cambridge University Press.

Sah, Raaj Kumar, and Joseph E. Stiglitz. 1988. “Committees, Hierarchies and Polyarchies.” *The Economic Journal* 98 (391): 451. https://doi.org/10.2307/2233377.

Scott, Erin L., Pian Shu, and Roman M. Lubynsky. 2020. “Entrepreneurial Uncertainty and Expert Evaluation: An Empirical Analysis.” *Management Science* 66 (3): 1278–99. https://doi.org/10.1287/mnsc.2018.3244.

Thiel, Peter, and Blake Masters. 2014. *Zero to One: Notes on Startups, or How to Build the Future*. 1. ed. New York, NY: Crown Business.

Van den Steen, Eric. 2004. “Rational Overoptimism (and Other Biases).” *American Economic Review* 94 (4): 1141–51. https://doi.org/10.1257/0002828042002697.

———. 2010. “Culture Clash: The Costs and Benefits of Homogeneity.” *Management Science* 56 (10): 1718–38. https://doi.org/10.1287/mnsc.1100.1214.

Verhoeven, Dennis, Jurriën Bakker, and Reinhilde Veugelers. 2016. “Measuring Technological Novelty with Patent-Based Indicators.” *Research Policy* 45 (3): 707–23. https://doi.org/10.1016/j.respol.2015.11.010.

Wu, Andy. 2016. “Organizational Decision-Making and Information: Angel Investments by Venture Capital Partners.” *Academy of Management Proceedings* 2016 (1): 11043. https://doi.org/10.5465/ambpp.2016.4.

Wuebker, Robert, Todd Zenger, and Teppo Felin. 2023. “The Theory-based View: Entrepreneurial Microfoundations, Resources, and Choices.” *Strategic Management Journal* 44 (12): 2922–49. https://doi.org/10.1002/smj.3535.

Zhao, Eric Yanfei, Greg Fisher, Michael Lounsbury, and Danny Miller. 2017. “Optimal Distinctiveness: Broadening the Interface between Institutional Theory and Strategic Management.” *Strategic Management Journal* 38 (1): 93–113. https://doi.org/10.1002/smj.2589.

Zuckerman, Ezra W. 1999. “The Categorical Imperative: Securities Analysts and the Illegitimacy Discount.” *American Journal of Sociology* 104 (5): 1398–1438. https://doi.org/10.1086/210178.

Tables and Figures

Table 1: Descriptive Statistics

Variable	N	Mean	Median	Std. dev.	Min	Max	Description
Number of Competitions	67						
Competition year	67	2015	2015	2	2011	2020	Year in which competition takes place.
Rounds per competition	67	1.76	2.00	0.72	1.00	4.00	Number of judging stages in each competition.
Number of Competition rounds	118						
Start-ups per round	118	41.19	24.00	44.76	5.00	204.00	Number of ventures in each competition round.
Judges per round	118	19.89	13.50	16.82	3.00	130.00	Number of judges in each competition round.
Number of Ventures	2650						
Number of competitions for repeat participants	619	2.85	2.00	1.34	2.00	11.00	Count of distinct competitions entered by repeat participants.
Matched with Crunchbase	2650	40%			0	1	Venture successfully matched to a Crunchbase (CB) profile.
Matched with Pitchbook	2650	28%	0		0	1	Venture successfully matched to a Pitchbook (PB) profile.
Founding year	1827	2012	2013	5	1985	2019	Year the business was founded (Source: competition data, CB, PB).
Out of business (as of 2023)	1175	19%			0	1	Venture went out of business as of 2023 (Source: CB, PB)
> 1m yearly est. revenues (as of 2023)	586	62%			0	1	Venture has at least $1m of yearly revenues as of 2023, as estimated by CB.
Millions in raised funding (as of 2023)	672	8.40	0.84	40.50	0.00	850.00	Millions of dollars raised in funding (Source: CB, PB).
Acquired (as of 2023)	1168	10%			0	1	Venture was acquired after the competition as of 2023 (Source: CB, PB).
Likelihood of successful exit (as of 2023)	492	40.33	18.50	43.88	0.00	100.00	Likelihood that the venture will successfully exit (Source: PB). Already acquired or IPO'd ventures score 100, closed ventures score 0.

