Blinded by the person? Experimental evidence from idea evaluation

Citation for the original published paper (version of record):
Dahlander, L., Thomas, A., Wallin, M. et al (2023). Blinded by the person? Experimental evidence from idea evaluation. Strategic Management Journal, 44(10): 2443-2459.
http://dx.doi.org/10.1002/smj.3501

Received: 11 June 2020 | Revised: 1 March 2023 | Accepted: 6 March 2023

DOI: 10.1002/smj.3501

SHORT ARTICLE

STRATEGIC MANAGEMENT JOURNAL WILEY

Blinded by the person? Experimental evidence from idea evaluation

Linus Dahlander[1] | **Arne Thomas**[2] | **Martin W. Wallin**[3,4] | **Rebecka C. Ångström**[5]

[1]ESMT Berlin, Berlin, Germany

[2]University of Amsterdam, Amsterdam Business School, Amsterdam, The Netherlands

[3]Department of Technology Management and Economics, Chalmers University of Technology, Gothenburg, Sweden

[4]Department of Management, Technology, and Economics, ETH Zurich, Zurich, Switzerland

[5]Stockholm School of Economics, House of Innovation, Stockholm, Sweden

Correspondence
Linus Dahlander, ESMT Berlin, Schlossplatz 1, 10178 Berlin, Germany.
Email: linus.dahlander@esmt.org

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 411843692; Institute for Management of Innovation and Technology (IMIT)

Abstract

Research Summary: Seeking causal evidence on biases in idea evaluation, we conducted a field experiment in a large multinational company with two conditions: (a) blind evaluation, in which managers received no proposer information, and (b) non-blind evaluation, in which they received the proposer's name, unit, and location. To our surprise—and in contrast to the preregistered hypotheses—we found no biases against women and proposers from different units and locations, which blinding could ameliorate. Addressing challenges that remained intractable in the field experiment, we conducted an online experiment, which replicated the null findings. A final vignette study showed that people overestimated the magnitude of the biases. The studies suggest that idea evaluation can be less prone to biases than previously assumed and that evaluators separate ideas from proposers.

Managerial Summary: We wanted to find out if there were biases in the way managers evaluate ideas from their employees. We did a field experiment in a large multinational technology company where we tested two different ways of evaluating ideas: one where managers did not know anything about the person who came up with the idea and one where they

did know the person's name, which unit they worked for, and where they were located. The results were surprising. We did not find any bias against women and employees that did not work in the same location and unit as the evaluator. Managers are advised that hiding the identity of idea proposers (from idea evaluators) may not be a silver bullet to improving idea evaluation.

KEYWORDS

bias, field experiment, idea evaluation, innovation, online experiment

1 | INTRODUCTION

The literature on idea evaluation cautions that evaluators can be biased toward certain proposers, meaning that the same idea would receive different evaluation scores depending on who proposed it. Indeed, evaluators often do not base their evaluation solely on the *idea itself* but also on *whose idea* it is (Fuchs, Sting, Schlickel, & Alexy, 2019; Menon & Blount, 2003; Reitzig & Sorenson, 2013). Prior work on idea evaluation has, for example, explained how biases could arise from hierarchy (Keum & See, 2017; Schweisfurth, Schöttl, Raasch, & Zaggl, 2023), sequence (Bian, Greenberg, Li, & Wang, 2021; Criscuolo, Dahlander, Grohsjean, & Salter, 2021), and nepotism (Reitzig & Sorenson, 2013). Knowing who proposed an idea can provide important information (Chaiken, 1980; Pornpitakpan, 2004), yet relying on such source-based heuristics can lead to biases that disadvantage women and people far away from the decision-makers (Banaji & Hardin, 1996; Blair & Banaji, 1996; Stangor, Lynch, Duan, & Glas, 1992). We focus on three potential biases, namely that evaluators provide systematically lower evaluation scores to (a) female idea proposers; and to idea proposers from (b) other units, and (c) other locations.

To empirically assess whether and to what degree these biases are at play in idea evaluation, we conducted a field experiment based on a simple intervention: blinding that withholds information about the idea proposer from evaluators. Prior work has speculated that blinding is a light-touch intervention to remove biases from idea evaluation (Grohsjean, Dahlander, Salter, & Criscuolo, 2022). Blinding might mitigate evaluator biases, ensuring that ideas are evaluated on an equal footing, and has been deployed in diverse settings, such as blind auditions, blind recruitment, and (double-)blind academic peer review. To test blinding in idea evaluation, we conducted a preregistered field experiment inside a large multinational company in the information and communication technology sector. We asked innovation managers to evaluate real business ideas from our partner organization. We expected to identify biases and that blinding would reduce them. To our surprise, blinding the evaluators for the idea proposers' identity had no effect. Acknowledging the limitations of the field experiment and improving generalizability, we replicated the results in an online experiment.

2 | THEORY AND HYPOTHESES

Ideas are the seeds of innovation, but not all seeds bear fruit. It is inherently challenging to assess the potential of new ideas because they are surrounded by market and technological uncertainty. When deciding on new ideas, organizations are likely to make costly errors in the form of false positives (investing in ideas that ultimately fail) and false negatives (missing out on ideas that ultimately become a hit). Ideas' uncertainty can lead to evaluation biases distorting organizational outcomes (Criscuolo et al., 2021; Keum & See, 2017; Reitzig & Sorenson, 2013). Just as idea proposers systematically overestimate the value of their ideas (Fuchs et al., 2019), innovation managers make mistakes in evaluating ideas (Boudreau, Guinan, Lakhani, & Riedl, 2016; Criscuolo, Dahlander, Grohsjean, & Salter, 2017). One challenge for evaluators is to separate ideas from the person who generated them; some idea proposers get the benefit of the doubt, whereas others struggle to get recognized despite having a good idea. For instance, prior work suggests that women and proposers from other units and locations receive lower idea evaluations (see, e.g., Reitzig & Sorenson, 2013).

