



Organization Science

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Peer Evaluations: Evaluating and Being Evaluated

Helge Klapper, Henning Piezunka, Linus Dahlander

To cite this article:

Helge Klapper, Henning Piezunka, Linus Dahlander (2023) Peer Evaluations: Evaluating and Being Evaluated. Organization Science

Published online in Articles in Advance 08 Dec 2023

. <https://doi.org/10.1287/orsc.2021.15302>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2023 The Author(s)

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Peer Evaluations: Evaluating and Being Evaluated

Helge Klapper,^a Henning Piezunka,^{b,*} Linus Dahlander^c

^aPurdue University, West Lafayette, Indiana 47907; ^bINSEAD, 77305 Fontainebleau, France; ^cESMT Berlin, 10178 Berlin, Germany

*Corresponding author

Contact: hklapper@purdue.edu,  <https://orcid.org/0000-0002-1253-9180> (HK); henning.piezunka@insead.edu,

 <https://orcid.org/0000-0003-3956-1828> (HP); linus.dahlander@esmt.org,  <https://orcid.org/0000-0003-3527-7440> (LD)

Received: June 3, 2021

Revised: June 29, 2022; September 29, 2022;
May 2, 2023

Accepted: October 10, 2023

Published Online in *Articles in Advance*:
December 8, 2023

<https://doi.org/10.1287/orsc.2021.15302>

Copyright: © 2023 The Author(s)

Abstract. Peer evaluations place organizational members in a dual role: they evaluate their peers and are being evaluated by their peers. We theorize that when evaluating their peers, they anticipate how their evaluations will be perceived and adjust their evaluations strategically to be evaluated more positively themselves when their peers assess them. Building on this overarching claim of role duality resulting in strategic peer evaluations, we focus on a dilemma that evaluating members face: they want to leverage their evaluations of peers to portray themselves as engaged and having high standards, but at the same time, they must be careful not to offend anyone as doing so may cause retaliation. We suggest that organizational members about to be evaluated resolve this dilemma by participating in more peer evaluations but carefully targeting which evaluations they participate in. We test our theory by analyzing peer evaluations on Wikipedia, supplemented by in-depth semistructured interviews. Our study informs research on peer evaluation and organizational design by revealing how being an evaluator and evaluated can make evaluations more strategic.



Open Access Statement: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. You are free to download this work and share with others for any purpose, except commercially, if you distribute your contributions under the same license as the original, and you must attribute this work as “*Organization Science*. Copyright © 2023 The Author(s). <https://doi.org/10.1287/orsc.2021.15302>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by-nc-sa/4.0/>.”

Supplemental Material: The e-companion is available at <https://doi.org/10.1287/orsc.2021.15302>.

Keywords: organization design • crowdsourcing • peer evaluation • transparency • novel forms of organizing • Wikipedia

Evaluating organizational members is challenging yet essential (March and March 1977). Evaluations determine who climbs to the top, gets extended contracts, receives what remuneration, and obtains funds and resource allocation to their projects. Traditionally, organizations have entrusted hierarchically superior managers to evaluate people within their scope of control (Barnard 1968). There is, however, a good chance for such evaluations to be inaccurate; managers can be biased, have limited information about the person they evaluate, and lack the necessary (often local) knowledge (Aghion and Tirole 1997). Peer evaluations are an alternative to managerial evaluations that many organizations have adopted to grapple with some of these challenges.

Through peer evaluations, organizations source and aggregate the opinions of many—rather than from a single manager—when evaluating a member. As research on information aggregation shows, tapping into the knowledge of a larger set of people often leads to more accurate assessments (Christensen and Knudsen 2010, Woolley et al. 2010, Csaszar and Eggers 2013, Becker et al. 2017, Piezunka et al. 2022). While peer evaluation

constitutes a form of information aggregation, it differs from information aggregation in other contexts where organizational members evaluate some external alternative in one crucial respect: organizational members *evaluate and are evaluated*.

This role duality, specifically the outlook of being evaluated themselves, affects how members evaluate their peers. When there is at least some transparency about peer evaluations, that is, peers know how a member has evaluated others in the past, peers may take the evaluations that the member has provided of others into account when evaluating other members. They may, for example, if the focal member evaluated peers in ways they disagree with, sanction the member by giving that person a bad evaluation. Knowing that peers pay attention to how a member evaluates others, members are likely to anticipate how their evaluations will be perceived by their peers and adjust their evaluations strategically.¹ Rather than being sincere and providing merit-based evaluations, they evaluate peers in ways that are supposed to positively influence how they themselves are seen and evaluated by their peers.

Building on the overarching argument that role duality renders organizational members (more) strategic in their evaluation of peers, we theorize *how* they adjust their evaluations. We focus on the dilemma that organizational members need to appear engaged and have high standards but also need to avoid offending anyone and causing retaliation. We suggest that organizational members cope with this dilemma by evaluating more peers but that they target carefully whose evaluations they participate in. They focus on those where retaliation is unlikely and are particularly careful when evaluating peers negatively.

Testing our theory on the effect of role duality on evaluation behavior requires comprehensive data on peer evaluations. We find such data in the context of Wikipedia. Wikipedia's decision on whom to grant administrative rights to is based on peer evaluations. We examine whether and how (the outlook of) being evaluated—and thus the occupation of dual roles—affect a member's evaluations of peers. In the buildup to their own evaluations, the role duality becomes salient. In the case of Wikipedia, peer evaluations are transparent in that peers can see the complete evaluation history of a member (i.e., how and which peers a member has evaluated in the past).² For 2003–2014, we examine 10,660 evaluating members, making 187,898 peer evaluations, which we match with 116,703 comparable nonevaluations, involving 3,518 evaluation processes and 2,883 evaluated members. Our quantitative analysis is complemented by 24 in-depth semistructured interviews with Wikipedia members, giving us a deep, contextual understanding of how members think about their upcoming evaluations and how they evaluate their peers.