# of Venture-Competition obs.	3795						
>=1 patent granted (before competition)	2307	17%			0	1	Venture was granted at least one patent before the competition (Source: competition data).
Venture score dispersion	3795	0.95	0.87	0.45	0.01	2.92	Std. deviation of the judges' scores assigned to the venture in the competition.
Differentiation score	2074	0.15	0.15	0.03	0.04	0.25	1- maximum similarity of focal startup's description relatively to other startups
Venture average grade	3795	3.87	3.93	0.87	1.02	6.77	Average of the judges' scores assigned to the venture in the competition.
Judges per startups	3795	6.26	5.00	3.77	3.00	35.00	Number of judges grading each startup.
Venture grade asymmetry	3720	0.74	0.72	0.21	0.00	2.52	Std. deviation of scores across dimensions
Number of Venture-Round obs.	4860						
Number of Evaluation Criteria per Venture-Round	4860	22.81	24.00	5.82	1.00	38	Count of distinct evaluation dimensions that the judges used to score a startup in a given competition round.
Number of Judges	1054						
Matched with LinkedIn	1054	83%			0	1	Judge successfully matched to a LinkedIn profile.
Judge is male	1047	75%			0	1	The judge is male.
Judge has MBA	761	32%			0	1	The judge has an MBA (source: LinkedIn).
Judge has PhD	761	11%			0	1	The judge has a PhD (source: LinkedIn).
Judge has founded a startup	828	34%			0	1	The judge has founded a startup (source: LinkedIn).
Judge invested in 1+ startup	1054	35%			0	1	The judge invested in at least a startup (source: competition data).

Table 2: Disagreement predicts future success

	(1) Log funding+1	(2) Log funding+1	(3) Crunchbase rank quantile	(4) At least $1m yearly revenues	(5) Likelihood of success
Std. dev. scores	1.737***	2.184**	4.234*	0.118**	9.832*
	(0.484)	(0.815)	(1.856)	(0.043)	(4.414)
Average Score	1.364***	0.999+	10.929***	0.107***	15.635***
	(0.283)	(0.604)	(1.234)	(0.030)	(3.173)
Constant	4.347**	5.054*	-3.769	0.027	-33.988*
	(1.423)	(2.457)	(5.264)	(0.132)	(14.643)
Observations	825	316	1,256	731	546
R-squared	0.173	0.254	0.212	0.180	0.228
Competition Fixed Effects?	YES	YES	YES	YES	YES
Industry Fixed Effects?	YES	YES	YES	YES	YES
Founding Year Fixed Effects?	YES	YES	YES	YES	YES
Has Patent Granted Fixed Effects?	YES	YES	YES	YES	YES
First funding after competition?	NO	YES	-	-	-

Standard errors clustered at startup level.
I only include startups judged on or before 2019 to reduce truncation concerns. An individual observation is a startup-competition (e.g, X Inc's score at the Innovation Challenge).
*** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table 3: Differentiation predicts disagreement

	(1) Std. deviation of scores	(2) Std. deviation of scores	(3) Std. deviation of scores	(4) Std. deviation of scores
Differentiation	0.527*	0.489+	0.504+	0.764*
	(0.265)	(0.278)	(0.281)	(0.362)
Business uniqueness *(Guzman/Li 2023)*				0.022
				(0.097)
Average Score			0.034***	0.019
			(0.010)	(0.014)
Constant	0.824***	0.832***	0.695***	0.694***
	(0.041)	(0.043)	(0.060)	(0.088)
Observations	2,058	1,919	1,712	978
R-squared	0.037	0.070	0.066	0.058
Competition Fixed Effects?	YES	YES	YES	YES
Industry Fixed Effects?	NO	YES	YES	YES
Founding year Fixed Effects?	NO	YES	YES	YES
Has Patent Granted Fixed Effects?	NO	NO	YES	YES

