Check for updates

Received: 4 October 2025 | Revised: 26 July 2026 | Accepted: 29 July 2026

DOI: 10.1002/sej.70042

SPECIAL ISSUE ARTICLE

SMS | Strategic Entrepreneurship Journal

WILEY

Decomposing venture evaluation: Learning about belief formation

Thomas Åstebro | Andrew Funck | Francesco Giordano

HEC Paris, ION Management Science Lab, Paris, France

Correspondence
Andrew Funck, HEC Paris, ION Management Science Lab, Paris, France.
Email: andrew.funck@hec.edu

Funding information
ION Management Science Lab

Abstract

Research Summary: Evaluating early-stage ventures is difficult because observable signals of future quality are weak and noisy. How closely are evaluators' assessments related to venture quality, and do their subjective judgments add to the information already available about a venture? Using more than 1600 screening decisions from a venture incubator, we examine how judges use application information to evaluate ventures by rating them on a set of 14 criteria and providing an overall recommendation. Judges' beliefs as recorded in the grades of the criteria explain their recommendations well, but they are only weakly associated with venture quality. We document some systematic miscalibration in how criteria are weighted, though the miscalibrations largely even out when examining the correlation between recommendations and quality. Within the models we estimate, beliefs and recommendations predict venture quality no better than ventures' observable attributes. Experienced judges are more selective but identify high-quality ventures only slightly better than novice ones.

Managerial Summary: Accelerators and incubators rely on expert judges to screen early-stage ventures, but how much does this expertise add? Using more than 1600 screening decisions at a venture incubator, we assess how well judges identify high-quality ventures and where their judgment

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Strategic Entrepreneurship Journal* published by John Wiley & Sons Ltd on behalf of Strategic Management Society.

falls short. We find that judges' recommendations are only modestly accurate. Judges form a recommendation from venture information by first rating a set of 14 criteria, but the beliefs formed are themselves only weakly tied to venture quality. Judges also misweight the evidence: several characteristics that matter most for quality carry little weight in their recommendation, while others sway decisions despite mattering little. Experience helps at the margin; seasoned judges are more selective and better at filtering out weak ventures, but they don't spot winners better than novices. A simple model using only the objective data that ventures report in their applications predicts quality as well as the judges do.

KEYWORDS
accelerators, belief formation, expert forecasting, information aggregation, venture evaluation

1 | INTRODUCTION

Selecting high-potential ventures is a first-order decision in entrepreneurial ecosystems. Resources are scarce, and their allocation by incubators, accelerators, funding agencies, and investors shapes which ideas scale and which do not (Cohen & Hochberg, 2014; S. T. Howell, 2017; Kerr & Nanda, 2015).[1] How, then, do evaluators assess the quality of new ventures?

Forecasting future outcomes is notoriously difficult for humans. A large body of research in psychology demonstrates that individuals often struggle to make accurate predictions across a wide range of domains (Dawes et al., 1989; Grove & Meehl, 1996). Venture evaluation is especially challenging because evaluators face severe uncertainty, limited historical information, weak accounting signals, and outcomes that depend on future actions by multiple stakeholders. Unsurprisingly, prior research reports substantial prediction errors in venture selection settings. In a Nigerian business plan competition, trained expert judges failed to predict subsequent venture success (McKenzie & Sansone, 2019). Similarly, Zacharakis and Shepherd (2001) found that experienced Silicon Valley venture capitalists performed poorly when forecasting venture outcomes, achieving prediction accuracies well below that of simple mechanical rules and below picking winners at random.

Yet poor prediction performance is not universal. Analysts at the Canadian Innovation Centre forecasted the commercial success of inventions with approximately 80% accuracy (Astebro, 2004), and more sophisticated statistical models of judges' decisions have achieved predictive accuracies exceeding 90% in some contexts (Elhedhli et al., 2014). Existing evidence therefore presents a puzzle. Human judgment appears highly unreliable in some venture evaluation settings but substantially more accurate in others. At present, the literature contains too few comparable studies to determine whether these differences reflect variations in evaluator skill, decision environments, venture types, or the quality of available information.

Several broader mechanisms may explain why venture evaluation is difficult. Prediction accuracy deteriorates when individuals must integrate many noisy signals (Dawes, 1975; Enke & Zimmermann, 2019),[2] when decision environments are complex (Enke, 2024; Enke & Graeber, 2023; Molavi et al., 2023), when feedback is delayed or biased (Åstebro & Koehler, 2007; Dahlin et al., 2018), or when outcomes materialize far into the future (Fischhoff, 1975).

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

All of these conditions characterize entrepreneurial evaluation. By contrast, in environments with rapid and reliable feedback, such as weather forecasting, expert probabilistic judgments tend to be well calibrated (Murphy & Winkler, 1984). Although experts are often believed to outperform novices (Kahneman & Klein, 2009), evidence on learning and expertise in entrepreneurial forecasting remains limited.

This paper examines venture evaluation at a large incubator. Like other entrepreneurial ecosystem actors, the incubator seeks to identify ventures with the highest potential for future success among a pool of highly uncertain applicants. We study judges' beliefs about a venture, as recorded in ratings on a set of evaluation criteria, and how these recorded beliefs relate to the recommendations judges ultimately provide to accept or reject ventures.[3]

Our analysis examines venture evaluation in greater detail than prior work. We proceed in several steps. We first assess how accurately judges' recommendations identify top-quality ventures. We then ask how heavily the recommendation relies on the judge's beliefs as expressed by the ratings, and how much of each rating's association with quality the recommendation actually carries. Finally, we benchmark all of this against a set of observable attributes ventures report in their applications, asking how much the grades on criteria and the recommendation add to the informativeness of the raw attributes for venture quality.

We further investigate heterogeneity across judges, focusing on professional background and opportunities for improvement through experience. In particular, we examine whether more experienced judges or venture capitalists and business angels are better forecasters than other judges at the identification of venture quality.

We find that judges are generally poor at classifying venture quality. Their prediction performance exceeds that reported by Zacharakis and Shepherd (2001) and McKenzie and Sansone (2019), but remains well below the forecasting environment studied by Astebro (2004). Judges' beliefs as expressed in their ratings largely explain their recommendations, and most of a recommendation's predictive content is in turn carried by those same beliefs. Within the linear specifications we examine, a regression model built from the beliefs performs about as well as the recommendation itself. Both alternatives have classification accuracies just slightly higher than 60%. Because observable application data predict venture outcomes as accurately as judges' subjective beliefs in the models and information sets we test, we conclude that the grading and recommendation stages add little additional predictive value beyond what is available in the raw application data.

When forming their recommendations, judges weight their beliefs in ways that depart systematically from their association with venture quality. The weight a criterion receives in the recommendation is often poorly aligned with how it relates to venture quality: the recommendation over-weights most criteria, loading heavily on dimensions whose association with quality is weak, while several of the criteria most associated with quality receive comparatively little weight.

Recommendations are broadly conservative, with false negatives exceeding false positives on average. This conservatism deepens with experience: more experienced judges are increasingly selective and form recommendations relying more on information orthogonal to their criteria ratings. In all, experience is associated with better precision in excluding low-quality applications. Our findings contribute to the literature on entrepreneurial evaluation, belief formation, expert judgment, and decision-making under uncertainty. They suggest that evaluators may improve through calibration and learning from experienced peers. But the results also suggest that forecasting the future for ventures is very difficult. At the same time, the results raise the possibility that simple statistical prediction models based on objective application data can perform as well as human evaluators in venture screening contexts. The use of such models appears not to be motivated by obtaining any increased prediction accuracy, but by the opportunity to save time.

2 | RELATED LITERATURE

We first review related literature examining settings in which judges evaluate written materials only. Research shows that evaluators can predict with relatively high accuracy whether inventions will successfully reach the market

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(Åstebro & Elhedhli, 2006). The criteria used to assess inventions before market launch also predict the probability that inventions survive conditional on market entry (Åstebro & Michela, 2005). Similarly, Scott et al. (2020) show that judges can meaningfully distinguish high-quality deep-tech ventures, though they are less successful in evaluating consumer products and SaaS ventures when relying solely on short written abstracts. By contrast, judges in a Nigerian business plan competition failed to predict venture outcomes 3 years later (McKenzie & Sansone, 2019). In Ghana, however, business plan evaluations exhibited modest predictive validity even after controlling for founder ability and survey-based characteristics (Fafchamps & Woodruff, 2017).

Research on pitch competitions reaches similarly mixed conclusions. Judges in pitch competitions predict subsequent venture financing outcomes (S. T. Howell, 2020, 2021), but their evaluations also exhibit systematic gender bias (Kanze et al., 2018). The dispersion of judges' evaluations predicts subsequent venture success (Gius, 2025), and more diverse evaluation panels generate more informative predictions than homogeneous panels, although only for above-average ventures (Lane & Rietzler, 2026). Kalvapalle et al. (2024) provide a recent review of the pitching literature.

Taken together, prior evidence suggests that judges may or may not play an important role in distinguishing successful from unsuccessful ventures. Existing studies further suggest that evaluators may be less biased, and potentially more accurate, when reviewing written applications than when observing live pitch presentations. More generally, experts often outperform novices in forecasting uncertain outcomes (Chu et al., 2024; D'acunto et al., 2023; DellaVigna & Pope, 2018; Moore et al., 2017), although expertise develops slowly and requires substantial learning (Moore et al., 2017). Related work examines how venture capitalists' education and experience influence screening decisions (Moritz et al., 2022). Studies of learning curves further show that performance improvements tend to be steep initially but decline with repeated experience (Mazur & Hastie, 1978; Waldman et al., 2003); reviews are provided by Thompson (2010) and Dahlin et al. (2018). Additional research examines venture evaluation by venture capitalists (Gompers et al., 2020; Lyonnet & Stern, 2024) and the interaction between human judgment and AI-based venture evaluation systems (Lin et al., 2024).

Our analysis focuses exclusively on the venture selection stage rather than downstream incubation activities. For related work examining the broader acceleration process, see Avnimelech et al. (2025); Cohen et al. (2019); Sharapov and Dahlander (2025).

3 | DATA: ADMISSION PROCESS AND VENTURE CHARACTERISTICS

The HEC Incubateur is a venture acceleration program based in Paris. It caters primarily to young entrepreneurs living in Paris who have recently graduated from university, often from an entrepreneurship program at HEC, or from its partner institutions and universities in Paris. Each quarter (referred to as a batch), the incubator admits approximately 10 ventures, depending on the availability of slots and the quality of applications received. In the initial step of the admission process, applicants complete an online form. Some applicants are deemed ineligible and are automatically rejected. Those meeting the eligibility requirements proceed to the next stage where they are invited to complete a second application form. The application asks for information about the team, venture development stage, customers, market, business model, competition, product development, plans, funding, conversion rate, regulatory concerns, and impact. Summary statistics of venture characteristics from eligible applications across 9 batches between Summer 2021 and Winter 2024 are shown in Table 1, covering 1638 evaluations and 579 applicants.

Approximately 40% of the eligible applying ventures have at least one HEC Paris graduate among their founders, and 46% include at least one female co-founder. The average founder age is 33 years, and teams typically consist of four members, with two co-founders. At the time of application, 64% of ventures are incorporated. The dominant business model is subscription-based (54%), followed by marketplaces (21%) and e-commerce (15%). A smaller share of ventures follows service-oriented models, including agencies and training centers (4%), mobile apps (3%), and media or social networks (3%). Ventures span a range of sectors, with Software and IT (23.7%), Finance and Real

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70942 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Applying ventures.

Panel A: General statistics	
Number of companies	579
HEC graduate	40%
Age	33
Female	46%
Team size	4 (2)
Incorporated	64%
Panel B: Business model	
Subscription	54%
Marketplace	21%
(e)Commerce	15%
Agency/training center	4%
Mobile app	3%
Media/social networks	3%
Panel C: Maturity	
Minimum viable product	47%
Prototype	27%
Full working product	26%
Panel D: Sector	
Software IT	23.7%
Finance real estate	12.4%
Life science healthcare	11.4%
Hospitality education	9.7%
Consulting professional services	9.3%
Retail wholesale distribution	6.9%
Consumer goods	6.4%
Media entertainment culture	6.4%
Other	13.8%

Note: The table presents summary statistics of ventures that pass through the eligibility criteria at the HEC Incubateur. Panel A provides general statistics on the applicants. Panel B displays the distribution of business models among applicants, ranking them by prevalence. Panel C categorizes the maturity stage of the applicants' products, from early-stage prototypes to fully developed offerings. Panel D reports the sector where applying ventures operate.

Estate (12.4%), and Life Science and Healthcare (11.4%) being the most represented. The maturity of applicants varies: 47% have developed a minimum viable product, 27% are at the prototype stage, and 26% have a fully working product at the time of application.

Each application is assessed by at least two professional judges who are randomly assigned by the manager of the admissions process. Judges are not informed about the identity of the other judges on a submitted file. Judges are given 2 weeks to complete their assessments. For each batch, a judge first decides how many applications she can review, with an allocation ranging between 5 and 15 files. A judge can decide not to participate in making evaluations in a batch and will be asked again for the next batch. Judges are either employees of the incubator or volunteer external judges, representing, for example, venture capitalists, business angels, and entrepreneurs. Judges are asked to assess ventures on 14 criteria, covering aspects such as the clarity of the user pain point, feasibility of the

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

proposed solution, competitive advantage, founders' execution capacity, and financial viability. Each criterion is graded on a 1–3 scale, where 1 indicates the lowest and 3 the highest quality. Each grade of each criterion is specifically described in instructions to the judge. A venture needs to reach a specific target to be able to be given a higher grade. With "belief" we mean that the judge estimates that a specific grade (1, 2, or 3) has the highest probability of being true for a specific criterion for a specific venture. In the empirical analysis, we refer to these beliefs simply as "criteria grades" or even just "criteria." The complete list of evaluation criteria is reported in Appendix C. The order in Appendix C is given as in the form that the judges fill out, but they may fill out the criteria in any order they prefer. At the end of each evaluation, judges provide a trinary recommendation on whether the application should advance to the next stage (1 = No, 2 = Uncertain, 3 = Yes). For the analysis, we recode this recommendation as a binary variable, assigning 1 to "Yes" and 0 otherwise.

Interviews with judges, the manager of the admission process, and the manager of the incubator indicate that admitting high-quality ventures is a clearly stated primary concern for judges in the first stage. There are two additional stages leading to an admission or rejection: a group meeting among a subset of the judges that made the initial assessments, and a one-day in-person interview stage by an external Jury for a final group of applicants. In the second stage, the committee members are given a rank ordering from the top ranked to the lowest ranked venture by criteria grades from the first stage. The committee is asked to make a decision to promote a venture to the next stage only at the margin. Specifically, only 20 ventures on average are passed to the final Jury stage in each batch. However, if a venture above the cutoff is signaled by a committee member to have some critical flaw that the judges in the first stage did not know of, that venture is removed from consideration. In the final stage, an external Jury interviews all remaining applicants and makes a joint recommendation to admit or reject. The manager of the admissions process and the incubator manager then makes a final decision. These two stages contain different judgment processes and considerations and are therefore not part of the analysis in this paper. For example, strategic considerations are discussed in the group meeting. The committee further discusses whether they would be able to help an applicant, indicating an evaluation consistent with the social welfare impact of the program, not just picking the best applicants. An example may be if an applicant is too far developed in the commercialization process, so that they may not benefit much from being admitted. To repeat, all analysis in this article is on the first-stage recommendations.