The findings support the overarching argument that role duality and the outlook of being evaluated render organizational members more strategic; they adjust how they evaluate their peers. In line with our theory, members who are (about to be) evaluated participate in *more* peer evaluations. They are, however, *less* likely to participate in evaluations where their evaluation may offend someone. Specifically, they are less likely to participate in evaluations where their evaluation would be pivotal and focus their negative evaluations on inactive members. Interestingly, we find no evidence that they concentrate their positive evaluations on active peers—suggesting that they try to avoid negative reciprocity but do not attempt to invoke positive reciprocity. In a post hoc analysis, we find this strategic use of peer evaluations to be effective, as it renders members more likely to be evaluated positively and to get promoted by their peers. We discuss the implications of our results for research on peer evaluations and organizational design.

We contribute to research on organizational design and nonhierarchical forms of organizing (Christensen and Knudsen 2010, Puranam et al. 2014, Lee and Edmondson 2017, Ketkar and Workiewicz 2021). We point out how the formal structure of peer evaluations

renders organizational members subject to role duality—they evaluate and are evaluated. This feature—which differentiates peer evaluations from other forms of evaluation and information aggregation—distorts members' evaluations and thus undermines a fundamental input to information aggregation: the sincerity of evaluations by organizational members (Reitzig and Maciejovsky 2015, Keum and See 2017). Our data suggest strategic behavior in peer evaluations, specifically that organizational members may (ab)use peer evaluations to portray themselves. These strategic evaluations call into question whether a key assumption of information aggregation holds in the context of peer evaluations: that aggregating the opinions of many renders decisions more accurate (Condorcet et al. 1994, Csaszar and Eggers 2013, Becker et al. 2017). More broadly, our study complements research that describes how organizational design shapes and constrains behavior (Puranam et al. 2014, Joseph and Gaba 2020, Alexy et al. 2021, Levinthal and Newark 2023, Piezunka and Schilke 2023). Specifically, we highlight how peer evaluations lead organizational members to take into account their own (future) evaluation by others when evaluating. Finally, the specific organizational design of our empirical context—that is, the organization of Wikipedia—highlights the benefits and the potential limits of transparency in organizational design (Bernstein 2012, Obloj and Zenger 2022). Transparency helps to make evaluation more accountable but allows members to use their evaluations to portray themselves.

Theory

The Goal and Challenge of Establishing Merit-Based Evaluations

A fundamental problem for any organization is how to evaluate its members (March and March 1977, Richter et al. 2023). The outcome of evaluations is important as it determines who gains authority within the organization, which guides resource allocation and strategic decision making (e.g., Lazear and Rosen 1981, Bergeron et al. 2013, Rider et al. 2023). The evaluation process is important because if organizational members (or other stakeholders) perceive it as unfair, it can cause frustration and reduced motivation (Ambrose and Cropanzano 2003, Franke et al. 2013, Vough and Caza 2017). The criteria of evaluations are important as they reward and guide organization members' behavior (Ethiraj and Levinthal 2009, Obloj and Sengul 2020).

The sought-after yet often unachieved goal is the establishment of a meritocracy where members' performances and evaluations are tightly correlated (Blau 1964, Merton and Storer 1973). In meritocracies, the allocation of (positions of) power and rewards reflects the performance of their members. Beyond the momentary benefits, establishing a meritocracy provides incentives for organizational members to perform. Building up a

meritocracy requires members' evaluations to reflect their performances. Yet, more often than not, they deviate from each other (Peter and Hull 1969, Fairburn and Malcomson 2001, Grabner and Moers 2013, Benson et al. 2019). This has led organizations to consider alternatives to the traditional way where managers evaluate their members within their scope of control.

From Managerial to Peer Evaluations

Traditionally, the evaluation of employees (and corresponding reward and promotion decisions) has been made by managers one layer up in the hierarchy. Evaluating employees has been one of managers' primary activities and constitutes the basis for determining rewards and promotions (Simon 1947, Dornbusch et al. 1975, Magee and Galinsky 2008). However, managerial evaluations have been shown to fail often to reflect employees' actual performance. Decision makers typically lack sufficient information to evaluate an organizational member accurately (Kirsch 1996, Richter et al. 2023). Therefore, they may pay attention to traits and characteristics that do not necessarily indicate a person's performance (Brivot et al. 2014, Uribe et al. 2022). Managers may also fall victim to strategic impression management of employees who, given that a single manager evaluates them, may focus their efforts on appearing to perform in the eyes of that manager (Bolino et al. 2008, Proost et al. 2010). Managers are fallible and may be biased; even in organizations that want to promote meritocracy, managers display strong biases against women and minorities (e.g., Castilla 2008, Castilla and Benard 2010, Botelho and Abraham 2017, Castilla and Ranganathan 2020, Rider et al. 2023). The risk is evident: managers may promote people who appear to perform rather than the ones that do. Follett (1941) recognized such challenges early on, pointing out that managers are charged with tasks (in our case, the evaluation of employees) that they are often not qualified to fulfill, leading to other forms of evaluating organizational members.

One solution that has gained prominence is to rely on peer evaluations, where organizational members evaluate each other. Rather than having a single upper-level manager conducting the evaluation, multiple peers evaluate the focal member. Peer evaluations are frequently deployed in flat organizations such as Valve, GitLab, Spotify, ING, and, in our case, Wikipedia, but have also become common in more traditional hierarchical organizations (e.g., as 360-degree feedback) (Waldman et al. 1998, Brett and Atwater 2001, Puranam and Håkonsson 2015). Peer evaluations are also common on an industry level when evaluations take place across organizational boundaries, for example, in creative industries or academia (Lamont 2009, Li and Agha 2015, Askin and Bothner 2016, Boudreau et al. 2016). They are common in lateral organizations such as open-source software (O'Mahony and Ferraro 2007, Leskovec et al. 2010, Shah and Nagle 2020), user communities (Shah and Nagle 2020, Sherman

and Moeen 2022), or crowdsourcing contests (Tushman et al. 2012, Bayus 2013, Piezunka and Dahlander 2015). Peer evaluations are deployed for key decisions, for example, allocating funds, determining rewards and promotions, etc. (Lamont 2012, Bernstein et al. 2014, Rossman and Schilke 2014, Obloj and Zenger 2017, Choudhury et al. 2020).