Standard errors clustered by startup.
An individual observation is a startup-competition (e.g, X Inc's score at the Innovation Challenge).
*** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table 4: Differentiation mediates disagreement's predictive power

	(1) Log funding+1	(2) Crunchbase rank quantile	(3) At least $1m yearly revenues	(4) Likelihood of success
Std. dev. scores	1.824*	-1.576	-0.032	-10.035
	(0.802)	(3.837)	(0.098)	(9.791)
Std. dev. scores # Above Median Diff.	-0.609	12.915*	0.241*	34.777*
	(1.183)	(5.143)	(0.122)	(13.629)
Above Median Differentiation	0.396	-11.852*	-0.323*	-39.270**
	(1.295)	(5.063)	(0.129)	(13.156)
Average Score	1.828***	13.993***	0.105**	16.065***
	(0.391)	(1.586)	(0.039)	(4.253)
Constant	2.452	1.211	0.586*	28.122
	(2.669)	(10.255)	(0.278)	(28.370)
Observations	605	939	541	389
R-squared	0.206	0.246	0.225	0.256
Competition Fixed Effects?	YES	YES	YES	YES
Industry Fixed Effects?	YES	YES	YES	YES
Founding Year Fixed Effects?	YES	YES	YES	YES
Has Patent Granted Fixed Effects?	YES	YES	YES	YES

Standard errors clustered at startup level.
I only include startups judged on or before 2019 to reduce truncation concerns. An individual observation is a startup-competition (e.g, X Inc's score at the Innovation Challenge).
*** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table 5: Disagreement does not capture venture risk or asymmetric outliers

	(1) Log funding+1	(2) Crunchbase rank quantile	(3) At least $1m yearly revenues	(4) Likelihood of success	(5) Out of Business
Std. dev. scores	1.780***	4.569*	0.115*	9.719*	-0.016
	(0.486)	(1.870)	(0.045)	(4.698)	(0.030)
Venture Grade Asymmetry	1.612	4.862	0.093	-0.436	0.118+
	(0.983)	(4.500)	(0.110)	(10.645)	(0.067)
Average Score	1.454***	11.394***	0.109***	15.008***	-0.020
	(0.301)	(1.287)	(0.032)	(3.434)	(0.018)
Constant	2.827	-9.343	-0.043	-30.864	0.221*
	(1.784)	(6.742)	(0.169)	(19.028)	(0.101)
Observations	818	1,245	724	540	1,415
R-squared	0.178	0.215	0.178	0.222	0.091
Competition Fixed Effects?	YES	YES	YES	YES	YES
Industry Fixed Effects?	YES	YES	YES	YES	YES
Founding Year Fixed Effects?	YES	YES	YES	YES	YES
Has Patent Granted Fixed Effects?	YES	YES	YES	YES	YES

Standard errors clustered at startup level.
I only include startups judged on or before 2019 to reduce truncation concerns. An individual observation is a startup-competition (e.g, X Inc's score at the Innovation Challenge).
*** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table 6: Disagreement decreases in later competition rounds.

VARIABLES	(1) Std. deviation of scores	(2) Std. deviation of scores	(3) Std. deviation of scores
Final competition round	-0.193***	-0.215***	-0.193***
	(0.022)	(0.026)	(0.035)
Average Score	-0.002	-0.068*	0.039
	(0.013)	(0.031)	(0.035)
Constant	0.920***	1.219***	0.769***
	(0.052)	(0.134)	(0.144)
Observations	2,393	1,675	1,161
R-squared	0.057	0.333	0.402
Competition Fixed Effects?	YES	YES	YES
Industry Fixed Effects?	YES	YES	YES
Startup Fixed Effects?	NO	YES	YES
All different judges each round?	NO	NO	YES

Standard errors clustered by startup.
Individual observations are at the startup-competition round level (e.g, X Inc's score at the Innovation Challenge's semifinal).
*** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table 7: Former founders disagree the most with the other judges