Given such biases in idea evaluation, we considered a simple intervention to hide the identity of the proposer through blinding. Work on blinding is not new. An influential study on blinding comes from Goldin and Rouse (2000), often quoted to show that the introduction of blind auditions—a blind screen between the jury and an auditioning musician—increased the admission of women to music schools.[1] Most work on blinding comes from studies in academia, where it is a common practice when evaluating papers and grant applications. Early work by Blank (1991, p. 1042) argues that double-blind reviewing in academia "minimizes undesirable referee bias." More recent work by Kolev, Fuentes-Medel, and Murray (2019) found that female authors received lower scores on their grant proposals to the Gates Foundation, even with blinding. Controlling for applicant quality and proposal text suggests biases at a fundamental level, punishing women for producing a different *type* of research rather than research of lesser quality. Evidence from academia thus shows that blinding has the potential to alleviate some biases but not necessarily all types of biases. Inspired by such work on blinding in academia, research on idea evaluation has speculated that blinding could also remove biases in corporate idea evaluation (Grohsjean et al., 2022), yet the evidence to date is scarce.

2.1 | Main effect: Blinding in idea evaluation

Blinding can affect idea evaluation because evaluators may rely on social cues about the person proposing the idea as a heuristic device (see, e.g., Gigerenzer & Gaissmaier, 2011). Heuristics provide mental shortcuts that can save effort by focusing only on the issue's most relevant aspects and ignoring other information. While effective in some regards, heuristics can bias decision-making (Tversky & Kahneman, 1974). Source-based heuristics have been investigated in research on information processing, demonstrating how a source's attributes influence how information is perceived and valued (e.g., Chaiken, 1980; Pornpitakpan, 2004).

The tendency to use information about an idea proposer as a signal for idea quality may be reinforced when evaluators lack information, expertise, or resources to assess an idea's details.

[1]After adding controls for musicians, the study yields mixed results. In some stages, women did worse in blind auditions, which is explained by a potential drop in the quality pool of female candidates after adopting blind auditions. It also shows that the results are more nuanced than often cited (see also Gelman, 2019).

Source-based evaluation heuristics can arise when evaluators have too little information *and* when they have too much. Information overload and time pressure can induce people to rely on simple heuristics (Hansen & Haas, 2001). From the evaluator's perspective, idea evaluation is both information-deprived and information-overloaded—deprived because of the small and standardized information bits that idea descriptions typically hold; overloaded because of the large number of ideas (Piezunka & Dahlander, 2015).

While the effects of blinding are contingent on *what* is blinded (elaborated in the moderation effects below), research suggests that its baseline effect is negative. For instance, acceptance rates are lower and referee reports are less favorable when academic reviewers do not know who the authors are (Blank, 1991; Okike, Hug, Kocher, & Leopold, 2016), and customers evaluated products more favorably when having identity-revealing information on the seller (Forman, Ghose, & Wiesenfeld, 2008). Such findings could be driven by (perceived or actual) selection into blinding by which worse authors or less-trustworthy sellers choose to be blinded. Similarly, the evaluator may perceive identity-revealing information on the idea proposer as a positive signal that the proposer is committed and serious about the idea. Blinding proposer information should thus lead to lower evaluation scores.

Hypothesis (H1). Innovation managers rate ideas lower in blind evaluation.

2.2 | Moderation effects: Who benefits from blinding in idea evaluation?

Studies of academic reviews and hiring decisions suggest that blinding can alter evaluations and potentially overcome biases. We focus on three characteristics that research has found important for idea evaluation: gender, same unit, and shared location (see, e.g., Criscuolo et al., 2017, Reitzig & Sorenson, 2013). Below we elaborate on our expectations of lower scores for women and higher scores for proposers from the same unit and location (as the evaluator), which would disappear if the idea proposer's identity was blinded.[2]

2.3 | Blind to help women in idea evaluation?

Much research documents a tendency to evaluate men and women differently (e.g., Brooks, Huang, Kearney, & Murray, 2014; Heilman, 2001; MacNell, Driscoll, & Hunt, 2015). Gender is a highly visible characteristic that can compensate for unobserved information (Kunda & Spencer, 2003) and a common way to classify other people that occurs almost instantaneously (Brewer & Lui, 1989; Ridgeway, 2006; Stangor et al., 1992). In this process, gender roles and stereotypes are activated, which results in cognitive bias influencing judgment and evaluation (Ridgeway, 2006). Multiple studies have demonstrated how such biases work unfavorably toward women.

These patterns are particularly strong in the technology sector, where they impair women's chances of receiving entrepreneurial funding (e.g., Kanze, Huang, Conley, & Higgins, 2018),

[2]In the pre-analysis plan, we also theorized about potential effects of evaluation order. While recent research has shown that order can affect evaluations (Bian et al., 2021; Criscuolo et al., 2021), blinding is typically not proposed to overcome them. Therefore, we present arguments on evaluation order in the online appendix (Section A1).

progressing into managerial positions (e.g., Tai & Sims, 2005), and receiving equal pay (e.g., Bamberger, Admati-Dvir, & Harel, 1995). In the technology sector, women are confronted with strong male stereotypes (Del Carpio & Guadalupe, 2022), and these gender stereotypes can create role incongruities that work against them. Overloaded with fast-paced information but lacking granular and contextual information, evaluators risk falling back to decision heuristics, thus activating gender stereotypes and providing lower scores to women. In many ways, evaluating ideas is like evaluating entrepreneurial ventures. Research consistently suggests that ventures led by women are perceived as less viable (Lee & Huang, 2018), and that female entrepreneurs are evaluated worse by angel investors (Becker-Blease & Sohl, 2007; Brooks et al., 2014), venture capitalists (Greene, Brush, Hart, & Saparito, 2001; Nelson & Levesque, 2007), and CFOs (Graham & Harvey, 2001). These inequalities arise at least partially from role incongruity between female stereotypes and the images of successful entrepreneurs, although they can be mitigated by other factors such as framing (Lee & Huang, 2018). The same "lack of fit" may handicap women in idea evaluation, where the idea proposer is often expected to develop the idea further as an intrapreneur. We thus hypothesize that women are at a disadvantage compared to men and that blinding would remove this disadvantage.