The popularity of peer evaluations is likely rooted in their potential to provide an accurate and comprehensive assessment of peers (Saavedra and Kwun 1993, Cheng and Warren 2000, Brett and Atwater 2001). Sourcing and aggregating information from multiple peers are likely to result in a more comprehensive and less error-prone—and thus more accurate—evaluation than the assessment by a single hierarchical superior (Sah and Stiglitz 1986, Knudsen and Levinthal 2007, Csaszar and Eggers 2013). Beyond a broader information base, peer evaluations may also be more robust against impression management, as it is not enough to manipulate a single decision-making manager (Hastie and Kameda 2005). In summary, peer evaluations seem like a step toward merit-based evaluations.

Strategic Evaluations

For peer evaluations to foster meritocratic decisions, they rely on individual members being sincere and merit based. Peer evaluations, however, lead to a *role duality*: organizational members *evaluate and are being evaluated*. This role duality may result in organizational members engaging in *strategic evaluations* (at a higher tendency). We define strategic evaluations as evaluations where the evaluating members are, in their choice of whom they evaluate or how they evaluate (i.e., how positive, or negative), in part instrumental in improving their own chances of being evaluated positively.

A key assumption that underlies our argument is that peers pay attention to the peer evaluations a member has provided in the past. For that, peer evaluations need to be at least partially transparent; that is, peers have some information about how an organizational member has evaluated their peers so that peers can consider them when evaluating a member (Laat 2014). While organizations vary in the degree to which information about evaluation is transparent, there is typically at least some transparency about evaluations, because transparency is either an active part of the design of an organization (Bernstein 2012, 2017) or a given standard for a field (Lamont 2009). For example, B Corps, organizations that meet high environmental and social standards, have committed to high degrees of transparency (Pollack et al. 2021). Or, for organizations where transparency is not intended, rumors and gossip still allow for information to reach many organizational members (Noon and Delbridge 1993, Michelson and Mouly 2004). When there is at least some transparency, organizational members know that others can see their past evaluations when they themselves are evaluated, and they might be judged in part based on how they have evaluated their peers in

the past, and may thus engage in (more) strategic evaluations (Bowers 2019).

Being Strategic—But How? Drawing on the central premise that role duality (in combination with transparency) influences organizational members to become more strategic in their appraisal of peers, we posit that members confront a dilemma. On the one hand, they are inclined to evaluate many peers to convey a sense of engagement. On the other hand, by evaluating a larger pool of peers, they expose themselves to the possibility of offending those who may take issue with a member's assessment of a particular peer—potentially resulting in retaliation. We hypothesize that members navigate this dilemma by increasing the number of peers they evaluate (Hypothesis 1) but carefully targeting whom they evaluate, particularly when making negative peer evaluations (Hypotheses 2 and 3).

Hypotheses

Members Increasing Their Participation in the Evaluation of Peers

Organizational members often have discretion over whether to participate in or abstain from evaluating a peer. While it is plausible that members who are (about to be) evaluated would abstain from evaluating peers to avoid taking a clear position that peers disagree with (Padgett and Ansell 1993), we suggest that organizational members participate in the evaluation of more peers when they themselves are (about to be) evaluated.

We suggest that organizational members participate in the evaluation of peers strategically—using their evaluation of peers to portray themselves as being engaged on behalf of the organization. While organizational members' participation in peer evaluations is typically not a core feature of their work and not the primary basis of their evaluation (Brett and Atwater 2001), the very fact that it is not core may explain why peers pay attention to members' participation in peer evaluations: it allows them to detect whether a member engages on behalf of the organization and is not just self-serving (Lin-Healy and Small 2013). Members who are (about to be) evaluated may thus engage in noncore activities to appear engaged (Bode et al. 2022).

Beyond the outlined incentive to increase participation in general, organizational members are likely to benefit from the *positive evaluations* they provide (Aadland et al. 2019). If they evaluate a peer positively, that peer may reciprocate when the focal member is evaluated (Gouldner 1960, Willer et al. 2012, Nai et al. 2022). We expect that they increase the number of positive evaluations.

The logic that underlies the provisions of *negative evaluations* is less established. One may expect that the role duality would cause organizational members to avoid making any negative peer evaluations, following the logic of reciprocity we have outlined for positive evaluations. We

do, however, suggest that despite the risk of negative reciprocity, members anticipating their own evaluation do *not* forgo providing negative peer evaluations. Providing negative evaluations is essential for them to portray themselves as providing merit-based evaluations. If a member only rubber-stamped peers, they would consider it a disservice to the organization as the lack of differentiation undermines the goal of merit-based evaluation. Peers may also become suspicious that an organizational member unwilling to make negative evaluations is acting strategically—they may infer a selfish, instrumental, and unethical desire to elicit positive reciprocity (Blank 2007, Gagne et al. 2008, Lin-Healy and Small 2013). As strategically acting members do not want their strategic behavior to be detected, they take actions that—at first sight—appear nonstrategic. Using only positive evaluations also suggests having low standards (Filippas et al. 2018, Bowers 2019). Members who are (about to be) evaluated may therefore find themselves needing to evaluate both positively *and* negatively. We thus hypothesize:

Hypothesis 1. *Organizational members who are (about to be) evaluated increase their participation in evaluating peers (i.e., they provide positive and negative evaluations rather than abstain).*

Targeting Nonpivotal Evaluations

Our first hypothesis suggests that the role duality in peer evaluations (i.e., being the evaluator and the evaluated) leads organizational members to participate in evaluating more peers. While participation in more peer evaluations allows them to show their peers that they are engaged, and the provision of negative evaluations allows them to show that they have high standards, the outlined dilemma points to the potential downside of this approach: it may offend peers and cause a backlash. We suggest and outline in the following hypotheses that organizational members minimize the risk of backlash by carefully targeting *whom* they evaluate.

One way of targeting—we hypothesize—is avoiding evaluations where their input may be *pivotal* for the outcome of the evaluation of a peer. Evaluations of peers differ regarding how close their (expected) outcome is—and thus how pivotal a member's evaluation might be if she participated. We suggest that members avoid participating in evaluations where their input may be pivotal (i.e., tilting the overall assessment in one direction or another). If the eventual outcome is obvious, peers will hardly pay attention to an individual's evaluation as it does not make a difference (Weaver 1986). Their evaluation is thus unlikely to offend anyone. If, by contrast, the member makes a pivotal evaluation, peers will pay attention, and the numerous peers who disagree with the member's evaluation may be offended; they may express their discontent with a negative evaluation when the member will be evaluated. Experimental evidence suggests that some

people take the pivotality of their vote into account (Piezunka and Schilke 2023) and delegate making pivotal decisions to avoid punishment (Bartling et al. 2015).