We collect economic performance data of all applicants during the Summer of 2024. We use the personal name of the applicant (usually the CEO) and co-applicant and the name of the business in the application to manually search for venture information. Data are collected from LinkedIn, Crunchbase, venture, and personal web sites. Data on business performance include number of full-time employees, amount of funding collected, and business survival. Following past verified practice, we assume that the business has been closed if there is a missing value from all three data sources (Lerner & Malmendier, 2013). If instead we find evidence that the venture remains active but cannot verify its employee count or funding, we code the venture as open and set only the unverified metric to 0, leaving any verified metric at its recorded value. In particular, business activity is broadly retrieved through web searches.[4]

Table A1 presents key statistics on ventures' applications, survival rates, full-time equivalents (FTEs), and funding outcomes across all the batches and in total. Between approximately 50 and 100 applications are evaluated in each batch. As of Summer 2024, the proportion of ventures still active ranges from 56% to 85%, with an average survival rate of 73%. The average number of FTEs among surviving ventures is 6.7, with the 75th percentile reaching 8. Regarding funding outcomes, only 11% of ventures receive post-application funding on average. Among those that do receive funding, the average amount is €3.5 million. This figure is skewed upward by a few ventures securing exceptionally large amounts, particularly in the earlier batches.

Judges are tasked with identifying ventures with the highest economic potential. We measure venture quality using two outcomes observed as of 2024: the number of full-time employees and the cumulative amount of external funding raised (including grants). Ventures that have ceased operations receive a score of zero. Because ventures from earlier cohorts have had more time to grow, we remove cohort-level differences by regressing each outcome

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

on batch fixed effects and retaining the residuals. This removes the average differences across cohorts in the two quality measures; it does not equalize venture age or time since founding. We then rank ventures separately on each residualized measure and construct a composite performance index as the equally weighted average of the two ranks. A venture is classified as high quality if its composite index falls in the top 40% of the distribution, and low quality otherwise (thus, by construction, ventures that have ceased operations are classified as low quality). The 40% threshold corresponds to the average recommendation rate across all judge evaluations and reflects the admission capacity of the incubator's first selection stage. We consider alternative thresholds and weighting schemes in robustness checks.

3.1 | Notation

Before presenting the main analyses, we establish the notation used throughout. The indicator $v_i \in \{0, 1\}$ denotes the quality label of venture i, taking value $v_i = 1$ if venture i falls above the threshold. We denote by $Y_{i,j} \in \{0, 1\}$ the admission recommendation that judge j assigns to venture i, with $Y_{i,j} = 1$ indicating a recommendation for admission. The variable $X_{i,j,k} \in \{1, 2, 3\}$ denotes the grade on criterion k that judge j assigns to venture i.

4 | RESULTS

4.1 | Judges classification accuracy

How accurately do individual judges identify ventures with the greatest economic potential? Table 2 offers preliminary evidence, reporting descriptive comparisons of survival, FTEs, and funding by recommendation status.

TABLE 2 Outcomes differences by recommendation status.

			Survival		FTE		€Mln funding	
Batch	Evaluations	$P(Y_{i,j} = 1)$	Mean-diff	r_{pb}	Mean-diff	r_{pb}	Mean-diff	r_{pb}
2021 Fall	144	35%	15%*	0.14*	7.22***	0.29***	0.36	0.02
2021 Summer	156	40%	29%***	0.3***	7.55***	0.29***	3.77*	0.19**
2022 Spring	145	40%	14%**	0.19**	3.08***	0.29***	0.05	0.05
2022 Summer	283	36%	33%***	0.32***	5.33***	0.26***	0.33***	0.24***
2022 Winter	156	37%	17%**	0.18**	8.25***	0.31***	0.92**	0.21***
2023 Spring	219	47%	0%	0	1.85**	0.16**	0	0.01
2023 Summer	188	41%	14%**	0.16**	1.69***	0.25***	0.09**	0.17**
2023 Winter	154	40%	8%	0.1	2.24***	0.24***	0.06	0.11
2024 Winter	193	44%	9%	0.1	1**	0.16**	0.02	0.02
Overall	1638	40%	17%***	0.18***	3.87***	0.22***	0.51**	0.06**

Note: The table reports descriptive statistics by batch and for the overall sample, comparing ventures that received a positive recommendation with those that were rejected. Column 2 shows the number of evaluations in each batch. Column 3 reports the proportion of applications that received a positive recommendation. Columns 4, 6, and 8 present the difference in means between recommended and rejected applications for three key business outcomes: the probability of survival, the number of full-time equivalents (FTEs), and the amount of funding raised post-application, respectively. Columns 5, 7, and 9 report the point-biserial correlation coefficients between judges' recommendations and these same business outcomes. Asterisks indicate statistical significance.
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 Recommendation errors.

		Recommendation	
		Rejected	Recommended
Venture quality	Top	46%	54%
		$n = 323$	$n = 373$
	Low	70%	30%
		$n = 661$	$n = 281$

Note: The matrix compares judges' admission recommendations (columns) with ventures' measure of economic quality (rows) across 1638 evaluations. Each row sums to 100%; cell entries report the share of evaluations of a given quality resulting in each recommendation, with absolute counts in parentheses. Shaded green cells denote correct classifications (top ventures recommended, low ventures rejected); coral cells denote misclassifications.

Recommended ventures exhibit higher survival rates and employ more FTEs than those not recommended: on average, they have a 17-percentage-point higher survival rate and 3.9 additional FTEs. The mean funding gap is relatively large given that ventures are at an early stage (€ 0.51 million), but it is not always significant across batches. This limited association with funding may reflect both judges' constrained ability to predict fundraising potential and the scarcity of post-application funding data, as few ventures secure external capital at this stage.

One might raise endogeneity concerns: ventures receiving positive recommendations are more likely to be admitted, and admission to the incubator may in turn enhance venture performance. This does not invalidate the descriptive forecasting exercise, but it means that the resulting quality measure should not be read as an untreated or purely latent measure of venture quality. However, even if there is a treatment effect from admission, it only applies to between 10% and 20% of applicants of each batch. Furthermore, while treatment may shift up the quality measure for those admitted, in order to make our subsequent analysis corrupted, the treatment must perturb the association between, for example, a specific criterion or the recommendation and the quality measure. It should be noted that the venture attributes, criteria grades, and the recommendations are all Granger causal to the quality measures. We thus use the terminology "predictive" as the predictors occur before venture quality is measured. However, the criteria grades cannot be claimed to be causal to the recommendation, as both are recorded simultaneously. We prefer to claim that the recommendation is a simple aggregation of the beliefs a judge holds about the various criteria.

Do judges consistently identify top-quality ventures? Table 3 reports classification accuracy across all evaluations, comparing judges' recommendations with ventures' quality labels. Judges issued a correct positive recommendation in only 54% of evaluations of top-quality ventures, failing to identify the remaining 46%. Classification ability is stronger for low-quality ventures: judges issued a correct negative recommendation in 70% of such evaluations, misclassifying only 30%. This asymmetry yields an overall accuracy of 63%.

To quantify the predictive content of judges' recommendations, we estimate a linear probability model and assess its predictive accuracy via leave-one-venture-out cross-validation. For each venture i, we remove all of its recommendations from the sample, estimate

$$v_i = \alpha + b\,Y_{ij} + \epsilon_{ij}, \tag{1}$$

on the remaining recommendations, and generate out-of-sample predictions for the held-out venture's recommendations. Holding out all judgments of a given venture prevents information about a venture's realized quality from leaking into the prediction of its own recommendations. A recommendation is classified as positive whenever its predicted value exceeds the empirical base rate (i.e., the prediction should be above the 60th percentile). The model attains an out-of-sample balanced accuracy of 61.9%, compared with 50.5% for a random classifier predicting at the prior. Sensitivity (the true-positive rate) is 53.6% versus 41.7%, and specificity (the true-negative rate) is 70.2%

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

versus 59.3%. Judges' recommendations thus improve on chance by nearly 12 percentage points in balanced accuracy, yet leave close to half of all top-quality ventures undetected. The following section examines how the recommendation $Y_{i,j}$ is formed from judges' beliefs about the underlying criteria, and the extent to which those beliefs predict venture quality.

4.2 | Beliefs and recommendations

We begin by characterizing how judges aggregate the criteria into an admission recommendation. We estimate the linear probability model

$$Y_{i,j} = \alpha + \Sigma_k \gamma_k X_{i,j,k} + \epsilon_{i,j}, \tag{2}$$

in which the coefficient γ_k measures the weight a criterion receives in the summative judgment. Column 1 of Table 4 reports the estimates. The criteria account for roughly half of the variation in recommendations ($R^2 = 0.51$), and the recommendation loads most heavily on the perceived ability to attract venture-capital or business-angel financing, execution speed, expected sales volume, competitive advantage, and the business model, each significant at the 1% level. Even if the model is a reduced linearly additive version of their mental model, judges' final decisions are strongly rooted in the beliefs they form from the application.

Equation (2) also lets us ask how much of the recommendation's own correlation with venture quality is driven by the criteria, and how much is orthogonal to their linear projection. We relate venture quality to the recommendation split into the part explained by the criteria and an orthogonal residual,

$$v_i = \alpha + \delta_1 \widehat{Y}_{i,j} + \delta_2 Y_{i,j}^{\perp} + \epsilon_{i,j}, \qquad \widehat{Y}_{i,j} = \widehat{\alpha} + \Sigma_k \widehat{\gamma}_k X_{i,j,k}, \qquad Y_{i,j}^{\perp} = Y_{i,j} - \widehat{Y}_{ij}, \tag{3}$$

where δ_1 isolates the component of the recommendation driven by criteria, and δ_2 the residual component orthogonal to them. This component may reflect information available to the judge but not captured by the evaluation criteria, but it may also reflect nonlinear use of the recorded criteria, judge-specific thresholds, omitted variables, measurement error, or decision noise. Column 3 gives $\widehat{\delta}_1 = 0.343$ and $\widehat{\delta}_2 = 0.135$, both significant at the 1% level. Most of the recommendation's correlation with quality, roughly two-thirds, is already encoded in the criteria, while a nontrivial third operates independently of the grades.

The complementary question is how much of the quality-relevant information in the criteria judges actually use. We first estimate the weights on venture quality from a reduced form linear additive regression,

$$v_i = \alpha + \Sigma_k \beta_k X_{i,j,k} + \epsilon_{i,j}, \tag{4}$$

which measures each criterion's total association with realized quality (Column 4, $R^2 = 0.085$). We then condition on the recommendation,

$$v_i = \alpha + \delta Y_{i,j} + \Sigma_k \theta_k X_{i,j,k} + \epsilon_{i,j}, \tag{5}$$

so that θ_k captures the residual association of criterion k with quality that survives after the recommendation is accounted for (Column 5). Adding the criteria on top of the recommendation raises the fit from $R^2 = 0.058$ (recommendation only) to $R^2 = 0.093$, and several θ_k remain statistically significant, most notably sales volume, skills and connections, competitive advantage, and motivation and ambition. The criteria therefore retain an in-sample association with quality that is not channeled through the recommendation.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 Criteria, judges' recommendations and ventures' quality.

	$Y_{i,j}$	v_i				
Model	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Variables						
Skills and connections	0.0163			0.0412**	0.0390**	0.0412**
	(0.0112)			(0.0185)	(0.0185)	(0.0185)
Motivation and ambition	0.0189*			−0.0319*	−0.0344**	−0.0319*
	(0.0104)			(0.0166)	(0.0166)	(0.0166)
Sales volume	0.0586***			0.0537***	0.0458**	0.0537***
	(0.0122)			(0.0177)	(0.0178)	(0.0176)
VC and BA	0.1278***			0.0483***	0.0311*	0.0483***
	(0.0126)			(0.0184)	(0.0185)	(0.0181)
Competitive advantage	0.0490***			−0.0379**	−0.0445***	−0.0379**
	(0.0107)			(0.0163)	(0.0162)	(0.0161)
Business development status	0.0082			0.0289*	0.0277*	0.0289*
	(0.0109)			(0.0164)	(0.0163)	(0.0163)
Customer traction	0.0124			0.0368*	0.0351*	0.0368*
	(0.0126)			(0.0198)	(0.0197)	(0.0197)
User pain point	0.0102			−0.0058	−0.0071	−0.0058
	(0.0110)			(0.0162)	(0.0160)	(0.0160)
Business model	0.0422***			−0.0081	−0.0138	−0.0081
	(0.0117)			(0.0181)	(0.0182)	(0.0181)
Understanding of competition	0.0245**			0.0057	0.0024	0.0057
	(0.0105)			(0.0159)	(0.0158)	(0.0158)
Regulation	0.0020			−0.0001	−0.0004	−0.0001
	(0.0100)			(0.0153)	(0.0152)	(0.0152)
Tech product development status	0.0322**			0.0147	0.0103	0.0147
	(0.0128)			(0.0191)	(0.0191)	(0.0190)
Execution speed	0.0813***			0.0264	0.0154	0.0264
	(0.0122)			(0.0173)	(0.0175)	(0.0172)
Finance	−0.0043			−0.0127	−0.0121	−0.0127
	(0.0090)			(0.0143)	(0.0142)	(0.0142)
Y_{ij}		0.2421***			0.1352***	
		(0.0311)			(0.0373)	
$\hat{Y}_{ij}$			0.3432***			
			(0.0441)			
$Y_{ij}^{\perp}$			0.1352***			0.1352***
			(0.0379)			(0.0373)

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 (Continued)

	$Y_{i,j}$	v_i				
Model	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Fit statistics						
Observations	1638	1638	1638	1638	1638	1638
R^2	0.51403	0.05752	0.06813	0.08476	0.09348	0.09348

Note: This table reports OLS estimates from the following specifications. Column 1 projects the judge's recommendation onto evaluation criteria: $Y_{ij} = \alpha + \Sigma_k \gamma_k X_{ijk} + \epsilon_{ij}$. Columns 2–6 regress venture quality v_i on different combinations of recommendation components and criteria scores: Column 2, $v_i = \alpha + \delta Y_{ij} + \epsilon_{ij}$; Column 3, $v_i = \alpha + \delta_1 \widehat{Y}_{ij} + \delta_2 Y_{ij}^{\perp} + \epsilon_{ij}$; Column 4, $v_i = \alpha + \Sigma_k \beta_k X_{ijk} + \epsilon_{ij}$; Column 5, $v_i = \alpha + \delta Y_{ij} + \Sigma_k \theta_k X_{ijk} + \epsilon_{ij}$; Column 6, $v_i = \alpha + \delta_2 Y_{ij}^{\perp} + \Sigma_k \beta_k X_{ijk} + \epsilon_{ij}$. Y_{ij} is judge j's binary recommendation for venture i, v_i is a binary indicator of venture quality, and $X_{i\,jk}$ are standardized criterion scores. $\widehat{Y}_{ij}$ and $Y_{ij}^{\perp}$ denote the fitted value and residual from Column 1, representing the criteria-predictable and criteria-orthogonal components of the recommendation. By the Frisch–Waugh–Lovell theorem, $\beta_k = \delta\gamma_k + \theta_k$: each criterion's total correlation with quality (Column 4) decomposes into a share carried by the recommendation and a direct residual association. The reported R^2 is the adjusted R^2, which penalizes the inclusion of additional regressors and so does not mechanically increase with the number of explanatory variables, allowing comparison of fit across specifications of differing dimension. Standard errors clustered at the venture level in parentheses.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Finally, we combine the full association of the criteria with quality together with the part of the recommendation orthogonal to them in

$$v_i = \alpha + \delta_2 Y_{i,j}^{\perp} + \Sigma_k \beta_k X_{i,j,k} + \epsilon_{i,j}. \tag{6}$$

Because $Y^{\perp}$ is orthogonal to the criteria, the slopes on $X_{i,j,k}$ coincide with β_k of Equation (4), and Equation (6) is an exact reparameterization of Equation (5); the two specifications (Columns 5 and 6) attain the same and highest fit in the table ($R^2 = 0.093$). Equation (6) is the more interpretable of the two: β_k recovers each criterion's full association with quality, and the coefficient on $Y^{\perp}$ measures the predictive value of the recommendation that operates independently of the grades.