Hypothesis (H2). Innovation managers rate ideas that are proposed by female employees higher in blind evaluation.

2.4 | Blind to help people outside the unit in idea evaluation?

Scholars have repeatedly demonstrated that people are positively biased toward members of their group (see, e.g., Mullen, Brown, & Smith, 1992). One mechanism underlying in-group bias is that prior interaction increases comfort and reduces objectivity (Lawler, 1992; Zajonc, 1968). The preference for ideas from the same unit can also arise through categorization, as randomly assigning subjects to the same groups induces in-group preferences even in the absence of direct social interaction (Tajfel, Billig, Bundy, & Flament, 1971). In-group bias can arise for strategic reasons, where the evaluator looks to further his/her organizational unit. Pushing ideas from the own unit forward within the organization can bring additional resources and increase prestige, and successful ideas can help achieve the unit's strategic or business targets, which benefit the evaluator, especially if holding a managerial role. In-group bias can also arise because evaluators subconsciously prefer ideas from their unit. Evaluators likely perceive ideas from the same unit as more interesting and understandable because they have a shared understanding of key challenges, technologies, and market needs. Strategic considerations and subconscious processes can increase the scores evaluators give to ideas proposed by one of their own. In line with these arguments, Reitzig and Sorenson (2013) demonstrated that middle managers are biased in favor of ideas from their division. These findings align with the not-invented-here syndrome, in which groups believe they have a monopoly on knowledge, reject outside ideas, and promote their unit (Katz & Allen, 1982). While we expect in-group biases in idea evaluation, blinding counteracts them and thus reduces evaluation scores for ideas originating in the same unit as the evaluator.

Hypothesis (H3). Innovation managers rate ideas that are proposed by employees from the same organizational unit lower in blind evaluation.

2.5 | Blind to help people outside the location in idea evaluation?

Being in the same unit does not always imply sharing a location, which indicates that location and unit need to be analytically separated. However, similar arguments apply. A shared location typically implies a greater mutual understanding of cultural aspects and speaking the same language (metaphorically or literally). This makes ideas from colocated proposers and evaluators more relatable and accessible, reducing the cognitive burden when evaluating ideas. Moreover, identification and a sense of togetherness among colocated employees may trigger favoritism, like in-group bias. Feeling a closer emotional association with colleagues from the same location, evaluators may pay special attention to their ideas or give them the benefit of the doubt. Research on idea evaluation has found such tendencies to favor colleagues from the same location (Reitzig & Sorenson, 2013). Studies in international business have illustrated this at the firm-level, where a home country bias in R&D activities exists (Belderbos, Leten, & Suzuki, 2013). Blinding may thwart this inclination, resulting in lower evaluation scores for ideas from the same location.

> **Hypothesis (H4).** Innovation managers rate ideas that are proposed by employees located in the same country lower in blind evaluation.

3 | A FIELD EXPERIMENT IN A MULTINATIONAL COMPANY

Our field experiment tests the hypotheses in a real-world setting where managers have a stake in their decisions. We asked innovation managers at our partner company—a leading multinational company in the information and communication technology sector—to evaluate employee ideas proposed through the company's idea management system. We experimentally varied whether information on the idea proposer was blinded or not. The experiment was preregistered at the American Economic Association RCT Registry under the ID AEARCTR-0005439.[3]

3.1 | Participants

We recruited participants from the formal network of innovation managers operated by our industrial partner's internal accelerator. Like regular venture capitalists, the accelerator looks to develop employee ideas into new businesses and offers intrapreneurs funding and access to company personnel, expertise, and partnerships. The accelerator's process has several stages. Our field experiment is situated at the very beginning of that process—where innovation managers review and evaluate employees' initial ideas. The network of innovation managers consists of volunteers and all employees can apply, irrespective of position, unit, or location. Once accepted into the network, employees go through a short, formalized training on (a) user-generated, design-driven innovation, (b) the company's idea management system, and (c) business

[3]Note that we preregistered our study after the experiment started, but before we retrieved or inspected any data. The preregistration can therefore be classified as a registration prior to any human observation of the data (as defined in the OSF preregistration template, available at https://osf.io/prereg).

coaching. One task of the newly trained innovation manager is to support idea proposers in developing and improving their ideas. Most importantly for our study, the innovation managers regularly evaluate and give feedback on early ideas and recommend mature ideas for the company's internal funding process. When making a recommendation, they use the same criteria we employ in the experiment (see Section 3.5). Sixty innovation managers signed up for our experiment, thirty-eight completed it (63.3%), and eight evaluators (13.3%) started the idea evaluation but did not finish it.

Neither the innovation managers evaluating the ideas nor the employees proposing them were aware that their evaluations or ideas were part of an experiment. Instead, we communicated—in line with the messages of the internal accelerator—that the evaluations were an additional input to the company's effort to unlock the intrapreneurial spirit and improve idea evaluation. Even our contact persons were not aware of our exact research interest. We took great care not to reveal our research question or experimental manipulation to avoid experimenter demand effects (Rosenthal, 1966).

3.2 | Evaluation task

We used the survey tool Qualtrics to design an online evaluation interface, customizing its flow and visual appearance. We mirrored our industrial partner's corporate design to maximize the integrity and credibility of the online idea evaluation as an important organizational task. After a brief welcome screen, each idea was presented on an individual evaluation screen containing (in this order): a short request to evaluate the idea, information on the idea proposer depending on the treatment (see Section 3.3), the idea title, the idea description, five questions to rate the ideas (see Section 3.5), and a text field for open comments. We provide a stylized illustration of the idea evaluation screen in Figure 1 and the survey flow in Figure A1.[4]

Each innovation manager was asked to evaluate 48 ideas. The ideas came from a larger pool of ideas through the company's idea management system, an online platform where employees can submit ideas and interact with others to refine them. It is an important tool in our partner's innovation process. Ideas are evaluated regularly, and there is a budget specifically for their development. Evaluating these ideas is thus the first step toward possible larger, impactful investments down the line.