Hypotheses 2a and b. *While organizational members who are (about to be) evaluated increase their participation in the evaluation of peers (Hypothesis 1), they target both their (Hypothesis 2a) positive and (Hypothesis 2b) negative evaluations on nonpivotal evaluations (pivotality negatively moderates the positive and negative main effect of about-to-be-evaluated on positive evaluations).*

Organizational members are likely to make stronger adjustments for negative evaluations than for positive evaluations. A pivotal negative evaluation is more likely to be remembered and cause retaliation than a positive one (Margolis 2001, Brebels et al. 2008, Keysar et al. 2008). Negative evaluations have a stronger effect on the recipient than positive evaluations (Skowronski and Carlston 1989, Taylor 1991, Baumeister et al. 2001). They may be perceived as a rejection (Fernandez-Mateo and Coh 2015, Piezunka and Dahlander 2019) and even translate into a negative long-term relationship that hurts the evaluator (Labianca and Brass 2006, Brass and Borgatti 2019). They are also more impactful, as a few negative evaluations could alter the outcome of an evaluation process (Dlugos and Keller 2021). When a member is anticipating being evaluated, that member is unlikely to target negative evaluations in situations where the outcome is uncertain and the input could be pivotal. In sum, organizational members about to be evaluated increase their participation, but this increase is attenuated if the evaluation is expected to be pivotal, particularly for negative evaluations.

Hypothesis 2c. *The hypothesized moderation effect is greater for negative (Hypothesis 2b) than positive (Hypothesis 2a) evaluations.*

Targeting (In-)Active Peers

Another way of targeting—we hypothesize—is deciding what kind of members to evaluate. Increasing the number of positive and negative evaluations, as posited in Hypothesis 1, introduces a dilemma. While positive evaluations come with the chance to benefit from positive reciprocity, negative evaluations come with the risk of retaliation. Organizational members who are about to be evaluated are thus likely to target positive evaluations on peers who hope to benefit from positive reciprocity and negative evaluations of those peers where they do not expect retaliation. In this decision making, a peer's evaluation activity becomes paramount. Peers' activity levels tend to be highly skewed, in particular when organizations rely on self-selection, so that there is a very dedicated active group and a long tail of infrequent contributors (Dahlander and Frederiksen 2012, Setia et al. 2012, Kane et al. 2014,

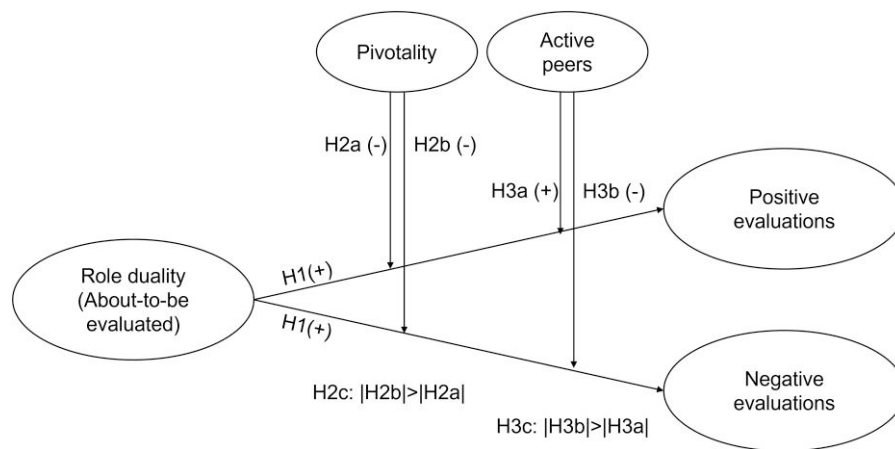
Safadi et al. 2021). Active peers, likely to reciprocate, receive more positive reviews due to their potential influence on the broader group (Aadland et al. 2019). However, giving negative feedback, especially to these active peers, carries inherent risks. A rebuff from an engaged member intensifies the immediate fallout and can trigger a cascading effect, prompting further unfavorable evaluations from others. This suggests that activity moderates the effect of positive and negative evaluations but in opposing ways.

Hypotheses 3a and b. *While organizational members who are (about to be) evaluated increase their participation in the evaluation of peers (Hypothesis 1), they target their positive evaluations on active peers (Hypothesis 3a), but their negative evaluations on inactive peers (Hypothesis 3b) (peer activity positively (negatively) moderates the main effect of about-to-be-evaluated on negative (positive) evaluations).*

We expect negative evaluations of active peers to be particularly unlikely as members who are (about to be) evaluated consider them dangerous and expect them to focus instead on negative evaluations of inactive peers. A negative evaluation of an active peer would stir strong negative emotions and translate into retaliation (Skowronski and Carlston 1989, Keysar et al. 2008). Moreover, the second-order cascading effect that we outlined is likely particularly strong. Negative evaluations are more salient than positive ones and, thus, are more likely to be noticed by peers (Baumeister et al. 2001). A negative evaluation of an active peer is especially noticeable and likely to be perceived as unfair and not based on merit (Bitektine and Haack 2015, Shaw et al. 2019, Safadi et al. 2021). If other peers believe a negative evaluator acted on criteria other than merit, they are likely to intervene (Bacharach and Lawler 1980). Also, active peers are likely to be well connected, and connected peers may retaliate on behalf of the negatively evaluated peer (Venkataramani and Dalal 2007). By contrast, inactive peers who struggle to be recognized (Cattani et al. 2014) are less likely to have peers who will support them and retaliate on their behalf. The negative evaluation of inactive peers is thus less likely to result in the outlined second-order effects. These arguments suggest that the moderation effect is greater for negative than positive evaluations because negative evaluations cause a stronger response.

Hypothesis 3c. *The hypothesized moderation effect is greater for negative (Hypothesis 3b) than for positive (Hypothesis 3a) evaluations.*

In sum, our hypotheses suggest that organizational members who are about to be evaluated increase their tendency to evaluate peers (negatively and positively). This is moderated by the pivotality of evaluations and by active peers (Figure 1). We return to the scope conditions in greater detail in the discussion.