These specifications are linked by an exact algebraic identity. By the Frisch–Waugh–Lovell theorem, the weight of each criterion on quality decomposes as follows:

$$\beta_k = \delta\gamma_k + \theta_k, \tag{7}$$

with δ the coefficient on the recommendation in Equation (5) (equivalently, on $Y^{\perp}$ in Equation 6). The term $\delta\gamma_k$ is the share of a criterion's association with quality that is routed through the recommendation, and θ_k is the share that is not. Two complementary views of this decomposition follow.

Figure 1 plots each criterion by the weight it receives in the recommendation against its weight on venture quality, with the dashed 45° line marking equal weighting. Points below the line are over-weighted by the recommendation relative to their predictive importance for quality; points above it are under-weighted. The recommendation over-weights most criteria: VC and BA, execution speed, competitive advantage, motivation and ambition, and the business model all sit well below the line, and the gap is statistically significant for each. Only a minority of criteria fall on the under-weighted side, and these are weighted close to proportionally: skills and connections, customer traction, and business development status lie modestly above the line, with none of these departures significant at conventional levels. For several of the over-weighted criteria, including competitive advantage, motivation and ambition, and the business model, the weight on venture quality is itself negative, so the recommendation loads positively on a criterion whose association with quality is actually negative.

Figure 2 splits each β_k into the part carried by the recommendation ($\delta\gamma_k$, light bars) and the part orthogonal to it (θ_k, dark bars). For the criteria most predictive of quality, that is, sales volume, VC and BA interest, skills and

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

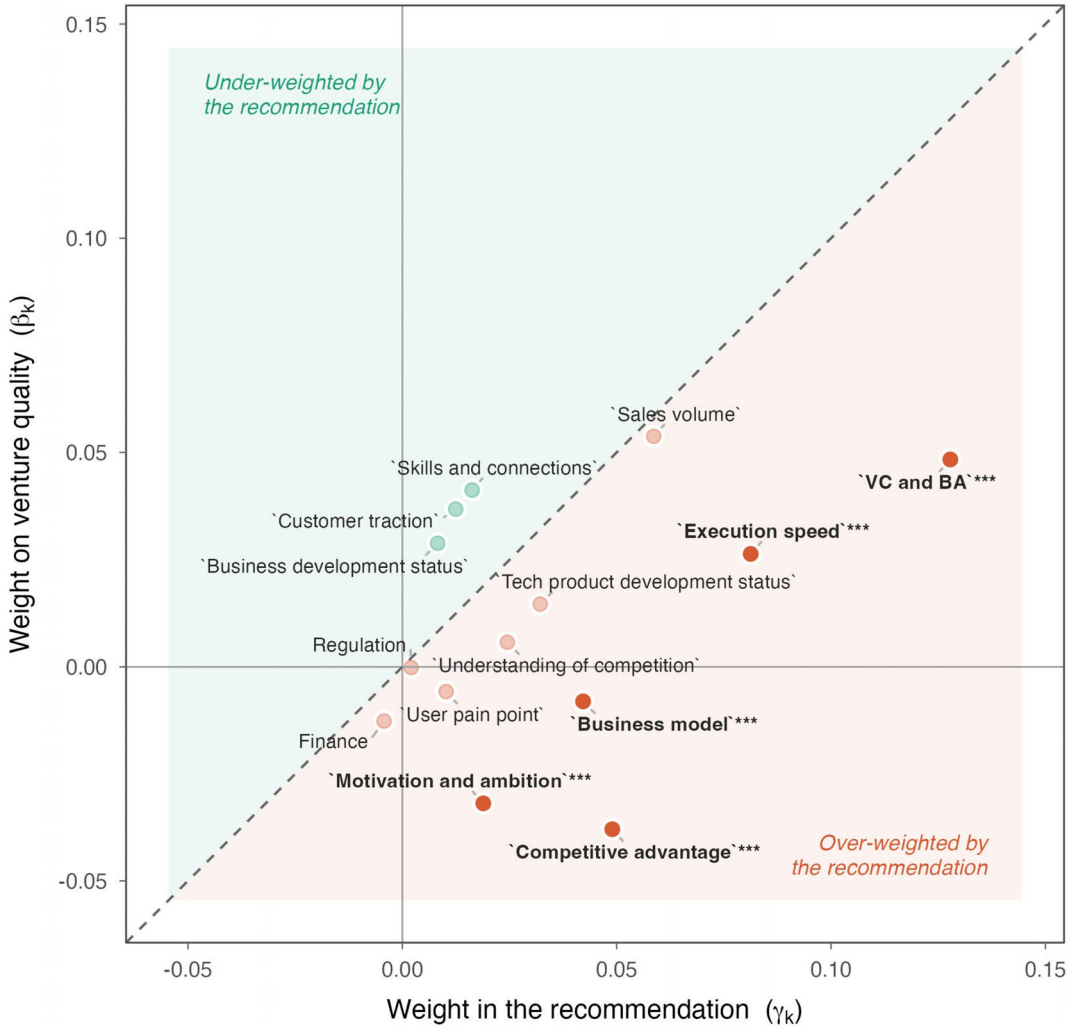


FIGURE 1 Criteria over- and under-weighting. The figure plots each evaluation criterion k in the space of its weight in the judge's recommendation, $\widehat{\gamma}_k$ (horizontal axis), against its weight on venture quality, $\widehat{\beta}_k$(vertical axis). Coefficients are obtained from the system of equations:

$$v_i = \alpha_1 + \Sigma_k \beta_k X_{ijk} + \epsilon_{ij},\ Y_{ij} = \alpha_2 + \Sigma_k \gamma_k X_{ijk} + \eta_{ij},$$

where v_i is a binary indicator of venture quality, Y_{ij} is judge j's binary recommendation for venture i, and X_{ijk} are standardized criterion scores. Both equations are estimated jointly via Seemingly Unrelated Regressions (SUR). The dashed 45° line marks equal weighting ($\gamma_k = \beta_k$), where the recommendation responds to a criterion in proportion to that criterion's predictive importance for venture quality. Points in the lower-right (shaded coral) region lie below the line and are over-weighted by the recommendation ($\gamma_k > \beta_k$); points in the upper-left (shaded teal) region lie above the line and are under-weighted ($\gamma k < \beta k$). Faded points denote criteria for which H_0: $\gamma_k = \beta_k$ cannot be rejected at the 10 % level. Bold labels with asterisks report the significance of H_0: $\gamma_k = \beta_k$, tested using the SUR cross-equation covariance matrix, which accounts for correlation of residuals across equations: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

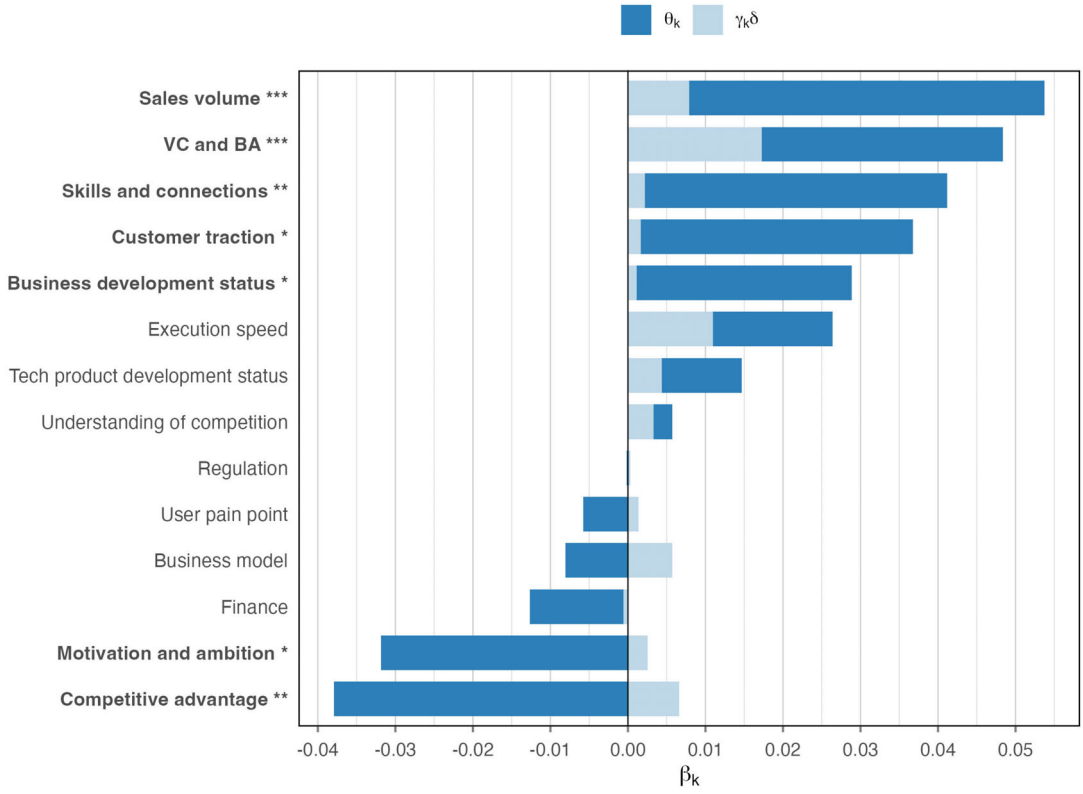


FIGURE 2 Criteria information and judges' recommendations. The figure displays the decomposition of each criterion's correlation with venture quality into a component captured by the judge's recommendation and a residual component. We estimate:

$$v_i = \alpha + \Sigma_k \beta_k X_{ijk} + \epsilon_{ij} \quad v_i = \alpha + \delta Y_{ij} + \Sigma_k \theta_k X_{ijk} + \epsilon_{ij} \quad Y_{ij} = \alpha + \Sigma_k \gamma_k X_{ijk} + \epsilon_{ij},$$

where v_i is a binary indicator of venture quality, Y_{ij} is judge j's binary recommendation for venture i, and X_{ijk} are standardized criterion scores. By the Frisch–Waugh–Lovell theorem, $\beta_k = \delta\gamma_k + \theta_k$ for each criterion k. The dark bars show θ_k, the association between criterion k and quality uncorrelated with the recommendation; the light bars show $\delta\gamma_k$, the share of the criterion's predictive content that is carried over by the recommendation. Standard errors are clustered at the venture level. Asterisks denote significance of β_k: $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

connections, customer traction, and business development status, the orthogonal component θ_k dominates: the recommendation transmits only a small fraction of their measured association with quality. The criterion with the largest share flowing through the recommendation is the assessment of whether a venture is investable (VC and BA), where over a third of the association flows through the recommendation, indicating that judges draw on investability-related information reflected both in the grades and, indirectly, in the recommendation itself. Descriptively, then, a substantial share of the criteria's association with quality runs orthogonal to the recommendation rather than through it.

Whether this residual represents a missed predictive opportunity is an out-of-sample question. In terms of overall discrimination, the answer is barely: augmenting the recommendation with the criteria grades ($Y \rightarrow X + Y^{\perp}$) raises balanced accuracy from 61.9% to 62.7%, leaves overall accuracy unchanged (63.1%–62.8%), and improves the Brier score only slightly (0.232–0.228, Table 5). But the composition of the gain matters more than its size. The grades re-

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5 Prediction performance at evaluation level.

			Classification				Scoring	
Predictors	Description	*N*	Accuracy	Bal. Acc.	Sensitivity	Specificity	Brier	AUC
–	*Random classifier*	*1638*	*0.518*	*0.505*	*0.417*	*0.593*	*0.244*	*0.500*
Y	Admission Recommendation	1638	0.631	0.619	0.536	0.702	0.232	0.376
X	Criteria grades	1638	0.615	0.619	0.642	0.596	0.230	0.649
$Y^{\perp} + \widehat{Y}$	Recommendation decomp.	1638	0.626	0.618	0.568	0.669	0.229	0.639
$Y + X$	Recommendation + criteria	1638	0.628	0.627	0.626	0.628	0.228	0.656
$X + Y^{\perp}$	Criteria + ort. Recommendation	1638	0.628	0.627	0.626	0.628	0.228	0.656
Z	Ventures' attributes	1638	0.611	0.612	0.621	0.604	0.232	0.652
$X + Z$	Criteria + Ventures' attributes	1638	0.615	0.616	0.622	0.610	0.230	0.665
$Z + Y$	Ventures' attributes + Recommendation	1638	0.617	0.617	0.618	0.617	0.227	0.675
$Z + X + Y^{\perp}$	Ventures' attributes + criteria + ort. Recommendation	1638	0.627	0.628	0.632	0.623	0.228	0.674

>*Note*: The table reports cross validation out-of-sample classification and scoring performance at the evaluation level across prediction models. The outcome is a binary indicator of venture quality, and the first column lists the regressors entered in the OLS prediction model. All models are assessed using leave-one-venture-out cross-validation: all evaluations of a held-out venture are removed before estimation, and the held-out evaluations are then classified according to the empirical prior. Classification metrics comprise Sensitivity (the true-positive rate), and Specificity (the true-negative rate), and their unweighted (Accuracy) and weighted, (Balanced Accuracy) means; scoring metrics comprise the Brier score and the area under the receiver operating characteristic curve (AUC). The first row reports a random classifier drawing at the empirical prior, which serves as a benchmark. Predictors are denoted Z (venture attributes), X (criteria grades), and Y (admission recommendation), with $Y^{\perp}$ marking a component orthogonal to the linear projection of criteria grades ($\widehat{Y}$). Higher values indicate better performance for all metrics except the Brier score, for which lower is better. Values in italics refer to classification metrics of the "random classifier". They were made in italics to differentiate this classifier from all others.

balance the error profile, lifting sensitivity from 53.6% to 62.6% at the expense of specificity, which falls from 70.2% to 62.8%. Whether this trade is worthwhile depends on the accelerator's objective: if it weights missing a top-quality venture more heavily than admitting a weak one the shift is valuable, since using the criteria directly recovers close to a fifth of the top ventures the recommendation alone would have rejected. The criteria's residual association with quality thus sharpens no overall discrimination, but may pay off on a margin the accelerator might care about; to that extent, aggregating the grades into a single recommendation leaves some quality-relevant information unused.