For our experimental manipulation (blinding) to work credibly, we needed early stage ideas unknown to the participants. Therefore, we considered only the 412 ideas proposed in the 6 months before our study. We left titles and descriptions unchanged.[5] In terms of content, most of the ideas were categorized under four headlines: autonomous vehicles (124 ideas), design thinking (87 ideas), logistics (86 ideas), and smart manufacturing (64 ideas). On average, the ideas had 120.24 words and received 3.21 comments. This shows the ideas had not received much attention prior to our experiment. We cannot share idea details or concrete examples because they are proprietary.

We received 1,942 idea evaluations; 1,824 (38 evaluators × 48 ideas) from innovation managers who finished the idea evaluation, and 118 from those who did not finish. We excluded

[4]Sections, tables, and figures with an "A" (e.g., Section A3, Table A1, Figure A7) are in the online appendix.

[5]In total, we retrieved 570 ideas that had been submitted between February 6, 2019, and October 7, 2019. Besides restricting the time frame (from April 8, 2019, to October 7, 2019), we took additional steps in selecting ideas. These steps are described in Section A2.

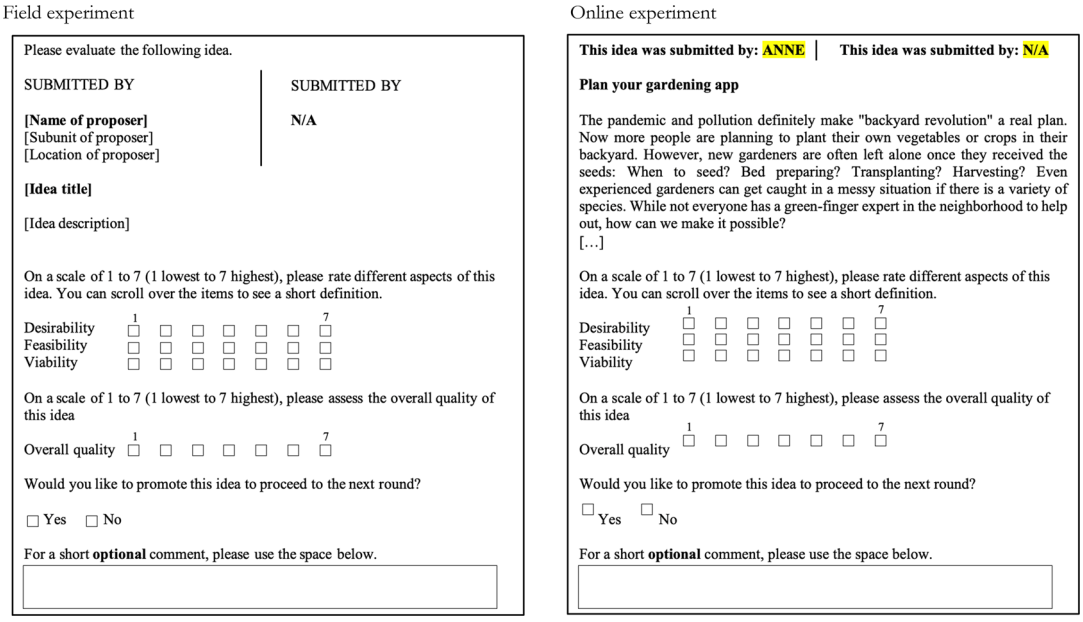

FIGURE 1 Stylized idea evaluation screen of field experiment and online experiment (non-blind condition left and blind condition right).

seven evaluations that were completed in less than 8 s. Our main sample contained 1,837 idea evaluations because in 98 cases the innovation managers did not rate the ideas on our main dependent variable.

3.3 | Treatment conditions

We used two conditions: *blind evaluation*, in which the innovation manager received no information about the idea proposer ("Submitted by: N/A"), and *non-blind evaluation*, in which the innovation manager received information about the idea proposer (name, unit, geographical location). We used a *within-subject design* in which each innovation manager evaluated ideas under both conditions.

3.4 | Randomization

Each innovation manager evaluated 48 ideas, which we randomly picked from the idea pool. To ensure each innovation manager evaluated ideas from employees with diverse backgrounds, we relied on stratified random assignment. Each idea was assigned to one of 20 strata based on the proposer's gender (2 strata) and unit (10 strata). We then randomly picked ideas from each stratum using a built-in function in Qualtrics. The number of ideas picked from each stratum was roughly proportional to the stratum size, although we oversampled small strata. After ideas were randomly picked, we randomized the evaluation order and blinded the evaluators at random.

3.5 | Variables

3.5.1 | Dependent variables

As specified in the pre-analysis plan, we used *Overall score* as our main dependent variable. For each idea, evaluators were asked: "On a scale of 1 to 7 (1 lowest to 7 highest), please assess the overall quality of the idea." In robustness checks, we used alternative-dependent variables regularly used at our partner firm (see Section A3).

3.5.2 | Treatment variable

We used an indicator variable *Blind* that switched to 1 if an evaluator evaluated an idea in the blind condition.

3.5.3 | Moderator variables

To test (H2), *Female proposer* switched to 1 if an idea was proposed by a woman. We coded gender based on name matching. If name matching did not yield a clear match, we consulted internal documents from our partnering firm to resolve ambiguity.

To test (H3), *Same unit* switched to 1 if an idea proposer was part of the same division, function, or geographical market as the innovation manager evaluating the idea. At our industrial partner, each employee belongs to one and only one division, function, or geographical market. We determined the employee's position based on an internal code, which we retrieved from the company's intranet.

To test (H4), *Same location* switched to 1 if an idea proposer was in the same country as the evaluator. We coded the location from the employees' addresses in the internal records.