VARIABLES	(1) Absolute deviation from other judges	(2) Absolute deviation from other judges
Male	-0.017	-0.035
	(0.027)	(0.026)
Has PhD	0.051	0.045
	(0.051)	(0.050)
Has MBA	-0.047	-0.050+
	(0.029)	(0.027)
Invested in 1+ Startups	-0.013	-0.005
	(0.029)	(0.027)
Has been a Founder	0.067*	0.063*
	(0.027)	(0.026)
Constant	0.915***	0.929***
	(0.032)	(0.030)
Observations	18,145	18,083
R-squared	0.222	0.355
Competition-Round Fixed Effects	YES	YES
Industry Fixed Effects	YES	YES
Startup Fixed Effects	NO	YES

Standard errors at judge and startup level.
Individual observation is a judgement of a startup during a competition round (e.g, Judge John's score of X Inc. during the Innovation Challenge's semifinal, relative to the other judges).
*** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table 8: What aspects of a startup idea are most polarizing?

VARIABLES	(1) Std. deviation of scores
Product / Solution	0.022*
	(0.009)
Market Validation & Analysis	0.035***
	(0.009)
Industry Attractiveness	0.055***
	(0.010)
Business Model	0.077***
	(0.009)
Team	Omitted baseline
Constant	1.150***
	(0.008)
Observations	62,322
R-squared	0.156
Competition-Round Fixed Effects?	YES
Startup Fixed Effects?	YES

Standard errors clustered by startup.
Individual observations are at the startup-competition round-dimension level (e.g, X Inc's Product score at the Innovation Challenge's semifinal).
*** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table 9: Disagreement increases after a venture is granted its first patent.

	(1) Std. deviation of scores	(2) Std. deviation of scores	(3) Std. deviation of scores	(4) Std. deviation of scores	(5) Std. deviation of scores
Has granted patent	0.048*	0.041+	0.096+	0.098+	0.086+
	(0.021)	(0.023)	(0.053)	(0.053)	(0.050)
Average score				0.014	-0.027
				(0.030)	(0.033)
Constant	0.998***	1.004***	1.019***	0.963***	1.142***
	(0.009)	(0.009)	(0.009)	(0.122)	(0.138)
Observations	2,305	2,216	1,395	1,395	1,108
R-squared	0.310	0.331	0.655	0.655	0.676
Competition Fixed Effects?	YES	YES	YES	YES	YES
Industry Fixed Effects?	NO	YES	YES	YES	YES
Founding year Fixed Effects?	NO	YES	YES	YES	YES
Startup Fixed Effects?	NO	NO	YES	YES	YES
Only include startups which applied for formal IP?	NO	NO	NO	NO	YES

Standard errors clustered by startup.
Individual observations are at the startup-competition level (e.g, X Inc's score at the Innovation Challenge).
In Column (3)-(5) I control for firm fixed effects: the identifying variation come from firms which took part to multiple competition, but were granted a patent only after one (or more) of the competitions had already taken place. In Column (5), I only include startups which either applied for or were granted IP protection (including copyright, patents, and trademarks) before every competition they took part in.
*** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Online Appendix

Table 1A: List of competitions

Competition	Distinct Editions	Years	Number of participants
Techriot XLR8	1	2016	49
Arizona Innovation Challenge	14	2012-2019	1906
Angel Capital Summit	3	2014-2016	103
Capital Championship	1	2016	49
Centura Patient Engagement Challenge	1	2018	11
Clean Energy Challenge	2	2013	44
Cleantech Open	2	2011	55
Colorado Capital Conference	2	2013, 2016	74
Colorado Impact Days	1	2016	196
CSU Challenges	2	2016	80
CU Challenges	2	2013, 2016-2018	73
DOE cleantech competition	1	2013	6
Energize	1	2013	21
Energy security prize	2	2013	25
FH innovationx	1	2016	21
Grubstake awards	2	2017-2018	36
IGEM grant	2	2013-2014	34
H2O Challenges	3	2014-2016	197
Innosphere	2	2013-2015	24
Launch Alaska	1	2017	14
Missouri & Ohio clean energy challenges	2	2013	24
OEDIT grant	2	2015	29
Prime Health Challenge	6	2014-2016, 2018-2019	192
SDBC	2	2014	12
Spark program	1	2018	49
Transtech energy conference	1	2012	19
Venture Madness	6	2015-2020	452