4.3 | Information extraction

The analyses so far take the criteria grades and the recommendation as the relevant information set for assessing ventures' quality. We now turn to the attributes ventures report in their applications and ask how the grades and the recommendation relate to them. Three possibilities arise. The criteria grades and the recommendation may reproduce the predictive of the attributes, neither adding nor discarding information about quality; they may extract less than the attributes contain, losing quality-relevant information in the course of evaluation; or they may add content beyond the attributes, drawing on information the recorded fields do not capture. We adjudicate among these by

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

comparing the out-of-sample predictive performance of the attributes alone against specifications that augment them with the criteria grades, the recommendation, and both.

These attributes are the raw information ventures supply in their applications, prior to any assessment by the judges. We take these fields directly from the application form: continuous fields are used as reported, and categorical fields are dummy-coded with missing responses assigned their own category. They comprise the gender composition of the founding team, founder roles, whether a founder attended HEC Paris and the type of degree obtained, years since graduation, the time since the team began working full-time on the project, the presence of a co-founders' pact, incorporation status and the time since company registration if incorporated, company type and business model, team size and sector experience, whether the venture already has customers, the number of customers interviewed, market potential, recent turnover, weekly growth and conversion rates, solution maturity, financing needs, funds raised, and runway.

Because the number of attributes is large relative to the number of ventures, we select an informative subset via LASSO rather than including every attribute. Selection is nested inside the leave-one-venture-out procedure we use to measure predictive accuracy: for each fold, attributes missing for more than 20 ventures are dropped and the remainder are standardized. We then estimate a LASSO regression of the quality label on the encoded attributes, selecting the penalty by fivefold cross-validation at the minimum Mean Squared Error (MSE) value, and retain those with nonzero coefficients. A core set of six attributes, the degree obtained, incorporation status, company type, business model, whether the venture already has customers, and the number of employees, is retained in the large majority of folds; a handful of additional attributes (founder education, months worked full-time, founder role, customers interviewed, and market potential) are retained only intermittently. Throughout the paper, the vector $\boldsymbol{Z}_i$ refers to this fold-specific, selected subset of structured, applicant i attributes. Appendix D reports the full encoding, LASSO specification, fold construction, and standardization procedure in detail.

We benchmark an OLS model using the attributes alone,

$$v_i = \alpha + \boldsymbol{\Phi}^{\intercal}\boldsymbol{Z}_i + \epsilon_{i,j}, \tag{8}$$

against specifications that augment it with the criteria grades, with the recommendation, and with both

$$v_i = \alpha + \boldsymbol{\Phi}^{\intercal}\boldsymbol{Z}_i + \Sigma_k \beta_k X_{i,j,k} + \delta_2 Y_{i,j}^{\perp} + \epsilon_{i,j}, \tag{9}$$

where $Y^{\perp}$ is the component of the recommendation orthogonal to the grades, defined in Equation (3). As before, all models' predictive performance is assessed out-of-sample by leave-one-venture-out cross-validation, and Table 5 reports performance for the nested predictor sets $\boldsymbol{Z}$, $\boldsymbol{Z} + X$, $\boldsymbol{Z} + Y$, and $\boldsymbol{Z} + X + Y^{\perp}$.

Taken on their own, the three information sets attain similar point estimates of out-of-sample predictive performance. The recommendation Y, the criteria grades X, and the attributes $\boldsymbol{Z}$ attain balanced accuracies of 61.9%, 61.9%, and 61.2%, respectively, with the latter two reaching comparable area under the curve (AUC) values (0.649 and 0.652)[5]; no single set shows consistently higher point estimates, and all three improve on the random benchmark by a similar margin. Combining sets adds remarkably little. Pairing the attributes with the criteria grades ($\boldsymbol{Z} + X$) leaves balanced accuracy essentially unchanged at 61.6%, adding the recommendation ($\boldsymbol{Z} + Y$) yields the best AUC (0.675) and the lowest Brier score (0.227), and the fullest specification, $\boldsymbol{Z} + X + Y^{\perp}$, lifts balanced accuracy only to 62.8%. Across every metric, the gain from adding a second or third information set to any single one is limited.

The one systematic exception is again sensitivity. Both the attributes and the criteria grades recover top-quality ventures at a markedly higher rate than the recommendation alone: where Y correctly flags 53.6% of top ventures, X and $\boldsymbol{Z}$ raise this to around 64% and 62%, respectively, each trading away some specificity in return. The recommendation is more conservative than the underlying information would warrant, missing top ventures that a simple linear model of both the attributes and the grades would have identified.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

All models in Table 5 exhibit similar point estimates, and although we do not formally test whether the observed differences are statistically distinguishable, these results admit a unified interpretation. The attributes **Z**, the criteria grades X, and the recommendation Y each predict venture quality at roughly the same level, and combining them improves discrimination only trivially. The recommendation thus adds little information, within the specifications we test, beyond what the attributes already reveal; rather, it compresses the available information into a single judgment about as predictive, in point estimates, as the two combined. In this sense Y is an aggregator of the application information within the linear models class, with one important qualification: in collapsing that information into a binary call, the recommendation errs toward caution, recovering fewer top-quality ventures than the grades or attributes.

4.4 | Experience

The limited baseline accuracy of the recommendation raises a natural question: is higher experience associated with better performance? We measure experience as the total number of evaluations a judge completes over the full sample period and we compare judges with different overall activity levels. We order the 72 judges in our sample by the number of evaluations they have ever completed and group them into deciles, each accounting for roughly a 10th of all evaluations, so that the lowest decile contains the historically least-active judges and the highest the most-active. Evaluation activity is highly concentrated: the least active quarter of judges together account for only 3.5% of all evaluations, while the most active 25% account for nearly three quarters (Figure B1). For each decile, Figure 3 reports the true-positive rate, the true-negative rate, their weighted mean (balanced accuracy), and the overall recommendation rate.

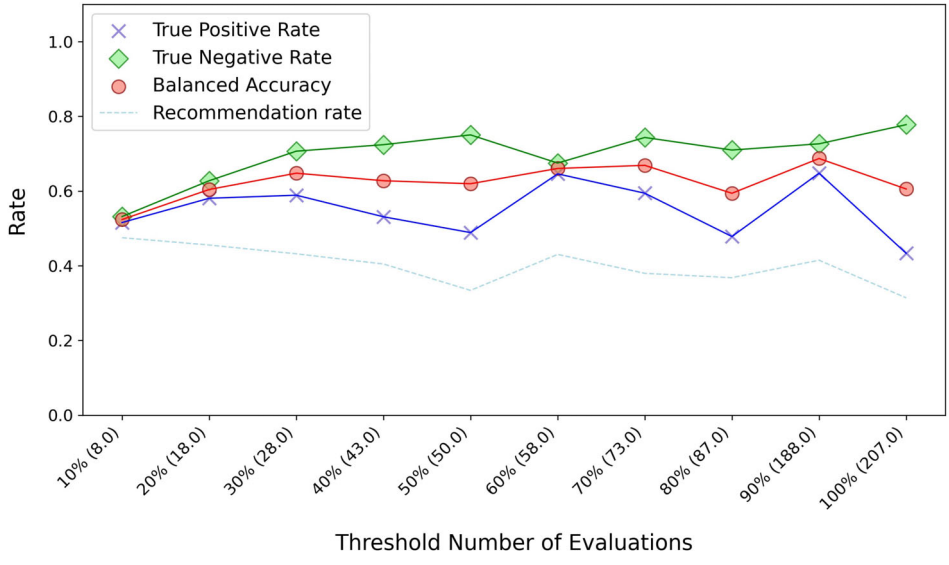


FIGURE 3 Decision accuracy by experience deciles. For each decile of judges, ordered by the number of evaluations completed, we plot: (i) the share of top-quality ventures that judges recommend (true-positive rate), (ii) the share of non-top-quality ventures that judges do not recommend (true-negative rate), (iii) their weighted mean (balanced accuracy), and (iv) the overall Recommendation rate. Judges are binned so that each decile accounts for ≈10% of all evaluations: The first decile contains the least-active judges (up to eight evaluations), the second decile the next group (≈ 9–19 evaluations), and so on. *X*-axis labels show the cumulative evaluation share (10%, 20%, …) with the corresponding evaluation-count threshold in parentheses.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Balanced accuracy rises with total experience, with the bulk of the gain realized between the lowest deciles and the middle of the distribution and little improvement thereafter. This pattern is consistent with a large number of studies showing steep learning gains with a few repetitions, later to flatten out with higher experience (Thompson, 2010), though as a comparison across judges it cannot distinguish learning-by-doing from other sources of variation across judges. The gain is driven almost entirely by the true-negative rate: judges with higher experience are associated with better screening out of low-quality ventures, while their true-positive rate is flat or, in the most-active deciles, slightly lower. This pattern could be explained by learning-by-doing. Since all judges, through the composition of applications and the random allocation procedure, see more low-quality than high-quality ventures, a learning process would be expected to raise accuracy faster for low-quality applicants than for high-quality ones. The recommendation rate falls in parallel, from roughly one-half among the least experienced judges to around one-third among the most active, consistent with increased selectivity as experience accumulates.

Table 6 confirms this pattern in the out-of-sample prediction exercise and extends it to two further dimensions of judge heterogeneity. Across the recommendation- and criteria-based models, balanced accuracy rises from the first through the third quartile of total experience and changes little in the fourth, mirroring Figure 3.

Two cross-sectional contrasts in point estimates stand out: external evaluators show higher balanced accuracy and AUC than internal ones, and non-investors show higher values than investors (though we only have a small number of investors) across nearly all specifications.

Finally, we ask whether the source of the recommendation's association with venture quality shifts with experience. Estimating the recommendation decomposition described in equation (3) separately by experience quartile, Figure 4 plots the fraction of quartile-specific R^2 attributable to the criteria-aligned component, relative to the orthogonal component. This criteria share is highest among the least experienced judges, close to unity in the first quartile, and lower from the second quartile onward, ranging between 80% and 90%, where the orthogonal component becomes statistically significant. For less experienced judges, the recommendation's association with venture quality is almost entirely attributable to codified criteria, whereas more experienced judges increasingly draw on information orthogonal to the linear projection of criteria grades.

Read together, the three results tell a coherent story about what experience is, and is not, associated with. Judges with higher experience exhibit higher balanced accuracy, but only up to the center of the experience distribution and only by sharpening the rejection of weak ventures; and their recommendations draw relatively less on the linear projection of codified criteria. What it leaves untouched is the detection of top-quality ventures.

4.5 | Robustness

This section examines the robustness of the prediction results reported in Table 5 to key modeling choices. Two main modeling choices are considered, with further robustness checks deferred to Appendix. Table 7 reports points estimates of out-of-sample prediction performance across different choices of the prior used as benchmark and the weights used to construct the venture quality index. Both are compared to the baseline case, in which the prior is set to 40%, the average recommendation rate (interpreted as capacity at the first decision stage) and equal weights are assigned to FTEs and funding. Table 7 presents results separately for alternative priors (specifically, 20% and 50%) and for alternative weights on FTEs relative to funding (25% and 75%).

Table 7 compares out-of-sample Balanced Accuracy and AUC. Across models, both metrics are highest at the lowest prior: balanced accuracy generally rises and AUC rises uniformly as the prior decreases. The pattern is likewise uniform across all models as the relative weight on FTEs increases: point estimates of both AUC and balanced accuracy rise.

Table A2 shows further robustness results, comparing recommendation decisions defined from two complementary specifications: (i) a trinary specification that explicitly treats intermediate uncertain recommendations as a distinct decision category, and (ii) a specification that drops observations with missing information rather than imputing

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6 Heterogeneity analysis of prediction performance at evaluation level.

		Baseline		Q1		Q2		Q3		Q4		Internal		External		Investor		No investor	
		($N = 1638$)		($N = 395$)		($N = 430$)		($N = 418$)		($N = 395$)		($N = 888$)		($N = 750$)		($N = 177$)		($N = 1461$)	
Predictors	Description	Bal. Acc.	AUC	Bal. Acc.	AUC	Bal. Acc.	AUC	Bal. Acc.	AUC	Bal. Acc.	AUC	Bal. Acc.	AUC	Bal. Acc.	AUC	Bal. Acc.	AUC	Bal. Acc.	AUC
–	*Random classifier*	*0.505*	*0.500*	*0.510*	*0.500*	*0.529*	*0.500*	*0.499*	*0.500*	*0.475*	*0.500*	*0.539*	*0.500*	*0.536*	*0.500*	*0.532*	*0.500*	*0.503*	*0.500*
Y	Admission recommendation	0.619	0.376	0.588	0.343	0.625	0.385	0.639	0.401	0.615	0.365	0.616	0.368	0.622	0.384	0.591	0.350	0.623	0.380
X	Criteria grades	0.619	0.649	0.569	0.615	0.598	0.612	0.644	0.672	0.601	0.615	0.604	0.632	0.619	0.659	0.554	0.554	0.621	0.652
$Y^{\perp} + \widehat{Y}$	Recommendation decomposition	0.618	0.639	0.597	0.602	0.626	0.630	0.628	0.657	0.615	0.610	0.613	0.634	0.625	0.632	0.561	0.562	0.625	0.646
$Y + X$	Recommendation +criteria	0.627	0.656	0.578	0.611	0.612	0.623	0.649	0.677	0.619	0.627	0.617	0.638	0.638	0.666	0.563	0.542	0.632	0.660
$X + Y^{\perp}$	Criteria + orthogonal recommendation	0.627	0.656	0.573	0.611	0.614	0.624	0.647	0.677	0.621	0.627	0.617	0.638	0.638	0.666	0.563	0.543	0.632	0.660
Z	Venture attributes	0.612	0.652	0.655	0.648	0.607	0.611	0.590	0.600	0.619	0.661	0.615	0.642	0.605	0.648	0.450	0.484	0.626	0.663
$X + Z$	Criteria + venture attributes	0.616	0.665	0.608	0.646	0.578	0.594	0.615	0.632	0.608	0.661	0.608	0.649	0.616	0.658	0.488	0.487	0.622	0.674
$Z + Y$	Venture attributes +recommendation	0.617	0.675	0.655	0.660	0.587	0.630	0.596	0.626	0.639	0.678	0.618	0.663	0.606	0.667	0.477	0.491	0.636	0.687
$Z + X + Y^{\perp}$	Venture attributes +criteria + orthogonal recommendation	0.628	0.674	0.603	0.645	0.572	0.600	0.632	0.645	0.619	0.667	0.605	0.657	0.626	0.665	0.488	0.481	0.634	0.685

Note: The table reports heterogeneity in cross validation out-of-sample prediction performance across subsamples of evaluators. The outcome is a binary indicator of venture quality, and the first column lists the regressors entered in the OLS prediction model. All models are assessed using leave-one-venture-out cross-validation: all evaluations of a held-out venture are removed before estimation, and the held-out evaluations are then classified according to the empirical prior. The table reports out-of-sample balanced accuracy (Bal. Acc.) and the area under the receiver operating characteristic curve (AUC) for a set of subsamples of evaluators: across quartiles of cumulative evaluator experience (Q1 the least experienced), by evaluator affiliation (internal vs external), and by investor status (investor versus non-investor); the Baseline columns pool all evaluators. The first row reports a random classifier drawing at the empirical prior, which serves as a benchmark. Predictors are denoted Z (venture attributes), X (criteria grades), and Y (admission recommendation), with $Y^{\perp}$ marking a component orthogonal to the linear projection of criteria grades ($\widehat{Y}$). Higher values indicate better performance for all metrics except the Brier score, for which lower is better. Values in italics refer to classification metrics of the "random classifier". They were made in italics to differentiate this classifier from all others.