4 | RESULTS

Table A1 provides an overview of the variables and their correlations. It shows that 50.2% of the ideas were rated in a blind evaluation and that 16.7, 12.9, and 9.5% came from a female proposer, the same unit, and the same location, respectively. On average, the ideas received an *Overall score* of 3.32. There was only modest consensus among the evaluators. In 20.75% of the cases, they agreed on the rating and the intraclass correlation coefficient for the ideas was 0.1496 (one-way random effects model).

As a first step, we conducted mean comparisons between the treatment conditions. Table 1 reveals that the difference was small (0.0636), and *t* tests failed to reject that blind and non-blind evaluation produce the same mean outcomes. For innovation outcomes, it is also important to consider extreme outcomes in the tails of the distribution. Figure A2 shows that the distributions of *Overall score* exhibit no clear differences between the blind and non-blind condition, either in the middle or in the tails. Overall, these first results provide no support for (H1).

To test our hypotheses more conclusively, we ran a series of ordinary least squares estimations that linked the *Overall score* that idea *i* received from innovation manager *j* to a treatment

TABLE 1 Mean comparison of treatment conditions in field and online experiment

DV: Overall score	Field experiment	Online experiment—Within design	Online experiment—Between design
Non-blind condition	3.353	4.666	4.575
Blind condition	3.289	4.608	4.587
Difference	0.0636	0.0580	−0.0123
t-statistic	0.90	1.62	−0.34
p-value	0.37	0.10	0.74
N	1,837	7,332	7,331

indicator *Blind* and fixed effects for the idea, the evaluator, and the display order. For (H2) to (H4), we also included variables indicating ideas from a *Female proposer*, the *Same unit*, and the *Same location* and a series of interaction terms with the treatment indicator. Because we could not separately blind one characteristic of an idea proposer (e.g., gender), we estimated their effects in one model[6]:

$$\begin{aligned} Evaluation\ score_{ij} = {} & \beta_0 + \beta_1 Blind_{ij} + \beta_2 Female\ proposer_i + \beta_3 Same\ subunit_{ij} \\ & + \beta_4 Same\ location_{ij} + \beta_5 Blind_{ij} * Female\ proposer_i + \beta_6 Blind_{ij} \\ & * Same\ subunit_{ij} + \beta_7 Blind_{ij} * Same\ location_{ij} + \beta_8 Idea\ FE_i \\ & + \beta_9 Evaluator\ FE_j + \beta_{10} Order\ FE_{ij} + \varepsilon_{ij}. \end{aligned}$$

Table 2 reports the regression results. Despite strong theoretical priors, we found no support for any of the hypotheses. On average, innovation managers rated the ideas only 0.0989 points lower (95% CI [−0.2241, 0.0264]) in the blind evaluation, providing no support for (H1). Regarding (H2), ideas from female proposers were rated only 0.109 points higher (95% CI [−0.1877, 0.4063]) in the blind evaluation. Regarding (H3), overall scores for ideas from the same unit exhibited practically no difference between the treatment conditions (point estimate of 0.0258; 95% CI [−0.4166, 0.4682]). Regarding (H4), ideas from the same location were rated even higher in the blind condition (by 0.145 points; 95% CI [−0.2493, 0.5387]). Overall, we found no support for our hypotheses in any of the three different analytical approaches: (a) mean comparisons and *t* tests, (b) visual inspection of distribution graphs, and (c) regression analyses.

4.1 | Post hoc analyses: Exploring and replicating the null finding

In line with *SMJ*'s guidelines, we assessed the robustness of our null finding in post hoc analyses using several alternative operationalizations, samples, and estimation techniques (see Section A4). The results hold when accounting for (a) alternative dependent variables, (b) differences in evaluation effort, (c) sample size and power, (d) idea quality, (e) distributional effects, (f) demand effects, (g) alternative time thresholds, and (h) experimental hurdles.

[6]We used the outlined model to test (H2) to (H4). To test (H1), we restricted the model to include only *Blind* and the fixed effects.

TABLE 2 Regression analyses for field and online experiment

	Field experiment		Online experiment—Within design		Online experiment—Between design	
DV: Overall score	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
H1: Blind	−0.0989 (.119)	−0.133 (.0770)	−0.0629 (.0497)		0.0124 (.717)	
Female proposer		−0.401 (.583)				
Same subunit		0.170 (.280)				
Same location		−0.235 (.211)				
H2: Blind × female proposer		0.109 (.462)				
H3: Blind × same subunit		0.0258 (.907)				
H4: Blind × same location		0.145 (.463)				
H2: Female name				0.0696 (.0877)		−0.00273 (.948)
H2: Male name				0.0562 (.129)		−0.0221 (.603)
Constant	2.696 (.000)	2.703 (.000)	4.947 (.000)	4.884 (.000)	4.860 (.000)	4.873 (.000)
Evaluator fixed effects	Yes	Yes	Yes	Yes	No	No
Idea fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Order fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	1,837	1,837	7,314	7,314	7,304	7,304
R^2	.4049	.4074	.1176	.1176	.1066	.1066

Note: Experiment and experimental design are given in the column headers. *p*-Values in parentheses are based on *SE*s clustered at the evaluator level. In the online experiment, we randomly assigned female and male names to the ideas. Therefore, we did not need to specify an interaction. Instead, the main effects of *Female name* and *Male name* give the difference between assigning female and male names and the blind condition, our baseline.

We conducted an additional online experiment on Prolific to address intractable challenges from the field experiment, detailed in Section A5. We replicated the field experiment as closely as possible, while also addressing its limitations. First, we increased the sample size and statistical power by recruiting 1,543 participants. Second, we reduced the number of ideas per evaluator to 10 and incentivized the participants with above-average compensation (yielding 14,663 evaluations). Third, we ensured high and constant idea quality by using successful ideas from crowdfunding and standardizing them in terms of length and presentation. Fourth, we replicated the within design of the field experiment and complemented it with a between design (in which evaluators evaluate only blinded or only non-blinded ideas). Fifth, we focused on the proposer's gender, the most meaningful aspect of the proposer's identity in the online setting

and the strongest effect in the field experiment. The evaluation setup, experimental conditions, and randomization closely resembled the field experiment.