Figure 1. Overview of Hypotheses


Methods

Research Context and Data

Data on peer evaluations are rare because organizations are concerned about the ethical, legal, and operational implications of sharing who evaluates whom and how. Thus, few studies have examined this kind of data (e.g., Li and Agha 2015, Aadland et al. 2019). We test our theory on the effect of the dual role structure on the strategic adjustment of peer evaluations using data from Wikipedia, the largest free online encyclopedia (Leskovec et al. 2010, Zhang and Zhu 2011, Gallus 2017, Piskorski and Gorbatai 2017, Klapper and Reitzig 2018, Nagaraj 2018). Wikipedia is contributed to and maintained by a vibrant community of self-selected volunteers (Hill 2013). It is open, as everybody can edit pages and contribute to the various articles, whether a registered member or not. To coordinate such broad participation, Wikipedia has adopted a structure where selected members take on the role of administrators with greater authority. For example, they can (un)block user accounts, restrict page edits, and delete pages.

We examine the peer evaluation-based process used to determine which members become administrators. The process is called “Request for Adminship” (RfA).³ It begins with a nomination, which candidates or others may send. Following the nomination but before the formal evaluation begins, candidates must respond to questions about their past activities on Wikipedia and their opinions of, and plans concerning, their prospective adminship. Then, the evaluation process moves to peers deciding whether and how to evaluate the candidate. Members have access to various sources of information, most notably the member’s user pages, which are akin to a private web page on Wikipedia, past contributions, and, crucial for our theory, the complete record of the member’s past evaluation of RfA candidates. Members may then participate in evaluating the candidate.

At the end of the process, a small group of selected Wikipedia members with even further privileges—the so-called “bureaucrats”—determine whether there is consensus regarding the nomination. Typically, an approval rating of over 75% will lead to a successful RfA and below 65% will not. For the gray zone, Wikipedia states, “RfAs that finish between 65 and 75% support are subject to the discretion of bureaucrats.”⁴ This implies that negative evaluations are asymmetrically important in affecting the outcome of the RfA because a threshold of more than 25% negative evaluations triggers bureaucrats’ intervention. The consequences of a failed RfA are not severe but can be emotionally draining, as our interviews suggest. Candidates can try again but are encouraged to wait at least three months.

Members, whether evaluating or being evaluated, put a lot of effort into RfAs. In online communities, members seek such positions as a means toward power or higher status, or because they are important to their identity (Vaast and Levina 2015). While successful candidates benefit from additional authority, many evaluating members want to ensure the evaluation process is fair and accurate. As one of our interviewees said, “When someone becomes an administrator on Wikipedia, there is no term limit. They have the position for life, literally. There’s no retirement policy, there is no standard process for removing the power of the administrator. [...] If you did get it wrong, then there’s no good process for reversing this.” Whereas the admin, short for administrator, position was originally considered by Wikipedia’s founder Jimmy Wales as merely a technical support function (“not a big deal”),⁵ the importance of the position and the standards and scrutiny applied to selection have increased substantially over time (Kittur et al. 2007, Jemielniak 2014). Thus, deciding whether a member becomes an administrator is thorough and aligned with Wikipedia’s focus on professionalism, objective evaluation,

and civil debates (Greenstein and Zhu 2016). Our interviewees underscored the importance of the high standards, given what is at stake on Wikipedia: they pointed to the importance of “getting it right” and “to maintain a standard” as administrators obtain additional privileges, which are difficult to revoke. While the thoroughness of the process may help the organization, it creates a burden on the side of the candidates. Some of our interviewees explained that they decided not to initiate an RfA, or at least to delay the process. The interviews also echoed that they thought carefully about how they carried themselves in the lead-up to their own evaluation.

We relied on the archived backup of the English-language version of the Wikipedia database in 2015, provided by the Wikimedia Foundation. Wikimedia creates monthly backups of all wikis that include all publicly accessible information. Those backups contain information on all interactions, for example, the current page content, logs on activities (e.g., user being created, user being promoted to admin), and changes to pages. In addition to contributions to Wikipedia pages, we focused on all RfA evaluation processes on Wikipedia starting in 2003 and continuing through 2014. We excluded RfAs that were stopped by administrators because the candidate either broke the rules or had no realistic chance of succeeding.⁶ For each RfA we determined all evaluations and matched them to an evaluator and candidate.

Research Design and Variables

As we theorize how role duality affects whether and how members evaluate, we take the perspective of a member and how they adjust to evaluating peers when anticipating their own RfA (i.e., when the member becomes a candidate) compared with times when they are not expecting to be evaluated. Beyond examining whether members participate when they are about-to-be candidates (Hypothesis 1), we also examined how and whom they evaluate (Hypotheses 2 and 3). To handle all three questions, we chose the member-evaluation pair as our unit of analysis.

To construct our variables, we collected information about the evaluating member (who may become a candidate in the future), the evaluated peer (current RfA candidate), their dyadic relation, and the evaluation process. We traced the full prior activity of the evaluating and evaluated peers, for example, where and how much they contributed to Wikipedia pages, how they evaluated past RfAs, or how they interacted with each other. Unless otherwise specified, all variables were constructed by calendar month. The sample comprises 3,434 RfAs with 187,898 evaluations of 2,883 candidates from 10,660 members. About 46% of RfAs are successful, indicating that these evaluations are important. We track RfAs from the first one observable, in 2003, until the end of 2014.

We also gathered qualitative data to become more familiar with the evaluation process. To gain access to interviewees, we contacted the head of research of the Wikimedia Foundation, a nonprofit organization that provides the infrastructure that underlies Wikipedia. Following her guidance, we publicly announced our study on a Wikipedia forum and provided contact information. We conducted semistructured interviews with 24 active members.⁷ The average interview lasted 46 minutes, and interviews were transcribed verbatim, totaling 284 pages. Seventeen interviews were conducted in English, and seven were conducted in German (see Table 1 for an overview). We drew on these interviews to inform our operationalization of the variables and to interpret our results, as described.

Dependent Variable

The Decision of Whether and How to Evaluate a Peer.