19324434x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

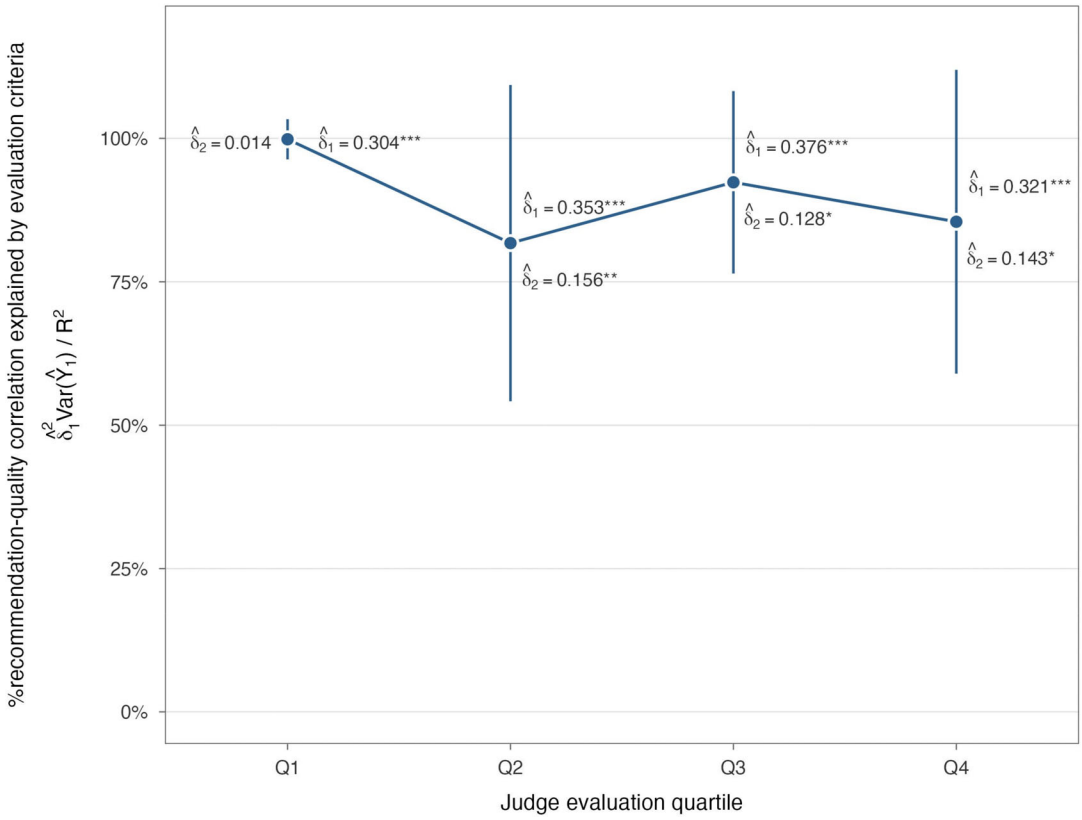


FIGURE 4 Criteria, recommendation and experience. The figure displays the share of the recommendation–quality relationship attributable to the criteria-aligned component, $\hat{\delta}_1^2 \widehat{\text{Var}}\left(\hat{Y}_{ij}\right)/R^2$, across quartiles of judge evaluation experience, where quartile 1 contains the least experienced judges. Coefficients are obtained from the regression:

$$v_i = \alpha + \delta_1 \hat{Y}_{ij} + \delta_2 Y_{ij}^{\perp} + \epsilon_{ij},$$

where $\hat{Y}_{ij}$ is the fitted value and $Y_{ij}^{\perp}$ the residual from the first-stage projection of the judge's recommendation onto standardized criteria scores:

$$\hat{Y}_{i,j} = \hat{\alpha} + \Sigma_k \hat{\gamma}_k X_{i,j,k}, \qquad Y_{i,j}^{\perp} = Y_{i,j} - \hat{Y}_{ij.}$$

Both regressions are fully interacted with judge evaluation quartile indicators, so $\hat{\delta}_1$, $\hat{\delta}_2$, and all variance terms below are quartile-specific. $\hat{Y}_{ij}$ and $Y_{ij}^{\perp}$ are orthogonal within quartile q, and the quartile-q R^2 decomposes additively:

$$R_q^2 = \underbrace{\frac{\hat{\delta}_{1,q}\,\widehat{\text{Var}}\left(\hat{Y}_{ij}\,|q\right)}{\widehat{\text{Var}}(v_i|q)}}_{\text{Criteria aligned}} + \underbrace{\frac{\hat{\delta}_{2,q}\,\widehat{\text{Var}}\left(Y_{ij}^{\perp}\,|q\right)}{\widehat{\text{Var}}(v_i|q)}}_{\text{Orthogonal}}$$

The plotted share is the criteria-aligned term divided by R^2: The fraction of the recommendation's explanatory power over venture quality that operates through the observable criteria. Point estimates and 95% confidence intervals are reported for the share in each quartile, computed via the delta method. The level of each component $(\hat{\delta}_1, \hat{\delta}_2)$ is annotated alongside each point estimate. Standard errors are clustered at the venture level. Asterisks denote significance of each component: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 7 Robustness analysis of prediction performance at evaluation level.

		Baseline			Prior = 0.2		Prior = 0.5		Weights = 0.25		Weights = 0.75	
Predictors	Description	*N*	Bal. Acc.	AUC	Bal. Acc.	AUC	Bal. Acc.	AUC	Bal. Acc.	AUC	Bal. Acc.	AUC
–	*Random classifier*	*1638*	*0.505*	*0.500*	*0.520*	*0.500*	*0.498*	*0.500*	*0.511*	*0.500*	*0.511*	*0.500*
Y	Admission recommendation	1638	0.619	0.376	0.646	0.417	0.604	0.353	0.606	0.360	0.616	0.372
X	Criteria grades	1638	0.619	0.649	0.657	0.709	0.602	0.628	0.596	0.622	0.630	0.664
$Y^{\perp} + \widehat{Y}$	Recommendation decomposition	1638	0.618	0.639	0.653	0.684	0.600	0.622	0.606	0.622	0.621	0.648
$Y + X$	Recommendation +criteria	1638	0.627	0.656	0.660	0.712	0.610	0.634	0.606	0.629	0.635	0.666
$X + Y^{\perp}$	Criteria + orthogonal recommendation	1638	0.627	0.656	0.660	0.712	0.611	0.634	0.605	0.629	0.634	0.666
Z	Venture attributes	1638	0.612	0.652	0.646	0.730	0.589	0.637	0.601	0.627	0.620	0.665
$X + Z$	Criteria + venture attributes	1638	0.616	0.665	0.661	0.747	0.616	0.657	0.609	0.654	0.614	0.679
$Z + Y$	Venture attributes +recommendation	1638	0.617	0.675	0.679	0.751	0.622	0.660	0.630	0.653	0.641	0.681
$Z + X + Y^{\perp}$	Venture attributes +criteria + orthogonal recommendation	1638	0.628	0.674	0.671	0.752	0.616	0.664	0.621	0.661	0.629	0.684

Note: The table reports robustness checks for the cross validation out-of-sample classification and scoring performance of the different models at the evaluation level. The outcome variable is a binary indicator of venture quality, and the first column lists the regressors included in the OLS prediction model. All models are assessed using leave-one-venture-out cross-validation: all evaluations of a held-out venture are removed before estimation, and the held-out evaluations are then classified according to the empirical prior. The table reports out-of-sample balanced accuracy and the area under the receiver operating characteristic curve (AUC) under alternative assumptions regarding the prior probability (0.2 and 0.50) and the weight assigned to FTEs relative to funding capital (0.2 and 0.75) in the construction of the economic quality measure used to define the binary quality label. These results are compared with the baseline specification, which assumes a prior probability of 0.40, consistent with the judges' recommendation rate, and equal weights (0.50 each) on FTEs and funding capital. Values in italics referred to classification metrics of the "random classifier". They were made in italics to differentiate this classifier from all others.

19324343x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esc Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

them as closed. These results are compared with the baseline specification, in which recommendation decisions are binary and missing information is handled in a conservative manner (e.g., assuming non-operating status when no information on venture activity is available). The trinary specification leaves point estimates close to baseline throughout, with balanced accuracy and AUC shifting by only small amounts across models. The Dropped NA specification instead shows a larger decline in point estimates across every model, most pronounced for the attribute-based models. As a result, the ordering of models by balanced accuracy is not preserved in the Dropped NA sample: the attribute-based models, broadly comparable to Y and X at baseline, fall to the bottom of the ranking once observations with missing information are dropped. The smaller estimation sample in the Dropped NA specification likely contributes to this pattern.

An additional issue to consider is measurement error. The introduction of this article stated that "venture evaluation is especially challenging because evaluators face severe uncertainty, limited historical information, weak accounting signals, and outcomes that depend on future actions by multiple stakeholders," pointing to the possibility of potentially severe measurement error. In regression models with measurement error, it is crucial to distinguish between the dependent and the independent variables. It is well known that measurement error in the dependent variable reduces model accuracy but introduces no bias in the estimates (Greene, 2003). Measurement error in the independent variables, by contrast, results in attenuation bias, pulling the estimated regression coefficients toward zero as the measurement error grows (Greene, 2003). Given the likely scope for measurement error in this setting, we assess its consequences directly.

Tables A3 and A4 present results obtained by correcting for potential measurement error in criteria ratings and recommendation decisions, following the methodology of Gillen et al. (2019), namely using peer evaluations as instruments. Specifically, for every evaluation grade on each criterion, as well as for the decision recommendation, we construct a first-stage regression that uses as instruments the grade on that criterion (respectively, the admission recommendation) from the set of other judges who see and evaluate the same venture. The fitted value from the first stage is then used as a predictor. The econometric specification of the first stage is as follows. Let N (i) denote the pool of judges that evaluate an individual venture i, let $m_i = |\mathrm{N}(i)|$ be the cardinality of the set of judges evaluating a given venture, and for each judge j, order the other judges by grade as j', ..., $j'_{(m_i-1)}$. For each pool size m, we run a separate regression pooling all (i, j) such that $m_i = m$, as follows:

$$X_{ijk} = \alpha_k^m + \Sigma_{l=1}^{m-1} \beta_{lk}^m X_{ij'_l k} + \epsilon_{ijk},$$

$$Y_{ij} = \alpha^m + \Sigma_{l=1}^{m-1} \beta_l^m Y_{ij'_l} + \epsilon_{ij},$$

The ORIV first-stage measurements are then derived as the OLS fitted values:

$$X_{ijk}^{\mathrm{ORIV}} = \widehat{\alpha}_k^m + \Sigma_{l=1}^{m-1} \widehat{\beta}_{lk}^m X_{ij'_l k},$$

$$Y_{ij}^{\mathrm{ORIV}} = \widehat{\alpha}^m + \Sigma_{l=1}^{m-1} \widehat{\beta}_l^m Y_{ij'_l}.$$

The approach uses each peer as a separate instrument and adopts an ordering rule for the peers by grade. Since peers are assumed to be exchangeable, individual coefficients of the first stage are not separately interpretable, only the fitted projection is.

Table A3 reports the corresponding coefficient estimates for the correlation of criteria grades with recommendation decisions and venture quality. correcting for measurement error. The R^2 of every regression is higher than in the corresponding baseline specification (Table 4), both for the recommendation decision and for venture quality. Overall, the sign and statistical significance of the coefficients remain stable compared to baseline. The magnitude of

the coefficients in Table A3 is generally larger when considering the recommendation decision as a dependent variable, comparable when regressing venture quality on ORIV criteria grades, and smaller when regressing venture quality on the ORIV recommendation decision, whether entered on its own or decomposed into its criteria—orthogonal and criteria—fitted components. Table A4 reports out-of-sample classification performance using the ORIV-corrected measurements in place of the raw criteria grades and recommendation decisions. Comparing predictive power after the first-stage ORIV correction, the models that use only the admission recommendation yield performance that is substantially unchanged (very similar balanced accuracy, 0.618 against 0.619), with higher sensitivity and lower specificity, lower Brier score, and a higher AUC. On the other hand, for the remaining models, the ORIV measurements yield higher accuracy. In the model that only uses criteria grades, the balanced accuracy rises from 0.619 in the baseline case to 0.629 under the ORIV measurement, with both sensitivity and specificity increasing. A similar pattern holds for the other models: slightly higher balanced accuracy, reflecting improved sensitivity and specificity, a lower Brier score, and a higher AUC.

Finally, Table A5 presents venture-level results obtained by aggregating evaluation-level predictions across judges. The table presents the prediction performance of the same sets of regressions displayed in Table 5, but where we aggregate data to the venture level. Each venture's evaluation-level predictions are averaged across its judges before classification.

Aggregating in this way smooths out judge-specific biases. The procedure lifts every balanced accuracy by a few points: the recommendation reaches 65.6% and the criteria grades 65.3%. That is, it seems that averaging across judges provides the most important statistical gain of all the models we test.