We took the same steps to analyze the data as in the field experiment: mean comparisons, distribution graphs, and regression analyses. Table 1 shows that the mean differences were small (0.0580 and −0.0123) and *t* tests failed to reject that blind and non-blind evaluations produce the same mean outcomes.[7] Figure A3 shows no notable differences between blind and non-blind evaluations and between the two experimental designs. Table 2 shows that blind evaluations were slightly lower in the within design (−0.0629; 95% CI [−0.1257, −0.0001]), but there was no difference in the between design (0.0124; 95% CI [−0.0549, 0.0798]).[8] In neither design did we find differences in the evaluation scores of ideas proposed by women or men, compared to those in the blind condition. Overall, these results thus confirm the null finding from the field experiment. A final vignette study, detailed in Section A6, showed that independent participants overestimated gender bias in idea evaluation.

5 | DISCUSSION

We conducted a field experiment at one of the world's leading technology firms, where we randomly assigned innovation managers to evaluate ideas in a blind and non-blind condition. Prior research had identified biases that could distort the evaluation of ideas (e.g., Boudreau et al., 2016; Criscuolo et al., 2017; Reitzig & Sorenson, 2013) and suggested blinding as a light-touched intervention to remove such biases (Grohsjean et al., 2022). We build on work that has identified gender, organizational structure, and location as sources of bias, but found no differences between blind and non-blind evaluations. This null finding was thoroughly assessed in post hoc analyses, replicated in an online experiment, and contributes to building cumulative knowledge in strategic management (Bettis, Ethiraj, Gambardella, Helfat, & Mitchell, 2016).

A fundamental question for strategy researchers is how to allocate resources to innovation projects (see, e.g., Criscuolo et al., 2017; Klingebiel & Rammer, 2014; Reitzig & Sorenson, 2013). These decisions have large consequences for companies, prompting scholarly work on how they are made and how to keep biases out of the process. We provide experimental evidence from one of the world's leading technology firms. Indeed, the lack of causal evidence from within-company settings limits scholarly understanding of innovation. Most experimental research on biases and the evaluation of intellectual work has been conducted in *non*-corporate settings, particularly in academic settings (see, e.g., Blank, 1991) that may limit the application of prior findings. One reason for the lack of corporate field experiments is that companies are generally reluctant to grant access to internal evaluation processes and provide data on the organization's choice set of ideas. Our industrial research partner opened their doors and provided us with an unfiltered set of ideas they had not previously evaluated, preventing selective sampling and success bias.

[7]Note that *Overall score* was slightly higher in non-blind evaluation for the within design, with a marginally statistically significant difference ($p = .10$). The mean comparisons also support the appropriateness of the within design used in the field experiment, as the average evaluation scores are very similar for the two experimental designs.

[8]Although statistically significant, the main effect of blinding in the within design is small and hardly economically meaningful. Its statistical significance is largely the result of a larger sample.

Considering our experimental findings and setup, we elaborate on four plausible reasons for our null finding: (a) organizational culture, (b) selection into the experiment, (c) separation of idea from person, and (d) shifting standards from evaluation to selection.

First, the null finding might be due to organizational culture. When we shared our results with our industrial partner, some explained them with an engineering culture in which "ideas matter more than people." Our partner firm is a prestigious employer scouting globally for the best engineering talent and using standardized hiring and selection policies, which ensures fit with organizational culture and homogenous evaluators. However, the online experiment replicated the field experiment but removed the engineering culture and increased variation in evaluator quality. The null finding persisted, suggesting that the organizational culture of our industrial partner cannot explain it.

Second, selection into the experiment could lead to the null finding. Studies of hiring discrimination, for instance, have found that blinding can make it harder for members of outgroups to be hired. Blinding may prevent "positive" discrimination, in which recruiters look, for example, to increase the number of women but can no longer be more generous toward them (see, e.g., Behaghel, Crépon, & Le Barbanchon, 2015). The risk for our study is that people positively inclined to give women and members of the outgroup higher evaluations would select to be part of our experiment. However, the firm identified the group of innovation managers participating in our experiment (limiting self-selection into the experiment). There is also no opportunity to select into the online experiment based on being more lenient toward disadvantaged groups. This makes us conclude that selection into the experiment cannot explain the null finding.

Third, blinding may be more effective when the person's identity is more tied to information deemed critical for assessing ideas. Our null finding contrasts CV experiments in which women with identical CVs often receive lower evaluations than men (e.g., Petit, 2007), and evidence from entrepreneurship, where woman-led ventures are perceived as less viable (Lee & Huang, 2018) and female-backed female entrepreneurs receive lower evaluations than men (Snellman & Solal, 2023). In all these cases, the evaluation process is at least as much about the person as it is about the idea, which may differ in our setting. Our null finding may arise because the idea takes precedence over the person; the proposer's identity does not evoke information deemed critical to idea evaluation. Even when we made the proposer's gender more salient in the online experiment, we found no gender differences. Based on these findings, there are good reasons to believe that blinding is not guaranteed to improve idea evaluation. It may be ineffective if the idea proposer's identity does not unlock strong schemas that blinding could curtail.

Fourth, people may apply different standards when evaluating than when selecting ideas, and biases only manifest themselves in the selection. The "shifting standards model" in social psychology (Biernat & Manis, 1994) suggests such a difference between evaluation and selection. For instance, when evaluating job candidates, a female candidate may be seen as "good for being a woman." However, in selection decisions where candidates (or ideas) are pitted against each other as there are limited resources, there are usually greater biases (Joshi, Son, & Roh, 2015; Koch, D'Mello, & Sackett, 2015). We study the evaluation of early stage ideas, which still have a long way to go to eventual selection. The evaluators do not make the final selection and have few budget constraints, which may reduce biases. This could explain why our results differ from previous work focused on selection (e.g., Reitzig & Sorenson, 2013). This opens the question of how to design evaluation and selection, as the ultimate selection requires favorable evaluations along the way.