Our categorical dependent variable captures a member's evaluation of a peer. All registered members on Wikipedia can self-select to evaluate any candidate. We categorized the decision to participate as (0) abstain or participate. Active participation can be (1) negative or (2) positive.⁸

One challenge is that we do not directly observe members abstaining; they can self-select into participating in evaluations but do not record their abstentions. The potential risk set of abstentions is huge because any registered member could evaluate in any RfA, resulting in more than one billion abstentions. This would be nearly impossible to handle empirically, and including such a large number of abstentions could result in oversampling unlikely events (Maoz and Russett 1993). To create a more realistic risk set, we reduced the number of potential participating members—those with the right to participate, whether they do so or not—and used a random sample of nonevaluations for each observed evaluation (King and Zeng 2003). Such choice-based sampling strategy is common in studies of network formation (e.g., Sorenson and Stuart 2001, Hallen 2008). We first created a risk set for all members who could have potentially participated in an RfA. We included all members who had been active—that is, who had edited at least one Wikipedia page—before the RfA and who had during their tenure shown an interest in evaluating candidates, operationalized as participation in at least one RfA. For each RfA in our analysis, we select all members who evaluated the candidate. We include, on average, five members who did not evaluate that candidate but were part of our risk set for each evaluation. This implies that we create a data set of people who could abstain or engage at any time point without including the massive number of possible abstentions in the data set. Thus, we arrive at 1,013,305 nonevaluations to 187,898 evaluations.

Table 1. Overview of Interviews

Length of interview (hours:minutes:seconds)	Language	Transcript length (pages)	Active since	Admin on Wikipedia
0:46:18	English	11	Long time ago	No
0:49:18	English	16	2004	Yes
0:43:05	English	16	2013	Yes (since 2015)
0:45:30	English	17	2014	No
1:28:19	German	12	2014	Yes
0:28:32	English	7	2012	No
0:58:41	English	19	2010	No
0:48:23	English	13	2006	Yes (since 2008)
0:47:12	German	9	2006	No
0:31:20	German	6	2006	Yes (in 2007)
0:39:00	German	8	2004	No
0:42:07	German	10	2003	Yes (since 2006)
0:44:45	English	12	2017	Yes
1:00:50	German	10	2011	No
0:59:39	English	17	2006	Yes (9 years)
0:38:23	English	11	2006	No
0:47:36	English	12	2006	No
0:27:11	English	10	2017	Yes
0:42:23	English	14	2005	Yes
0:59:41	English	14	2007	Yes (in 2009)
1:04:10	German	10	Long time ago	No
0:43:42	English	13	2005	Yes
0:47:19	English	11	2005	Yes
Email interview	English	6	2008	Yes

Independent Variable

About-to-Be Candidate (Hypothesis 1). We theorize that an important situation in which role duality matters is when members are about to be candidates and, thus, change how they evaluate their peers. We focus on the time period leading up to their candidacy. It is in this period that members who are about to be candidates are very likely aware that their evaluations of peers may affect how they are evaluated. We deploy a dummy that takes the value one for all evaluations in which members could have potentially participated during the six months before their own RfA.⁹

Our assumption that the outlook on being evaluated soon affects the evaluation behavior of members is supported by our interviews. As one interviewee recalled, “When coming up for evaluation, I was so anxious. I would have trouble sleeping. [...] Your whole history is out there for someone to judge you, and something that you think is not a problem, other people might think it is and then your whole campaign can come crashing down, because of one thing.” Therefore, we assume that most candidates prepare for their RfA, become conscious of their behavior, and adjust it. This was reflected in our interviews. For example, one interviewee stated, “I was planning to go for an RfA. It tended to mean that every time I was on Wikipedia, I was conscious of my behavior and the quality of my edits.” Another interviewee explained that “because it’s like, a normal society, they want to go upstairs. They want to be at a higher position ... It’s like in politics, they want to become mayor, they need some followers, somebody who support them. It’s the same as

politics, it’s running for presidency.” On average, we expect members who intend to run an RfA in the future will be more strategic than in other situations.

The validity of the variable also depends on whether other peers would pay attention to past evaluations of a candidate. Our interviews revealed that people often knew about a candidate’s evaluation track record: “All decisions are made transparent, and the same is with voting. So, everything is open, that’s the idea behind an open system, and so you can see how you voted in something.” They either remembered it or looked it up. Some, perceiving that past evaluations were documented as cumbersome, even created a software tool that made it easier to study the candidate’s previous evaluations of others. One interviewee explained that the RfA was a window into scrutinizing a candidate: “To me, it would be almost like an audition ... I actually get to see them and how they operate and what they do. It’s a little insight or a window into how they view the role of administrators. It gives you a chance to see who they think might be able to perform the job that now they are aspiring to and what levels or what criteria they will look at.” In sum, our robustness checks in the online appendix and our interviews support the idea that members changed their behaviors when evaluated and that those evaluating the candidates carefully investigated their prior track record.

Moderator Variables

Pivotality of Evaluations (Hypothesis 2). Although any evaluation affects the outcome of the evaluation process,

the degree of pivotality varies. Because the RfA process is transparent and members can see prior evaluations, they can see to what degree an evaluation is pivotal. We classify RfAs in which the current share of positive evaluations is close to 70% as being more pivotal. If it is above 75% or below 65%, the candidate is confirmed or rejected, respectively, without further discussion. Between 65% and 75%, the decision goes to the bureaucrats—members of the Wikipedia leadership team with special decision rights—who decide whether to promote the candidate. Thus, if the current share of positive evaluations approaches 70%, any evaluation is particularly pivotal. In contrast, members can offer more ritualized evaluations in evaluation processes with less uncertainty (Bartling et al. 2015). We measure the degree of pivotality with a continuous variable that takes the value one if the current approval percentage is 70% (most pivotal) and decreases to zero if the current approval percentage is at 100% or less than 40% (least pivotal).

Activity of Evaluated Peer (Hypothesis 3). Having hypothesized that about-to-be candidates consider how active a candidate is, we measure the degree to which those candidates participated in peer evaluations as the number of candidates they evaluated in the lookback period of six months divided by the total number of evaluation processes in which they could have participated. This activity is salient because it captures the extent to which an individual is perceived to be inactive and peripheral in evaluations of others, which was also echoed in our qualitative interviews: “If you have never shown interest in any elections, I doubt that you’re the kind of person that takes enough ... interest in community affairs in general.”