Table 2A: Breakdown by industry

Industry	Freq.	Percent
Advanced Manufacturing & Materials	138	5.53
Aerospace/Defense	55	2.20
Bio & Life Sciences	280	11.22
Cleantech/Renewable Energy	440	17.63
IT - Hardware	108	4.33
Medicine & Health	286	11.46
Software / Consumer Web	863	34.58
Other	326	13.06
Total	2496	100.00

Table 3A: Disagreement by industry

	(1) Std. deviation of scores
Advanced Manufacturing & Materials	0.067*
	(0.028)
Aerospace/Defense	0.089**
	(0.034)
Bio & Life Sciences	0.023
	(0.020)
Cleantech/Renewable Energy	0.015
	(0.025)
IT - Hardware	0.039
	(0.031)
Medicine & Health	-0.007
	(0.032)
Other	-0.021
	(0.036)
Software	Omitted baseline
Constant	0.934***
	(0.012)
Observations	3,637
R-squared	0.271
Competition Fixed Effects?	YES

Standard errors clustered by startup.
An individual observation is a startup-competition (e.g, X Inc's score at Innovation Challenge).
*** p<0.001, ** p<0.01, * p<0.05, + p<0.1

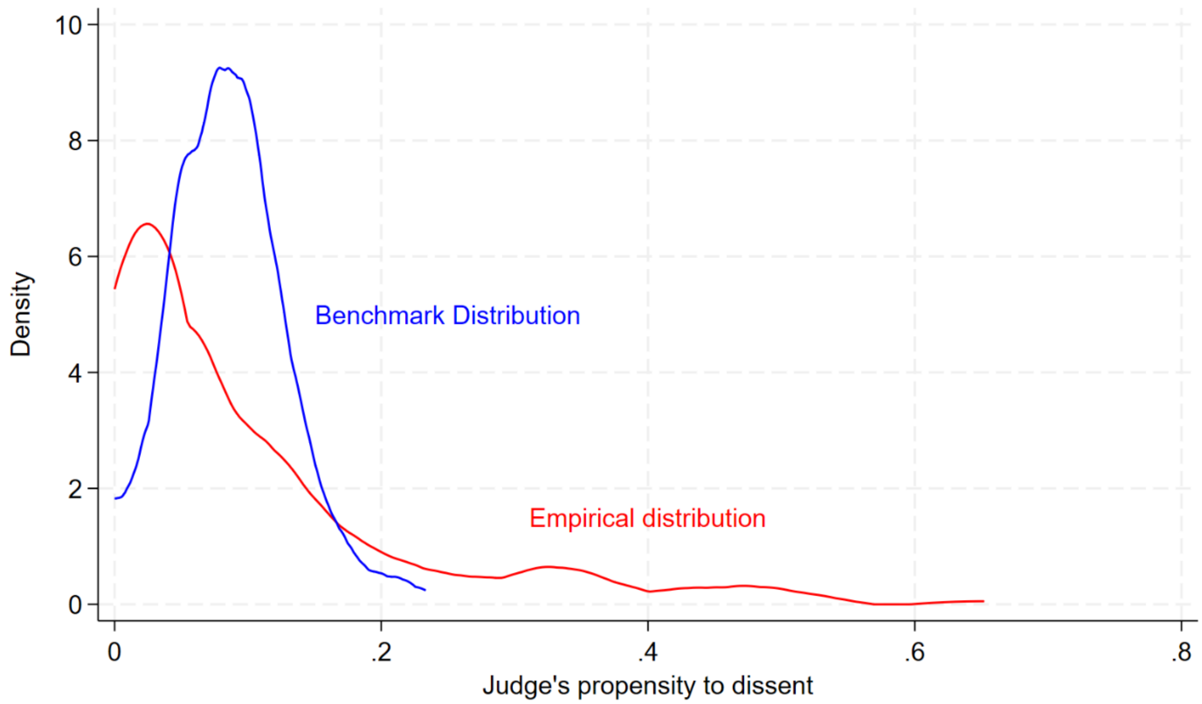
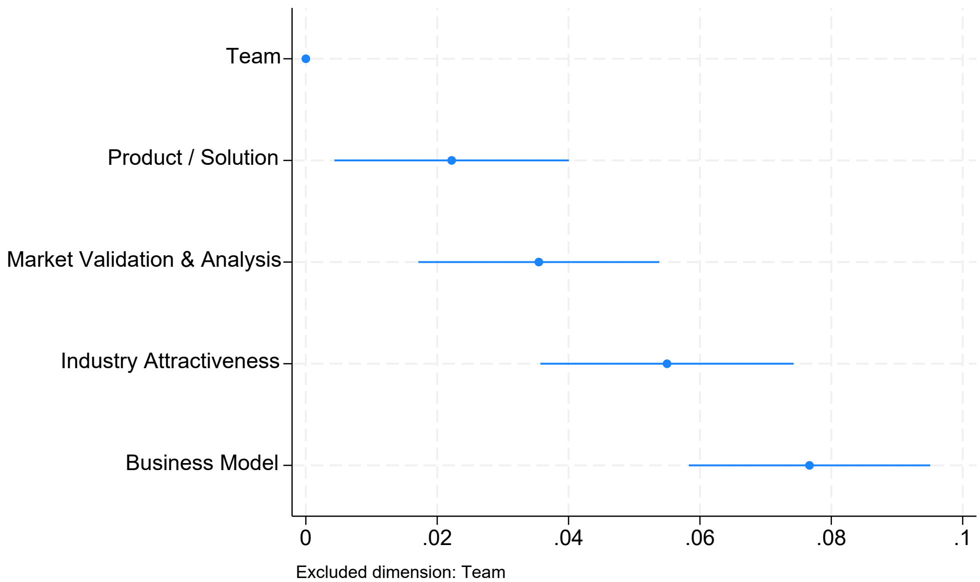