We invite future research to explain why our surprising null finding occurs. We acknowledge that the null finding is more robust for gender than that for organizational unit and location because we have replicated it in the online experiment. Future research on blinding is warranted to help explain our null finding. Our theorizing distinguished between strategic favoritism and subconscious preferences. Blinding can only be effective against biases caused by the withheld information. We blinded information on the idea proposer but left idea descriptions unchanged. To the degree to which biases are prompted by, for example, more interesting, exciting, and accessible idea descriptions, blinding is ineffective. Similarly, blinding could be ineffective because the evaluator has learned about an idea before, a common complication for academic peer review, where reviewers may have seen the paper presented at a conference. The same can happen with ideas from the same unit or location. Future research could further investigate such boundaries to effective blinding and the relative importance of strategic vs. subconscious biases. Given that blinding did not improve idea evaluation in our experiment, future research should investigate the costs of blinding. While blinding ideas is technically straightforward and relatively costless to implement, missing out on blocked information has opportunity costs. It reduces the potential to connect employees with similar interests and learn from what other people are working on. After all, evaluation is not an end goal for companies but one of many steps from an idea to a successful product. Blinding is potentially most helpful on a smaller scale to determine whether, how, and where biases exist before scaling any idea evaluation initiative within or across organizations.

Finally, the design and implementation of blinding may be less straightforward and more context dependent than most studies assume. Our field experiment and the follow-up online experiment are both "online," in the sense of displaying information as text-on-screen. This is a common practice to collect and assess ideas (Bayus, 2013; Beretta, 2019; Blohm, Riedl, Füller, & Leimeister, 2016; Poetz & Schreier, 2012), and future research will need to investigate potential differences between traditional idea evaluation where evaluation panels and idea proposers can meet up at the corporate headquarters and online idea evaluation—especially in terms of the schemas they unlock, and the potential to blind the information that prompts their unraveling.

ACKNOWLEDGMENTS

All authors contributed equally and are listed alphabetically. The authors are grateful for comments from Oliver Alexy, Gianluca Carnabuci, Florian Englmaier, Christoph Fuchs, Thorsten Grohsjean, Marcus Holgersson, Reddi Kotha, Tobias Kretschmer, Friederike Lenel, Markus Reitzig, David Ronayne, Ammon Salter, Henry Sauermann, Isabelle Solal, Martin Schweinsberg, Dashun Wang, and seminar participants at the Duke University field experimental workshop, Bath University, Ludwig Maximilian University of Munich, Singapore Management University, Stockholm School of Economics, TU Munich, Chalmers University of Technology, and the Imperial community. Hai Anh Vu provided outstanding research assistance. The project is funded by Deutsche Forschungsgemeinschaft (DFG) under project number 411843692 and the Institute for Management of Innovation and Technology (IMIT). The authors claim that all the errors are theirs alone. Open Access funding enabled and organized by Projekt DEAL.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data from the two online experiments is available at [https://osf.io/dkbpx/].

DATA AVAILABILITY STATEMENT

The field experiment is under NDA and we cannot share the data. The data from the online experiments can be fully shared (including the code). We would be very happy to share this data if anyone is interested!

REFERENCES

Bamberger, P., Admati-Dvir, M., & Harel, G. (1995). Gender-based wage and promotion discrimination in Israeli high-technology firms: Do unions make a difference? *Academy of Management Journal*, *38*(6), 1744–1761.

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, *7*(3), 136–141.

Bayus, B. L. (2013). Crowdsourcing new product ideas over time: An analysis of the Dell IdeaStorm community. *Management Science*, *59*(1), 226–244.

Becker-Blease, J. R., & Sohl, J. E. (2007). Do women-owned businesses have equal access to angel capital? *Journal of Business Venturing*, *22*(4), 503–521.

Behaghel, L., Crépon, B., & Le Barbanchon, T. (2015). Unintended effects of anonymous resumes. *American Economic Journal: Applied Economics*, *7*(3), 1–27.

Belderbos, R., Leten, B., & Suzuki, S. (2013). How global is R&D? Firm-level determinants of home-country bias in R&D. *Journal of International Business Studies*, *44*(8), 765–786.

Beretta, M. (2019). Idea selection in web-enabled ideation systems. *Journal of Product Innovation Management*, *36*(1), 5–23.

Bettis, R. A., Ethiraj, S., Gambardella, A., Helfat, C., & Mitchell, W. (2016). Creating repeatable cumulative knowledge in strategic management: A call for a broad and deep conversation among authors, referees, and editors. *Strategic Management Journal*, *37*(2), 257–261.

Bian, J., Greenberg, J., Li, J., & Wang, Y. (2021). Good to go first? Position effects in expert evaluation of early-stage ventures. *Management Science*, *68*(1), 300–315.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology*, *66*(1), 5–20.

Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology*, *70*(6), 1142–1163.

Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review. *American Economic Review*, *81*(5), 1041–1067.

Blohm, I., Riedl, C., Füller, J., & Leimeister, J. M. (2016). Rate or trade? Identifying winning ideas in open idea sourcing. *Information Systems Research*, *27*(1), 27–48.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science*, *62*(10), 2765–2783.

Brewer, M. B., & Lui, L. N. (1989). The primacy of age and sex in the structure of person categories. *Social Cognition*, *7*(3), 262–274.

Brooks, A. W., Huang, L., Kearney, S. W., & Murray, F. E. (2014). Investors prefer entrepreneurial ventures pitched by attractive men. *Proceedings of the National Academy of Sciences*, *111*(12), 4427–4431.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*(5), 752–766.

Criscuolo, P., Dahlander, L., Grohsjean, T., & Salter, A. (2017). Evaluating novelty: The role of panels in the selection of R&D projects. *Academy of Management Journal*, *60*(2), 433–460.

Criscuolo, P., Dahlander, L., Grohsjean, T., & Salter, A. (2021). The sequence effect in panel decisions: Evidence from the evaluation of research and development projects. *Organization Science*, *32*(4), 987–1008.