Control Variables: Evaluating Member. Members differ in their tendency to participate in evaluations. *Activity* of the evaluating member is calculated analogously to Hypothesis 3 (Activity of the Evaluated Peer) by measuring the number of evaluation processes in which the evaluating member registered a valid, nonwithdrawn evaluation during the lookback period of six months, divided by the total number of evaluation processes in which they could have participated. The overall amount and type of the evaluating member’s activity might also affect their tendency to evaluate the current candidate. Specifically, we control for the evaluating member’s experience in evaluating candidates by measuring the *Number of Evaluations* they participated in before the evaluation. The evaluating member’s *Number of Edited Pages* takes the logarithm of the number of unique pages that he revised during the lookback period (Piskorski and Gorbatai 2017). Next, we measure a member’s activity in relation to the overall activity on Wikipedia. The evaluating member’s *Edits* is the person’s share of total page edits in the lookback period (Aaltonen and Seiler

2016). Because the share of any one person’s page edits in relation to the edits by all Wikipedia members is very low, we standardized this variable. We also control how engaged members are on the Wikipedia pages they edit compared with other members (*Engagement*). A member might be more active on pages with few other members and therefore be less likely to have missed the work of candidates also active on those pages. The variable *Share of Edits* shows how active the evaluating member was on edited pages compared with other editing members. For example, a member could make small contributions to many pages or large contributions to only a few pages. To calculate this variable, we use the pages the evaluating member edited in the lookback period and took the mean of his share of edits (the number of his edits divided by the total number of edits). We also control for *Tenure*, calculated as months since the evaluating member’s registration.

Members can also be characterized by their previous RfA(s). If they were successful, being an administrator (*Admin*) may affect their likelihood of evaluating and their evaluation criteria. The other possibility is that they tried to become an administrator but failed (*Previous Failed Candidacy*). During the RfA process, members can change their evaluation, for example, when they see convincing arguments for or against the candidate that were unknown at the time of their first evaluation. Therefore, to further classify a member’s evaluation pattern, we calculate the tendency for *Reversed Evaluations*, measured as how often the member changed a positive evaluation to a negative one (or vice versa) during an RfA. We also calculate a member’s likelihood of deviating from the majority position (*Deviation from Majority*), that is, the distance from current approval percentage, which is the current number of positive evaluations divided by positive and negative evaluations. For example, if the evaluating member evaluates a candidate positively while the current approval percentage is 40%, this implies a “distance” of 0.6 (1–0.4). We average the distances for all prior RfAs in which the evaluating member participated. Finally, we measure when the evaluating member normally evaluated a candidate. For each evaluation process, we determine at what relative point in time they evaluated the candidate (*Relative Time to Evaluate*) and then calculate the mean of those calculations at the time of the evaluation. By measuring all these variables, we have accounted for differences across members regarding their tendency to participate in evaluations and to evaluate negatively.

Control Variables: Evaluated Peer. We collected corresponding information about the evaluated peer, that is, the current RfA candidate. Specifically, we control for the candidate’s *Number of Edits*, *Share of Edits*, and *Tenure*. We also calculate the *Average Similarity* of the candidates

to all (potential) evaluators as the mean of all dyadic similarity measures.

Although a candidate's quality, that is, their qualification for the work of administrator, is latent and evaluated differently by different evaluators, there are also several common indicators that Wikipedia evaluators use to assess a candidate's quality. The *Number of Edits* is important to evaluators, but it also matters which pages candidates contribute to. To categorize the candidate's activity, we utilize Wikipedia's own classification of each page, which is called namespace (NS).¹⁰ For example, NS0 refers to typical *Content pages*, NS1 to *Discussion pages*, NS2 and NS3 to *Users pages*, NS4 to pages about *Wikipedia*, NS6 to *File pages*, NS10 to *Template pages*, and NS14 to *Category pages*. Candidates are expected to contribute to many of these types of pages. We separate between these types of edits in the analysis to take into account that one type may be more important. Further, a candidate can either be nominated by another user or self-nominate. *Self-Nomination* can be perceived as a bad quality signal as the candidate does not have sufficient support to be nominated by someone else. Candidates are also expected to be active continuously; therefore, *Activity Gaps*, measured as the longest time in the lookback windows in which the candidate was not active on Wikipedia, are seen as a negative quality indicator. Evaluators also check whether the candidate has created new pages. This is perceived as an important task for Wikipedians, which we measure as the logarithm of the number of *Created Pages*.

Another quality indicator is the number of “reverts”—the number of other contributors' edits they have deleted. *Edits Reverted (Total)* is measured as the number of reverts, and *Edits Reverted (Share)* is measured as the ratio of reverts to overall edits in the lookback window. A very high ratio of deleted edits can be seen as controversial; it could result from cleaning up unproductive edits or vandalism by previous members, but it could also result from overzealous deleting of edits the candidate did not like. It can therefore lead to opposition. Another quality indicator is how thorough a candidate is when editing pages. Commenting on edits to explain the changes to future members facilitates coordination. Evaluators may therefore use the ratio of a candidate's edits commented on (*Edits Commented*) as an indicator of thoroughness (Collier et al. 2008). Finally, evaluators consider whether the candidate has a *Previous Failed Candidacy*.

Control Variables: Dyad (Evaluator-Evaluated-Pair). Previous interactions between members can be an important antecedent to how they evaluate other (Lerner and Lomi 2017). We control for interactions in previous RfAs. For example, if member A evaluated member B during B's RfA, member B may now choose to reciprocate, positively or negatively. This possibility is captured in the

dummy, *Preceding Peer Approval*, which is one if the peer had evaluated the evaluating member positively and zero if not. We also determine how often the evaluating member and the evaluated peer have deleted one another's edits to indicate potential prior disagreements. *Edits of Candidate Deleted* takes the logarithm of edits of the evaluated peer that the evaluating member deleted, whereas *Edits by Candidate Deleted* takes the logarithm of edits of the evaluating member that the evaluated peer deleted. We also account for the *Work Overlap* between the evaluating member and evaluated peer. We capture the degree to which they have edited the same pages by constructing a vector for the potential evaluator and one for the candidate, each including all Wikipedia pages. In these vectors, one indicates that the actor revised a page during the lookback period and zero indicates that the person did not. We then calculate the correlation coefficient for the two vectors.