Moreover, the relative ranking of information sets by balanced accuracy shifts, as the attribute-based specifications show lower point estimates in out of sample prediction than recommendation and criteria-based models. The sensitivity gap likewise persists, with the criteria grades recovering top-quality ventures at a markedly higher rate (73.6%) than the recommendation (63.6%).

5 | DISCUSSION

This paper examines how judges at an incubator evaluate early-stage ventures: how their beliefs about a venture, as recorded into grades on a set of pre-determined criteria with pre-defined levels, relate to their overall recommendation to accept or reject the venture, and how well those judgments align with venture quality. Judges are asked solely to assess the future potential of each venture, with no other considerations to take into account.

Our results suggest that judges' beliefs are only modestly predictive of venture quality. Regressing the quality label on the criteria grades achieves out-of-sample predictive accuracy that attains similar point estimates to that of judges' recommendations; within the models and the information sets we examine, compressing judges' multidimensional criteria beliefs into a single accept/reject decision does not seem to sacrifice much of their predictive content. The decomposition of the recommendation's predictive content into a criteria-aligned and an orthogonal component shows that most of the variance in venture quality explained by the recommendation is carried by the recorded grades, with a smaller share attributable to the orthogonal component. The beliefs judges form about venture attributes therefore appear to contain only limited information about venture quality. Nevertheless, it would be premature to conclude from this statistical exercise that the grading of the criteria is useless for belief formation about venture quality. It could be that judges would be much noisier in their recommendation if they did not go through the rating of the criteria. The process might still inform their mental model of venture assessment. The appropriate experiment to test whether filling in the criteria generates higher recommendation accuracies would be to randomly exclude the criteria for half the judges and see how predictive the recommendations then are.

In addition, we examine how the observable venture attributes reported in the application predict quality directly. The most predictive attributes for venture quality attain predictive accuracy with similar point estimates to the recommendation and to the judges' criteria beliefs, and combining the three information sets adds little

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

additional predictive value within the specifications we test. This represents a more critical test of the value of collecting judges' beliefs. This comparison provides a demanding benchmark for the value of collecting judges' beliefs: a simple linear, additive model built from the application file attains predictive accuracy comparable, in point estimates, to the recommendation. There are two systematic exceptions to these conclusions, each providing conflicting information. First, judge conservatism: the recommendation recovers fewer top-quality ventures than either the grades or the attributes would. This suggests that either grades or venture attributes are useful, and that directly using the recommendation loses accuracy at the top of the quality distribution. Second, aggregation across judges: averaging grades and recommendations across the multiple judges who review the same venture improves classification ability, suggesting that combining recommendations across judges reduces noise relative to relying on any single judge's recommendation. This finding is robust to a complementary correction for measurement error that instruments each judge's grades and recommendation with those of peer judges evaluating the same venture, following the Obviously Related Instrumental Variables approach of (Gillen et al., 2019).

Comparing coefficient estimates across two models—one in which the criteria grades predict the recommendation, the other in which they predict venture quality—highlights systematic differences in how judges weight criteria relative to their association with quality. Judges place disproportionate weight on the perceived investability of the venture, its execution speed, competitive advantage, motivation, and business model, criteria whose association with quality is weak and, for several, of the opposite sign. By contrast, they place comparatively little weight on several criteria that are more strongly associated with quality, including expected sales volume, skills and connections, customer traction, and business development status. The over-weighting is both more pronounced and statistically sharper than the under-weighting: the gap between recommendation weight and quality association is significant only among the over-weighted criteria. One interpretation is that judges are overconfident in the predictive value of their own assessments, consistent with broader evidence that individuals hold overly narrow confidence intervals and overstate their relative accuracy (Moore & Healy, 2008; Moore & Schatz, 2017).

One reason for these results might be that the judges may not understand the signals of venture quality that the application contains. The judges' beliefs about the criteria will then be noisy predictors, something often noted (Kahneman et al., 2021), especially in complex decision environments (Enke, 2024; Enke & Graeber, 2023; Enke & Zimmermann, 2019). In that case, neither the criteria nor an intuitive summary judgment based on them will be predictive of future venture success. Our results showing low classification accuracies of both these sets of beliefs point in this direction.

Alternatively, the judges may assess the criteria appropriately, but lack calibration when intuitively combining them into an overall judgment, which would make their recommendations noisier classifiers than a model that weights the criteria to optimize prediction of quality. We do not find that there is much difference in predictive accuracy between neither the recommendations nor a prediction model of venture quality based on the criteria, nor from the venture attributes themselves, so that supposition does not find strong support.

A final possibility is that judges combine their belief in a systematically distorted way relative to a model that weighs them to optimize prediction of quality. There is evidence of such patterns in other judgment settings, summarized for example in (Dawes et al., 1989; Kahneman, 2011; Kahneman et al., 2021). Consistent with this, we observe an asymmetric weighting of the evaluation criteria: a majority receive substantially greater weight in judges' decisions than their association with venture quality would warrant, whereas the underweighted criteria deviate from proportional weighting by comparatively little. This asymmetry suggests an optimistic tendency, although these distortions largely offset one another in the aggregate, leaving overall predictive accuracy close to that of an optimally weighted model. Our analyses do not allow us to determine precisely why this pattern of misweighting emerges.

Our findings further reveal that judges are relatively more effective at screening out weak ventures than at identifying the strongest opportunities. This may in part reflect the lower fraction of high-quality ventures they have observed in the past; thinner experience with such ventures would yield noisier predictions for those types (Dahlin et al., 2018; Thompson, 2010). The conservative stance is reinforced with experience: more experienced judges display higher accuracy, driven mainly by better rejection of low-quality projects, while their identification of top

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ventures does not show a comparable association with experience. Recommendation rates are lower among judges with more accumulated evaluations. This pattern is mirrored in the composition of the recommendation. With experience, the share of the recommendation's association with venture quality that operates orthogonally to the linear projection of criteria grades shifts from roughly 0% among the least experienced judges to about 15% among the most experienced.

Although balanced classification accuracy is only moderate, the objective of the evaluation under study is primarily to screen out weaker applications rather than to identify the highest-potential ventures with precision. Applicants advancing beyond this stage are evaluated through subsequent ranking and interview processes that provide richer information for distinguishing among ventures near the admission margin.

Our exercises are associative. The models we estimate are reduced form ordinary least squares linear probability models. The paper observes assessments and recommendations linked through statistical associations, not the mental models used by each judge: we do not know and cannot infer which mental model each judge uses from the observed correlations. It may, for example, be possible that after reading an application, a specific judge decides whether to recommend or reject a venture, and then fills in grades on the criteria to be consistent with this judgment. Other judges may use other mental models. But the goal of this article is not to estimate the mental models of individual judges, but to ask whether judges' beliefs, as recorded in the evaluation criteria, are informative of their recommendations and useful for predicting venture quality; how much they add to the information already available about ventures for that prediction; whether they are used optimally in forming recommendations; and how these patterns vary across judges with different levels of total evaluations. To that end, the associations are useful and paint a consistent story.

Prior research suggests substantial heterogeneity in the predictive accuracy of venture evaluators across settings. Judges evaluating business plans have been found to possess little or no ability to predict subsequent venture success in some contexts (McKenzie & Sansone, 2019), while in others their assessments remain predictive even after controlling for observable venture characteristics (Fafchamps & Woodruff, 2017). Evaluators using only brief venture abstracts can identify high-potential deep-tech ventures but appear less effective when assessing SaaS and consumer-product ventures (Scott et al., 2020). In contrast, evaluators of inventions have demonstrated relatively high accuracy in forecasting commercialization outcomes (Åstebro & Elhedhli, 2006), and judges in pitch competitions also generate information that predicts subsequent venture performance (S. T. Howell, 2021). Taken together, these findings suggest that the accuracy of venture evaluation depends critically on the evaluation context, the information available to evaluators, and the nature of the opportunities being assessed. Despite growing interest in entrepreneurial screening and selection, our understanding of how evaluators form judgments, weight available signals, and translate those beliefs into predictions remains limited. The field is therefore still at an early stage of developing a systematic understanding of when, how, and why human evaluators successfully identify entrepreneurial potential.

Taken together, our results show that predicting venture quality is difficult both for humans and for models estimated on humans' beliefs. Judges with more experience are better, but the gains are modest and concentrated in avoiding poor ventures rather than in identifying the best ones. The information contained in the criteria grades is limited, and judges' use of it departs systematically from optimal weighting in some respects, even though these departures largely offset one another in aggregate. Accuracy could be improved at the margin if less experienced judges learned from their more seasoned peers, if evaluators recalibrated the weights they assign to different criteria, or if they could observe more applications with high-quality ventures to train on, but such gains are unlikely to be large. One alternative is to train judges to make more calibrated decisions through managerial interventions. Such training has proved feasible but with rather limited effects (Moore et al., 2017). Training entrepreneurs in the scientific method has also shown some positive results (Camuffo et al., 2024).

Our results are broadly consistent with past research on venture evaluation and venture classification/prediction by humans that suggests that this is indeed a very difficult task (Astebro, 2004; Gompers et al., 2020; Zacharakis & Shepherd, 2001). The results point to humans as biased and noisy forecasters that perform as well as a linear additive

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

prediction model using optimally the information the judges use to form their beliefs. A more radical alternative would be to deploy large language models tuned to the decision context to assess applications. Evaluating the feasibility and effectiveness of such approaches is left for future research.

ACKNOWLEDGMENTS

We thank feedback from Joshua Gans, Fabian Gaessler, Scott Stern, and participants at the second Bayesian Entrepreneurship Conference and the Barcelona Summer Forum, Entrepreneurship workshop. During the preparation of this work, the authors used ChatGPT and Claude in order to improve wording and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article. Open access publication funding provided by COUPERIN CY26.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Thomas Åstebro https://orcid.org/0000-0002-8077-7631
Andrew Funck https://orcid.org/0009-0007-0945-8674

ENDNOTES

[1] Ventures refer to businesses in the process of being created.

[2] For theory see Aragones et al. (2005).

[3] We do not observe the process by which judges form or update these beliefs, nor the mental model that maps the criteria grades into their overall recommendation. We observe recorded grades and the final recommendation and measure the reduced-form correlations between these measures.

[4] Several articles have further developed the method of using defunct web activity as a measure of venture failure, and we follow this practice. See for example S. T. Howell (2017), Kerr et al. (2014), and Signori and Vismara (2018).

[5] The recommendation-only model's low AUC reflects a leave-one-venture-out systematic bias in single-predictor binary group-mean estimators, not weak discrimination.

REFERENCES

Aragones, E., Gilboa, I., Postlewaite, A., & Schmeidler, D. (2005). Fact-free learning. *American Economic Review*, *95*(5), 1355–1368.

Astebro, T. (2004). Key success factors for technological entrepreneurs' R&D projects. *IEEE Transactions on Engineering Management*, *51*(3), 314–321.

Åstebro, T., & Elhedhli, S. (2006). The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science*, *52*(3), 395–409.

Åstebro, T., & Koehler, D. J. (2007). Calibration accuracy of a judgmental process that predicts the commercial success of new product ideas. *Journal of Behavioral Decision Making*, *20*(4), 381–403.

Åstebro, T., & Michela, J. L. (2005). Predictors of the survival of innovations. *Journal of Product Innovation Management*, *22*(4), 322–335.

Avnimelech, G., Dushnitsky, G., Ellsaesser, F., & Fitza, M. (2025). Are accelerators akin to breweries or wineries? A Bayesian variance decomposition of accelerator and cohort effects. *Strategic Management Journal*, *46*(2), 534–579.

Camuffo, A., Gambardella, A., Messinese, D., Novelli, E., Paolucci, E., & Spina, C. (2024). A scientific approach to entrepreneurial decision-making: Large-scale replication and extension. *Strategic Management Journal*, *45*(6), 1209–1237.

Chu, J. Y., Voelkel, J. G., Stagnaro, M. N., Kang, S., Druckman, J. N., Rand, D. G., & Willer, R. (2024). Academics are more specific, and practitioners more sensitive, in forecasting interventions to strengthen democratic attitudes. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(3), e2307008121.

Cohen, S., Fehder, D. C., Hochberg, Y. V., & Murray, F. (2019). The design of startup accelerators. *Research Policy*, *48*(7), 1781–1797.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Cohen, S., & Hochberg, Y. V. (2014). *Accelerating startups: The seed accelerator phenomenon*. SSRN working paper.
D'Acunto, F., Hoang, D., Paloviita, M., & Weber, M. (2023). IQ, expectations, and choice. *Review of Economic Studies*, *90*(5), 2292–2325.
Dahlin, K. B., Chuang, Y.-T., & Roulet, T. J. (2018). Opportunity, motivation, and ability to learn from failures and errors: Review, synthesis, and ways to move forward. *Academy of Management Annals*, *12*(1), 252–277.
Dawes, R. M. (1975). Graduate admission variables and future success: We cannot tell whether the standard selection measures used by graduate schools are valid. *Science*, *187*(4178), 721–723.
Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674.
DellaVigna, S., & Pope, D. (2018). Predicting experimental results: Who knows what? *Journal of Political Economy*, *126*(6), 2410–2456.
Elhedhli, S., Akdemir, C., & Åstebro, T. (2014). Classification models via Tabu search: An application to early stage venture classification. *Expert Systems with Applications*, *41*(18), 8085–8091.
Enke, B. (2024). *The cognitive turn in behavioral economics*. Working paper.
Enke, B., & Graeber, T. (2023). Cognitive uncertainty. *The Quarterly Journal of Economics*, *138*(4), 2021–2067.
Enke, B., & Zimmermann, F. (2019). Correlation neglect in belief formation. *The Review of Economic Studies*, *86*(1), 313–332.
Fafchamps, M., & Woodruff, C. (2017). Identifying gazelles: Expert panels vs. surveys as a means to identify firms with rapid growth potential. *The World Bank Economic Review*, *31*(3), 670–686.
Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*(3), 288–299.
Gillen, B., Snowberg, E., & Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, *127*(4), 1826–1863.
Gius, L. (2025). Disagreement predicts startup success: Evidence from venture competitions. *Strategy Science*, *10*(2), 93–108.
Gompers, P. A., Gornall, W., Kaplan, S. N., & Strebulaev, I. A. (2020). How do venture capitalists make decisions? *Journal of Financial Economics*, *135*(1), 169–190.
Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.
Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impression-istic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323.
Howell, S. T. (2017). Financing innovation: Evidence from R&D grants. *American Economic Review*, *107*(4), 1136–1164.
Howell, S. T. (2020). Reducing information frictions in venture capital: The role of new venture competitions. *Journal of Financial Economics*, *136*(3), 676–694.
Howell, S. T. (2021). Learning from feedback: Evidence from new ventures. *Review of Finance*, *25*(3), 595–627.
Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515.
Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.
Kalvapalle, S. G., Phillips, N., & Cornelissen, J. (2024). Entrepreneurial pitching: A critical review and integrative framework. *Academy of Management Annals*, *18*(2), 550–599.
Kanze, D., Huang, L., Conley, M. A., & Higgins, E. T. (2018). We ask men to win and women not to lose: Closing the gender gap in startup funding. *Academy of Management Journal*, *61*(2), 586–614.
Kerr, W. R., Lerner, J., & Schoar, A. (2014). The consequences of entrepreneurial finance: Evidence from angel financings. *The Review of Financial Studies*, *27*(1), 20–55.
Kerr, W. R., & Nanda, R. (2015). Financing innovation. *Annual Review of Financial Economics*, *7*, 445–462.
Lane, J., & Rietzler, N. (2026). *Designing evaluation panels: When does professional panel diversity improve startup selection?* (Tech. Rep.).
Lerner, J., & Malmendier, U. (2013). With a little help from my (random) friends: Success and failure in post-business school entrepreneurship. *The Review of Financial Studies*, *26*(10), 2411–2452.
Lin, C., Spens, R., Wagh, P., Wang, P.-H., Lane, J. N., Boussioux, L., Ayoubi, C., & Chen, Y. H. (2024). Narrative AI and the human-AI oversight paradox in evaluating early-stage innovations. Working Paper.
Lyonnet, V., & Stern, L. H. (2024). Machine learning about venture capital choices. Working paper.
Mazur, J. E., & Hastie, R. (1978). Learning as accumulation: A reexamination of the learning curve. *Psychological Bulletin*, *85*(6), 1256–1274.
McKenzie, D., & Sansone, D. (2019). Predicting entrepreneurial success is hard: Evidence from a business plan competition in Nigeria. *Journal of Development Economics*, *141*, 102369.
Molavi, P., Tahbaz-Salehi, A., & Vedolin, A. (2023). Model complexity, expectations, and asset prices. *The Review of Economic Studies*, *91*(4), 2462–2507.
Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/smj.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Moore, D. A., & Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass*, *11*(8), e12331.