Del Carpio, L., & Guadalupe, M. (2022). More women in tech? Evidence from a field experiment addressing social identity. *Management Science*, *68*(5), 3196–3218.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, *19*(3), 291–313.

Fuchs, C., Sting, F. J., Schlickel, M., & Alexy, O. (2019). The ideator's bias: How identity-induced self-efficacy drives overestimation in employee-driven process innovation. *Academy of Management Journal*, *62*(5), 1498–1522.

Gelman, A. (2019). Did blind orchestra auditions really benefit women? Retrieved from https://statmodeling.stat.columbia.edu/2019/05/11/did-blind-orchestra-auditions-really-benefit-women/

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, *90*(4), 715–741.

Graham, J. R., & Harvey, C. R. (2001). The theory and practice of corporate finance: Evidence from the field. *Journal of Financial Economics*, *60*(2–3), 187–243.

Greene, P. G., Brush, C. G., Hart, M. M., & Saparito, P. (2001). Patterns of venture capital funding: Is gender a factor? *Venture Capital*, *3*(1), 63–83.

Grohsjean, T., Dahlander, L., Salter, A. J., & Criscuolo, P. (2022). Better ways to green-light new projects. *MIT Sloan Management Review*, *63*(2), 33–38.

Hansen, M. T., & Haas, M. R. (2001). Competing for attention in knowledge markets: Electronic document dissemination in a management consulting company. *Administrative Science Quarterly*, *46*(1), 1–28.

Heilman, M. E. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues*, *57*(4), 657–674.

Joshi, A., Son, J., & Roh, H. (2015). When can women close the gap? A meta-analytic test of sex differences in performance and rewards. *Academy of Management Journal*, *58*(5), 1516–1545.

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.

Katz, R., & Allen, T. J. (1982). Investigating the Not Invented Here (NIH) syndrome: A look at the performance, tenure, and communication patterns of 50 R&D project groups. *R&D Management*, *12*(1), 7–20.

Keum, D. D., & See, K. E. (2017). The influence of hierarchy on idea generation and selection in the innovation process. *Organization Science*, *28*(4), 653–669.

Klingebiel, R., & Rammer, C. (2014). Resource allocation strategy for innovation portfolio management. *Strategic Management Journal*, *35*(2), 246–268.

Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology*, *100*(1), 128–161.

Kolev, J., Fuentes-Medel, Y., & Murray, F. (2019). *Is blinded review enough? How gendered outcomes arise even under anonymous evaluation* (NBER Working Paper No. 25759). National Bureau of Economic Research, Cambridge, MA.

Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin*, *129*(4), 522–544.

Lawler, E. J. (1992). Affective attachments to nested groups: A choice-process theory. *American Sociological Review*, *57*(3), 327–339.

Lee, M., & Huang, L. (2018). Gender bias, social impact framing, and evaluation of entrepreneurial ventures. *Organization Science*, *29*(1), 1–16.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, *40*(4), 291–303.

Menon, T., & Blount, S. (2003). The messenger bias: A relational model of knowledge valuation. *Research in Organizational Behavior*, *25*(3), 137–186.

Mullen, B., Brown, R., & Smith, C. (1992). Ingroup bias as a function of salience, relevance, and status: An integration. *European Journal of Social Psychology*, *22*(2), 103–122.

Nelson, T., & Levesque, L. L. (2007). The status of women in corporate governance in high–growth, high–potential firms. *Entrepreneurship Theory and Practice*, *31*(2), 209–232.

Okike, K., Hug, K. T., Kocher, M. S., & Leopold, S. S. (2016). Single-blind vs double-blind peer review in the setting of author prestige. *Journal of the American Medical Association*, *316*(12), 1315–1316.

Petit, P. (2007). The effects of age and family constraints on gender hiring discrimination: A field experiment in the French financial sector. *Labour Economics*, *14*(3), 371–391.

Piezunka, H., & Dahlander, L. (2015). Distant search, narrow attention: How crowding alters organizations' filtering of suggestions in crowdsourcing. *Academy of Management Journal*, *58*(3), 856–880.

Poetz, M. K., & Schreier, M. (2012). The value of crowdsourcing: Can users really compete with professionals in generating new product ideas? *Journal of Product Innovation Management*, *29*(2), 245–256.

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, *34*(2), 243–281.

Reitzig, M., & Sorenson, O. (2013). Biases in the selection stage of bottom-up strategy formulation. *Strategic Management Journal*, *34*(7), 782–799.

Ridgeway, C. L. (2006). Gender as an organizing force in social relations: Implications for the future of inequality. In F. D. Blau, M. B. Brinton, & D. B. Grusky (Eds.), *The declining significance of gender* (pp. 265–287). New York, NY: Russell Sage.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. New York, NY: Appleton-Century-Crofts.

Schweisfurth, T. G., Schöttl, C. P., Raasch, C., & Zaggl, M. A. (2023). Distributed decision-making in the shadow of hierarchy: How hierarchical similarity biases idea evaluation. *Strategic Management Journal*, 1–28. https://doi.org/10.1002/smj.3497

Snellman, K., & Solal, I. (2023). Does investor gender matter for the success of female entrepreneurs? Gender homophily and the stigma of incompetence in entrepreneurial finance. *Organization Science*, *34*(2), 680–699.

Stangor, C., Lynch, L., Duan, C., & Glas, B. (1992). Categorization of individuals on the basis of multiple social features. *Journal of Personality and Social Psychology*, *62*(2), 207–218.

Tai, A. J. R., & Sims, R. L. (2005). The perception of the glass ceiling in high technology companies. *Journal of Leadership & Organizational Studies*, *12*(1), 16–23.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*(2, Pt. 2), 1–27.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Dahlander, L., Thomas, A., Wallin, M. W., & Ångström, R. C. (2023). Blinded by the person? Experimental evidence from idea evaluation. *Strategic Management Journal*, 1–17. https://doi.org/10.1002/smj.3501