Figure 1A: Empirical distribution of a judge's propensity to dissent (red). Counterfactual distribution if all judges were equally likely to dissent (blue).

Note: The sample only includes judges grading at least 20 distinct startups. A judge gives a dissenting grade if it is at least 2 grades (out of 7) away from the other judges' average. For each judge, I calculate the proportion of dissenting grades she assigns (the "propensity to dissent"): the red line reflects the empirical distribution of this score. As a benchmark, I plot the counterfactual distribution that I would observe if all judges were equally likely to give a dissenting grade. By construction, the average of the "actual" (red) and "benchmark" (blue) distributions are the same. Simulating a benchmark distribution is useful because part of the dispersion reflects sampling variation. Even in the case that each judge was equally likely to dissent, observing only 20 judgements per judge causes some judges to appear more contrarian than others. This is why the dispersion of the simulated distribution in blue is low but not zero. It only approaches zero if I gradually increase the number of judgements for each expert to infinity.



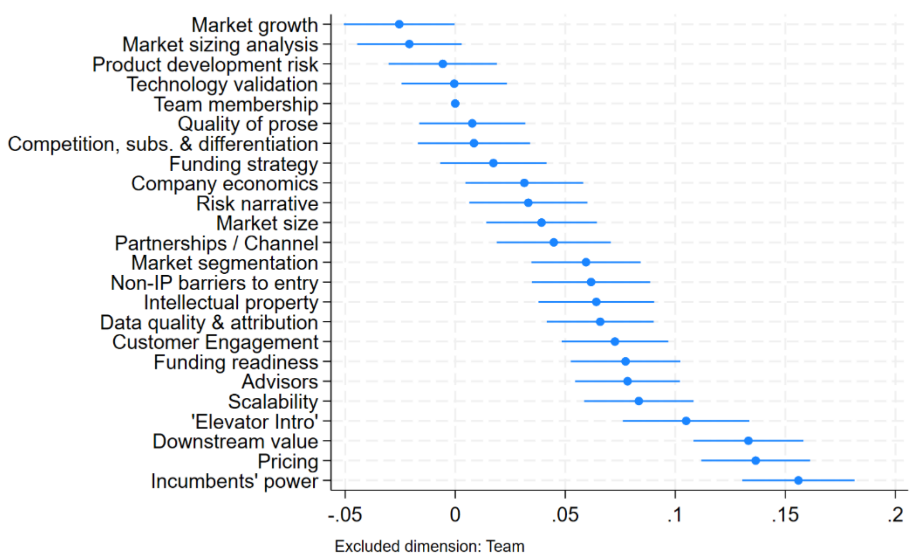


Figure 2A: Disagreement by grading dimension and subdimension (relative to team)

Note: The first regression is run at the startup-dimension-round unit of observation (see **Table 8**). The second is at the startup-subdimension-round unit of observation. Both include competition-round and startup fixed effects. Reported 95% confidence intervals are derived from standard errors clustered by startup.

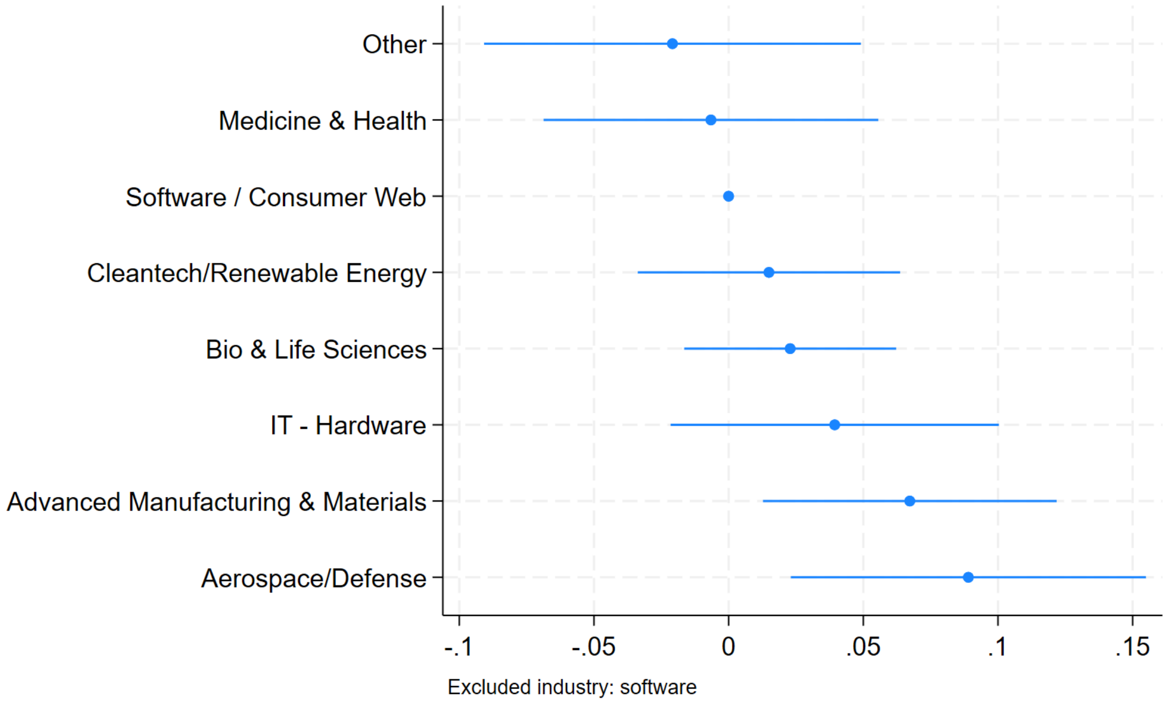


Figure 3A: Disagreement by industry (relative to software)

Note: The coefficients and confidence intervals are those reported in Table 3A.