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J., & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, *63*(11), 3552–3565.

Moritz, A., Diegel, W., Block, J., & Fisch, C. (2022). VC investors' venture screening: The role of the decision maker's education and experience. *Journal of Business Economics*, *92*(1), 27–63.

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association*, *79*(387), 489–500.

Scott, E. L., Shu, P., & Lubynsky, R. M. (2020). Entrepreneurial uncertainty and expert evaluation: An empirical analysis. *Management Science*, *66*(3), 1278–1299.

Sharapov, D., & Dahlander, L. (2025). Selection regimes and selection errors. *Organization Science*, *36*, 2324–2348.

Signori, A., & Vismara, S. (2018). Does success bring success? The post-offering lives of equity-crowdfunded firms. *Journal of Corporate Finance*, *50*, 575–591.

Thompson, P. (2010). Learning by doing. In *Handbook of the economics of innovation* (Vol. 1, pp. 429–476). Elsevier.

Waldman, J. D., Yourstone, S. A., & Smith, H. L. (2003). Learning curves in health care. *Health Care Management Review*, *28*(1), 41–54.

Zacharakis, A. L., & Shepherd, D. A. (2001). The nature of information and overconfidence on venture capitalists' decision making. *Journal of Business Venturing*, *16*(4), 311–332.

How to cite this article: Åstebro, T., Funck, A., & Giordano, F. (2026). Decomposing venture evaluation: Learning about belief formation. *Strategic Entrepreneurship Journal*, 1–36. https://doi.org/10.1002/sej.70042

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX A: ADDITIONAL TABLES

TABLE A1 Ventures survival rates, FTE, and funding.

Batch	N ventures	Survival Mean	FTE		€Mln funding		
			Mean	$P75$	$p > 0$	Mean| > 0	$P75$| > 0
2021 Fall	62	56%	10.9	15.5	11%	13	6.2
2021 Summer	49	71%	10.3	12	12%	12.7	2.9
2022 Spring	60	82%	5.9	9	15%	0.7	1
2022 Summer	96	64%	8.2	9	12%	1.2	1.6
2022 Winter	66	74%	10.2	13	23%	2.3	1.7
2023 Spring	72	85%	4.8	6	4%	0.4	0.6
2023 Summer	66	76%	3.9	5	9%	0.7	1
2023 Winter	54	81%	3.5	4	7%	0.6	0.7
2024 Winter	54	72%	3.8	5	7%	1	1
All	579	73%	6.7	8	11%	3.5	2

Note: The table presents summary statistics for all ventures evaluated across all batches in the dataset. Column 2 reports the number of ventures evaluated; Column 3 shows the survival rate, that is, the percentage of ventures in each batch still active as of Summer 2024. Columns 4 and 5 show the average and the 75th percentile of FTEs among surviving ventures. Columns 6–8 provide metrics on post-application funding in millions of euros: the probability of receiving any funding post-application, the mean funding amount if received, and the funding amount for the top 25% by post-application funding in the batch.

TABLE A2 Robustness analysis of prediction performance at evaluation level for trinary recommendation and alternative classification of missing values.

Predictors	Description	Baseline (N=1638)		Dropped NA (N=1389)		Y trinary (N=1638)	
		Bal. Acc.	AUC	Bal. Acc.	AUC	Bal. Acc.	AUC
–	*Random classifier*	*0.505*	*0.500*	*0.502*	*0.500*	*0.505*	*0.500*
Y	Admission recommendation	0.619	0.376	0.606	0.363	0.619	0.477
X	Criteria grades	0.619	0.649	0.599	0.622	0.619	0.649
$Y^{\perp} + \widehat{Y}$	Recommendation decomposition	0.618	0.639	0.606	0.615	0.614	0.641
$Y + X$	Recommendation + criteria	0.627	0.656	0.612	0.630	0.630	0.657
$X + Y^{\perp}$	Criteria + orthogonal recommendation	0.627	0.656	0.612	0.630	0.628	0.657
Z	Venture attributes	0.612	0.652	0.562	0.585	0.612	0.652
$X + Z$	Criteria + venture attributes	0.616	0.665	0.576	0.606	0.616	0.665
$Z + Y$	Venture attributes + recommendation	0.617	0.675	0.575	0.616	0.628	0.674
$Z + X + Y^{\perp}$	Venture attributes + criteria + orthogonal recommendation	0.628	0.674	0.577	0.617	0.622	0.673

Note: The table reports robustness checks for the cross validation out-of-sample classification and scoring performance of the different models at the evaluation level. The outcome variable is a binary indicator of venture quality, and the first column lists the regressors included in the OLS prediction model. All models are assessed using leave-one-venture-out cross-validation: all evaluations of a held-out venture are removed before estimation, and the held-out evaluations are then classified according to the empirical prior. The table reports out-of-sample balanced accuracy and the area under the receiver operating characteristic curve (AUC) across models, under two alternative specifications that each perturb a different object. The "Dropped NA" specification changes the construction of the quality label V: rather than imputing ventures with missing information as closed (as in the baseline), it drops these observations from the sample, yielding a smaller estimation sample. The "Y trinary" specification instead changes the recommendation predictor Y: rather than treating recommendations as binary (as in the baseline), it adds a third category for intermediate, uncertain recommendations. Both are compared with the baseline specification. Values in italics refer to classification metrics of the "random classifier". They were made in italics to differentiate this classifier from all others.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE A3 ORIV criteria, judges' recommendations, and ventures' quality.

	ORIV Y_{ij}	v_i				
Model	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Variables						
ORIV Skills and connections	0.0370			0.0408*	0.0369	0.0408*
	(0.0293)			(0.0231)	(0.0228)	(0.0229)
ORIV Motivation and ambition	0.0743**			−0.0266	−0.0343	−0.0266
	(0.0327)			(0.0234)	(0.0232)	(0.0232)
ORIV Sales volume	0.0739*			0.0760***	0.0683***	0.0760***
	(0.0405)			(0.0241)	(0.0238)	(0.0236)
ORIV VC and BA	0.2440***			0.0496**	0.0242	0.0496**
	(0.0391)			(0.0240)	(0.0243)	(0.0235)
ORIV Competitive advantage	0.1050***			−0.0670***	−0.0779***	−0.0670***
	(0.0352)			(0.0217)	(0.0214)	(0.0213)
ORIV Business development status	0.0710			0.0376*	0.0302	0.0376*
	(0.0462)			(0.0215)	(0.0218)	(0.0218)
ORIV Customer traction	0.0495			0.0339	0.0287	0.0339
	(0.0386)			(0.0259)	(0.0253)	(0.0253)
ORIV User pain point	−0.0044			−0.0008	−0.0004	−0.0008
	(0.0340)			(0.0208)	(0.0205)	(0.0205)
ORIV Business model	0.1227***			−0.0077	−0.0205	−0.0077
	(0.0441)			(0.0235)	(0.0227)	(0.0225)
ORIV Understanding of competition	0.0524*			0.0096	0.0042	0.0096
	(0.0295)			(0.0208)	(0.0204)	(0.0204)
ORIV Regulation	0.0515**			0.0061	0.0007	0.0061
	(0.0259)			(0.0194)	(0.0193)	(0.0193)
ORIV Tech product development status	0.0403			0.0092	0.0050	0.0092
	(0.0367)			(0.0240)	(0.0238)	(0.0237)
ORIV Execution speed	0.1352***			0.0299	0.0158	0.0299
	(0.0372)			(0.0233)	(0.0232)	(0.0228)
ORIV Finance	−0.0245			−0.0109	−0.0083	−0.0109
	(0.0254)			(0.0173)	(0.0175)	(0.0175)
ORIV Y_{ij}		0.1519***			0.1040***	
		(0.0183)			(0.0254)	
ORIV $\hat{Y}_{ij}$			0.1882***			
			(0.0227)			
ORIV $Y_{ij}^{\perp}$			0.1040***			0.1040***
			(0.0274)			(0.0254)

(Continues)

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Eesc Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE A3 (Continued)

	ORIV Y_{ij}	v_i				
Model	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Fit statistics						
Observations	1638	1638	1638	1638	1638	1638
R2	0.56869	0.09436	0.10146	0.12068	0.13977	0.13977

Note: This table reports OLS estimates from the following specifications controlling for measurement error in criteria ratings (X_{ijk}) and recommendation decisions (Y_{ij}) using the Obviously Related Instrumental Variables (ORIV) approach. Column 1 projects the judge's recommendation onto evaluation criteria: $Y_{ij}^{\text{ORIV}} = \alpha + \Sigma_k \gamma_k X_{ijk}^{\text{ORIV}} + \epsilon_{ij}$. Columns 2–6 regress venture quality v_i on different combinations of recommendation components and criteria scores: Column 2, $v_i = \alpha + \delta Y_{ij}^{\text{ORIV}} + \epsilon_{ij}$; Column 3, $v_i = \alpha + \delta_1 \widehat{Y}_{ij}^{\text{ORIV}} + \delta_2 Y_{ij}^{\perp\,\text{ORIV}} + \epsilon_{ij}$; Column 4, $v_i = \alpha + \Sigma_k \beta_k X_{ijk}^{\text{ORIV}} + \epsilon_{ij}$; Column 5, $v_i = \alpha + \delta Y_{ij}^{\text{ORIV}} + \Sigma_k \theta_k X_{ijk}^{\text{ORIV}} + \epsilon_{ij}$; Column 6, $v_i = \alpha + \delta_2 Y_{ij}^{\perp\,\text{ORIV}} + \Sigma_k \beta_k X_{ijk}^{\text{ORIV}} + \epsilon_{ij}$. Y_{ij}^{ORIV} is judge j's binary recommendation for venture i, v_i is a binary indicator of venture quality, and X^{ORIV} are standardized criterion scores. $\widehat{Y}_{ij}^{\text{ORIV}}$ and $Y_{ij}^{\perp\,\text{ORIV}}$ denote the fitted value and residual from Column 1, representing the criteria-predictable and criteria-orthogonal components of the recommendation. By the Frisch–Waugh–Lovell theorem, $\beta_k = \delta\gamma_k + \theta_k$: each criterion's total correlation with quality (Column 4) decomposes into a share carried by the recommendation and a direct residual association. The reported R^2 is the adjusted R^2, which penalizes the inclusion of additional regressors and so does not mechanically increase with the number of explanatory variables, allowing comparison of fit across specifications of differing dimension. Standard errors clustered at the venture level in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

TABLE A4 ORIV prediction performance at evaluation level.

			Classification				Scoring	
Predictors	**Description**	*N*	**Accuracy**	**Bal. Acc.**	**Sensitivity**	**Specificity**	**Brier**	**AUC**
–	*Random classifier*	*1638*	*0.513*	*0.503*	*0.434*	*0.571*	*0.244*	*0.500*
Y^{ORIV}	Admission recommendation^{ORIV}	1638	0.618	0.618	0.618	0.619	0.223	0.600
X^{ORIV}	Criteria grades^{ORIV}	1638	0.627	0.629	0.644	0.615	0.223	0.673
$Y^{\perp\,ORIV} + \hat{Y}^{ORIV}$	Recommendation^{ORIV} decomposition	1638	0.635	0.633	0.624	0.643	0.222	0.673
$Y^{ORIV} + X^{ORIV}$	Recommendation^{ORIV} + criteria^{ORIV}	1638	0.645	0.646	0.648	0.643	0.219	0.690
$X^{ORIV} + Y^{\perp\,ORIV}$	Criteria^{ORIV} + ort. recommendation^{ORIV}	1638	0.644	0.644	0.647	0.642	0.219	0.690

Note: This table reports out-of-sample cross-validation classification and scoring performance at the evaluation level across prediction models controlling for measurement error in criteria ratings and recommendation decisions using the Obviously Related Instrumental Variables (ORIV) approach. The outcome is a binary indicator of venture success, and the first column lists the regressors entered in the OLS prediction model. All models are assessed by leave-one-venture-out cross-validation: every evaluation of a held-out venture is removed before estimation, and the held-out evaluations are then classified against the empirical prior. Classification metrics comprise Accuracy, Balanced Accuracy, Sensitivity (the true-positive rate), and Specificity (the true-negative rate); scoring metrics comprise the Brier score and the area under the receiver operating characteristic curve (AUC). The first row reports a random classifier drawing at the empirical prior, which serves as a benchmark. Predictors are denoted Z (venture attributes), X^{ORIV} (fitted values of criteria grades in a first stage using peer judges as instruments), and Y^{ORIV} (fitted values of admission recommendations in a first stage using peer judges as instruments), with $Y^{\perp ORIV}$ marking a component orthogonal to the linear projection of criteria grades ($\hat{Y}^{ORIV}$). Higher values indicate better performance for all metrics except the Brier score, for which lower is better. Values in italics refer to classification metrics of the "random classifier". They were made in italics to differentiate this classifier from all others.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

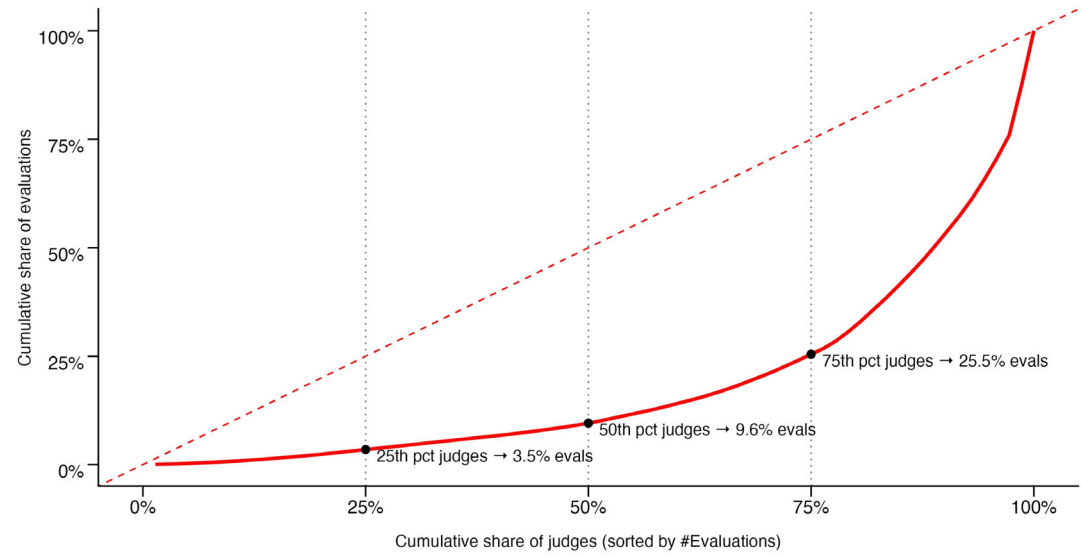
TABLE A5 Prediction performance at venture level.

			Classification				Scoring	
Predictors	**Description**	*N*	**Accuracy**	**Bal. Acc.**	**Sensitivity**	**Specificity**	**Brier**	**AUC**
–	*Random classifier*	*579*	*0.509*	*0.508*	*0.498*	*0.518*	*0.243*	*0.500*
Y	Admission recommendation	579	0.660	0.656	0.636	0.676	0.224	0.583
	Criteria grades	579	0.639	0.653	0.736	0.571	0.222	0.675
$Y^{\perp} + \widehat{Y}$	Recommendation decomp.	579	0.648	0.650	0.661	0.638	0.222	0.679
$Y + X$	Recommendation + criteria	579	0.648	0.658	0.720	0.597	0.219	0.688
$X + Y^{\perp}$	Criteria + ort. recommendation	579	0.648	0.658	0.720	0.597	0.219	0.688
Z	Ventures' attributes	579	0.608	0.608	0.611	0.606	0.233	0.645
$X + Z$	Criteria + ventures' attributes	579	0.625	0.631	0.661	0.600	0.227	0.665
$Z + Y$	Ventures' attributes + recommendation	579	0.630	0.633	0.649	0.618	0.225	0.674
$Z + X + Y^{\perp}$	Ventures' attributes + criteria + ort. recommendation	579	0.630	0.637	0.674	0.600	0.224	0.676

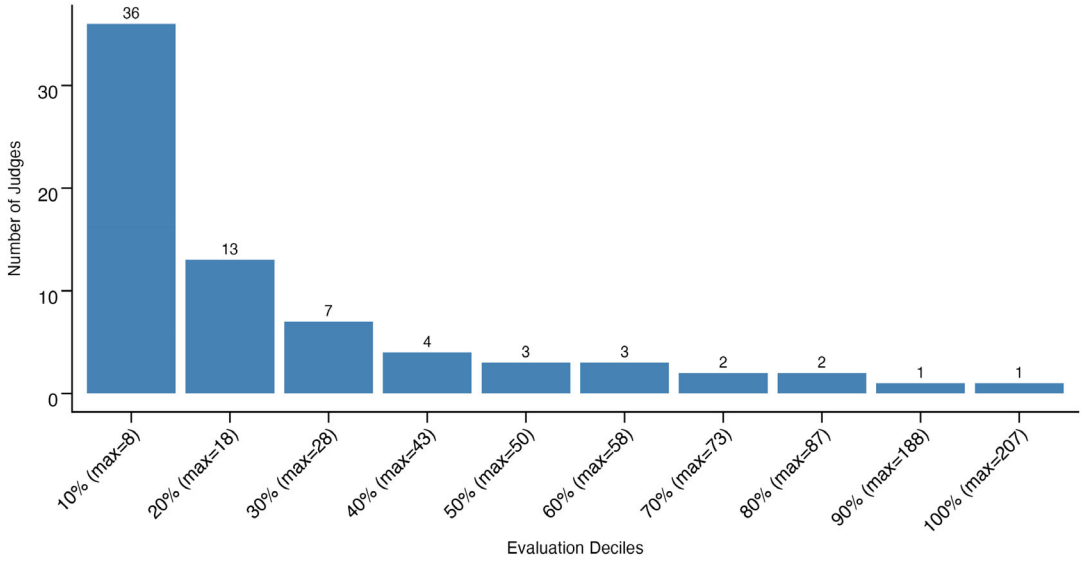
Note: The table reports cross-validation out-of-sample classification and scoring performance at the venture level across models. The outcome variable is a binary indicator of venture quality, and the first column lists the regressors included in the OLS prediction model. All models are assessed using leave-one-venture-out cross-validation: all evaluations of a held-out venture are removed before estimation, and the held-out evaluations are then scored using the model fitted on the remaining ventures. The evaluation-level predictions are aggregated to the venture level by averaging the predicted scores across a venture's evaluations, and the resulting venture-level score is classified according to the empirical prior. The table presents out-of-sample classification metrics, including Accuracy, Balanced Accuracy, Sensitivity (i.e., the true-positive rate), and Specificity (i.e., the true-negative rate), as well as scoring metrics, namely the Brier Score and the Area Under the Receiver Operating Characteristic Curve (AUC). Values in italics refer to classification metrics of the "random classifier". They were made in italics to differentiate this classifier from all others.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX B: ADDITIONAL FIGURES



(a) Lorenz Curve



(b) Number of Judges by experience

FIGURE B1 Judges' experience. The figures illustrate the distribution of evaluation activity across judges. (a) The cumulative share of evaluations as a function of the cumulative share of judges, sorted by experience. (b) The number of judges in each decile of the distribution of evaluations per judge, where each decile corresponds to successive shares of total evaluations, ordering judges by the number of evaluations performed. Overall, the results show that a large number of judges conduct only a few evaluations, while a small group of highly active judges account for the majority of evaluations.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX C: EVALUATION CRITERIA

1. Does this team have the necessary skills and connections?
2. Does this team have the sufficient motivation and ambition?
3. Does this project address a concrete user pain point?
4. Does this project have a viable business model?
5. Is the projected sales volume going to be excellent?
6. Would a VC or BA invest in this project?
7. Does the team have a clear understanding of competition?
8. Does this project have a clear competitive advantage based on some key differentiating factor?
9. Does this project address current or expected regulation affecting the business opportunity?
10. What is their technical product development status today?
11. What is their business development status today? (Partners, supply-chain, production, operations.)
12. Have the founders demonstrated that they can execute at high speed?
13. Has the project gained any customer traction at this time?
14. Does the project have the necessary financing in place for the next 12 months?

APPENDIX D: CLASSIFIER WITH VENTURES' ATTRIBUTES

This Appendix documents how venture attributes are encoded, how they are filtered and standardized, how the LASSO pre-selection step is nested inside the leave-one-venture-out (LOVO) cross-validation used to evaluate all predictive models, how the models are estimated and classified, and how selection frequency varies across folds.

Sources and encoding of candidate attributes

Candidate attributes are drawn from the incubator's application form. Free-text and bracketed survey responses are first mapped to numeric codes (ordinal brackets to integer codes; multi-select fields to indicator columns; free-text fields such as degree or company registration date to hand-reviewed or regular-expression-extracted values). The full candidate pool comprises 23 attributes, split into a *categorical pool* (14 attributes) and a *numeric-capable pool* (9 attributes).

Categorical pool. Table D1 lists the 14 categorical candidates and their encoding. When estimating classification models, every categorical attribute is dummy-coded.

Numeric-capable pool. Table D2 lists the nine numeric-capable candidates. Two (years since graduation, years since registration) are continuous by construction from parsed dates. One (financing amount needed) is a continuous euro figure hand-extracted from free text. The remaining six are engineered by mapping an ordinal survey bracket to a single representative point value, so that the attribute can be entered as a discrete regressor; the exact bracket-to-value correspondence is reported in Table D3.

Pre-filtering and standardization

Before any fold is formed, near-zero-variance attributes are dropped: an attribute is removed if it has fewer than two unique values in the estimation sample, or if a single category accounts for more than 98% of observations. This step is applied once to the full sample. Attributes with more than 20 missing values are also dropped. Numeric predictors are standardized once on the full analytic sample prior to the fold-wise loop.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE D1 Categorical attribute pool: Source and encoding.

Attribute	Original response format encoding	Encoding
Gender composition of co-founders	Closed categories: male only, female only, mixed, prefer not to answer	Dummy-coded
Applicant role	Free-text job title, keyword-classified into four mutually exclusive categories: Founder/CEO, Co-Founder/Executive, Manager/Director, Other/Nonexecutive	Dummy-coded
HEC Paris graduate	Free-text university field, matched on "HEC Paris" and known variants	Dummy-coded
Degree obtained	Free-text description, classified into: high school, bachelor's, master's, MBA, EMBA, other postgraduate	Dummy-coded
Signed a formal co-founder's pact	Closed categories: yes, no	Dummy-coded
Already incorporated company type	Closed categories: yes, no	Dummy-coded
Company type	Closed categories: SAS, SASU, SARL, foreign company, association, EI/EIRL/EURL, Société Européenne, entreprise à mission, profession libérale, fondation	Dummy-coded
Months worked full-time	Four closed brackets: <6, 7–12, 13–24, >24 on project months	Dummy-coded
Business model	Closed categories: media/social network, mobile app, (e-)commerce, subscription, agency/training center, marketplace	Dummy-coded
Already has customers planned	Closed categories: meetings with prospects, nurturing leads for a trial, first revenues, revenues increasing monthly	Dummy-coded
Weekly turnover growth	Closed categories: don't know, <2%, 2–rate 5%, >5%	Dummy-coded
Conversion rate	Closed categories: don't know, <5%, 5–10%, >10%	Dummy-coded
Solution maturity	Closed categories: idea/interactive demo, prototype, MVP, full working product	Dummy-coded
Startup runway	Closed categories: don't know, <3, 3–6, 6–12, >12 months	Dummy-coded

Classifier construction: Nested LASSO selection and leave-one-venture-out estimation

We evaluate all $\boldsymbol{Z}$-based models by leave-one-venture-out (LOVO) cross-validation: for each venture in turn, we take all evaluations *except* that venture's own and use them as the training fold, holding out all of that venture's evaluations for testing ($N = 579$ folds). Within each training fold, we first carry out a variable-selection step at the venture level. We estimate an ℓ_1-penalized (LASSO) regression of the venture-quality label v on the encoded attribute matrix,

$$\widehat{\phi} = \text{argmin}_{\phi} \frac{1}{2\, n_{\text{train}}} ||v_{\text{train}} - \boldsymbol{Z}_{\text{train}}\phi||_2^2 + \lambda||\phi||_1,$$

with the penalty λ chosen by fivefold cross-validation on that training fold at the minimum-MSE value (λ_{min}). An attribute is retained if any of its dummy columns has a nonzero coefficient at λ_{min}; for categorical attributes with multiple dummies, the attribute is kept as a whole if at least one level survives. This selection is computed independently in every one of the 579 outer folds, using only that fold's training ventures, so the selected attribute set can differ across folds. The held-out venture never enters either the fivefold selection step or the training data described next.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/smj.70042 by Andrew Funck - Eesc Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE D2 Numeric-capable attribute pool.

Attribute	Construction
Years since graduation	Submission year minus cleaned graduation year (continuous)
Years since company registration	Submission year minus registration year, parsed from a free-text registration date; set to missing if not incorporated
Number of employees	Bracket-to-headcount mapping, Table D3
Team sector experience	Bracket-to-years mapping, Table D3
Customers interviewed	Bracket-to-count mapping, Table D3
Market potential	Bracket-to-euro mapping, Table D3
Turnover, last 3 months	Bracket-to-euro mapping, Table D3
Financing amount needed	Continuous euro figure, hand-extracted from free text
Funds already raised	Bracket-to-euro mapping, Table D3

TABLE D3 Bracket-to-value mappings for engineered continuous attributes.

Attribute	Original bracket	Assigned value
Number of employees	1 person/2–4/5–10/more than 10	1/3/7.5/11
Team sector experience	1–3/4–6/7+ years	2/5/7 years
Customers interviewed	0–30/30–100/100–200/>200	15/50/120/200
Market potential	<500M/500M–1B/1–5B/5–10B	100M/750M/2.5B/7.5B
Turnover, last 3 months	0/1–5K/5–10K/10–30K/30–100K/>100K €	0/2500/7500/20,000/50,000/100,000 €
Funds raised	None/<50K/50–150K/150–500K/500K–1M/ 1–5M />10M €	0/5000/50,000/150,000/500,000/ 1,000,000/10,000,000 €

TABLE D4 Share of leave-one-venture-out folds in which each attribute is selected by the fold-wise LASSO ($N = 579$ folds).

Attribute	Share of folds selected
Already incorporated	100.0%
Company type	100.0%
Already has customers	100.0%
Number of employees	100.0%
Business model	99.0%
Degree obtained	93.1%
HEC Paris graduate	1.9%
Months worked full-time on project	1.7%
Customers interviewed	20.6%
Market potential	11.6%
Founder role: Co-Founder/Executive	0.3%
Founder role: Manager/Director	0.3%

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

With the surviving attributes, we then estimate an ordinary least squares (linear probability) model of v on the entire training fold and evaluate it on the held-out venture.

Robustness check: Categorical encoding of engineered attributes

As a robustness check, we re-estimate all specifications entering the six engineered numeric-capable attributes (Table D3) as dummy-coded categorical brackets rather than as imputed within-interval point values. Out-of-sample classification accuracy is unchanged under this alternative encoding.

Selection frequency across folds

Table D4 reports, for every attribute retained in at least one of the 579 outer folds, the share of folds in which it survives LASSO selection. A core set of six attributes is retained in the overwhelming majority of folds and accounts for the predictive content reported for ***Z*** in the main text; the remaining attributes are retained only intermittently and contribute little to predictive performance.

1932443x, 0, Downloaded from https://sms.onlinelibrary.wiley.com/doi/10.1002/sej.70042 by Andrew Funck - Esec Hec Paris , Wiley Online Library on [26/08/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License