Control Variables: RfA Evaluation Process. We create two time-invariant variables to classify the RfA. *Activity First Day* measures how many evaluations and comments were made on the first day of the RfA, indicating the overall interest in the evaluation process, which is not driven by the sequence of the RfA. For example, if many members participate immediately, this could be driven by the particular candidate or some prior controversy. By focusing on first-day activity, we can exclude the dynamic effect of early evaluations on later ones. We count the number of *RfAs Same Month* to account for any attention effects. If there are many elections simultaneously, the evaluating member could be more or less inclined to evaluate a candidate relative to others (Bowers 2015).

We also create three time-variant variables to measure the dynamic effects of each RfA. *Current Approval Rating* describes the level of support for the evaluated peer at the time of the evaluation (Botelho 2017). *Current Number of Evaluations* indicates how many evaluations have been made when a member makes their own. In addition to evaluations, members can comment on another member's evaluation, for example, criticizing it or asking for clarification. *Current Number of Comments* shows how many comments have been made during a member's evaluation and is a measure of ongoing discussion during the RfA process.

Data Analysis

A challenge related to nonevaluation is that we cannot observe whether members who did not participate were unaware of the RfA—for example, due to inactivity—or were aware and chose not to participate. Our theory focuses on abstaining—an active choice not to participate. To compare active decisions to participate in the evaluation and active decisions not to participate (abstentions), we restricted the sample of nonevaluations to instances in which members likely knew about the RfA. Because we

cannot observe such knowledge directly, we employ a matching to find nonevaluations that have characteristics similar to evaluations. We used propensity score matching (Rosenbaum and Rubin 1984) to find a subset of evaluations in which the members could have participated and evaluations in which they did participate. Having constructed a credible counterfactual by calculating the propensity for each potential evaluation to result in an actual evaluation, we could select nonevaluations that were similar to actual evaluations in their observable factors. We selected factors that we assumed to be correlated with awareness of the RfA: the potential evaluator’s activity, the candidate’s prominence, and the similarity between evaluator and candidate (see Online Appendix T5.1 for the full list of factors). We then used a logistic regression to determine how these factors explain whether potential evaluation would result in actual evaluations and used it to calculate the propensity for an evaluation. Next, we matched a maximum of one nonevaluation to each evaluation using nearest neighbor matching ($k = 1$) with replacement. To evaluate the effectiveness of the propensity score method, we compared the sample means of the matched nonevaluations to actual evaluations (Online Appendix T5.3) and compared the distribution of propensity scores before and after matching. Following the precautions described in the online appendix, we were able to rule out unawareness and focus on the effect of the decision to evaluate. In the end, we arrived at 187,898 evaluations and 116,703 comparable nonevaluations. As an alternative, we also used coarsened exact matching (see Online Appendix T5.5).¹¹

We estimate a multinomial logistic regression to test our hypotheses. The unit of analysis is a member’s decision on how to evaluate a candidate. A multinomial logistic regression captures the odds of selecting one option without assuming an ordinal ranking of the

options. This way, a positive effect of an independent variable on the evaluation decision—for example, a positive evaluation—does not necessarily imply a negative effect on an alternative decision—for example, a negative evaluation. The odds are always in comparison with a reference group. Borrowing from research in political science that also looks at forms of voting, we use abstaining as the baseline compared with different active choices (Peress 2011, Bagozzi and Marchetti 2017). We clustered standard errors on each evaluator, as evaluators can participate in multiple RfA processes and these evaluations are likely not independent.

Results

Tables 2 and 3 provide descriptive statistics for members, RfA elections, and variables at the evaluating member, evaluated peer, and dyad levels. On average, 50 members participated in an election, with 37 positive, 10 negative, and 3 neutral evaluations.

Tables 4 and 5 report the average marginal effects from the multinomial logistic regressions on the likelihood of evaluating a candidate positively, negatively, or neutrally compared with not evaluating at all. Each multinomial regression reports two columns—for negative and positive evaluations—compared with the baseline of no evaluation.

Before discussing the main results, we note that certain control variables shed light on the evaluation process. For example, higher tenure and activity (edited pages) by the evaluating member are associated with an increase in negative evaluations, potentially showing a rise in the standards of the more active and tenured members. Furthermore, the more active and tenured candidates are seen more positively, as is their due diligence (edits commented). On a dyadic level, the effect of positive reciprocity and negative retaliation is strong. At

Table 2. Descriptive Overview of Evaluator Groups

	Before ATBC		ATBC		After ATBC		Never ATBC	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Activity	0.194	(0.396)	0.408	(0.491)	0.209	(0.406)	0.113	(0.317)
Positive evaluations	0.147	(0.354)	0.299	(0.458)	0.154	(0.361)	0.081	(0.273)
Negative evaluation	0.036	(0.186)	0.079	(0.269)	0.042	(0.201)	0.025	(0.158)
Neutral evaluations	0.011	(0.106)	0.030	(0.170)	0.013	(0.113)	0.007	(0.081)
Edits	0.132	(1.706)	0.184	(0.898)	0.146	(0.966)	−0.119	(0.991)
Share of edits	0.310	(0.122)	0.317	(0.111)	0.303	(0.115)	0.282	(0.129)
Tenure	20.029	(15.507)	17.017	(14.664)	34.444	(23.174)	23.195	(19.841)
No. of elections participated	11.768	(18.294)	29.312	(52.258)	69.600	(99.449)	16.122	(39.848)
Admin	0.013	(0.115)	0.006	(0.080)	0.786	(0.410)	0.049	(0.216)
Relative time to evaluate	0.349	(0.125)	0.327	(0.122)	0.338	(0.121)	0.387	(0.208)
Previous failed candidacy	0.000	(0.000)	0.215	(0.411)	0.194	(0.396)	0.000	(0.000)
Reversed evaluations	0.420	(1.346)	0.692	(1.436)	0.653	(1.758)	0.127	(0.532)
Deviation from majority	0.206	(0.105)	0.199	(0.100)	0.200	(0.107)	0.249	(0.196)
Observations	21,265		69,417		315,992		481,334	

Note. Evaluators are categorized as follows: before they are anticipating a future evaluation (Before ATBC), while they are about-to-be candidates (ATBC), after they are about-to-be candidates (After ATBC), or as evaluators who were never candidates (Never ATBC).

Table 3. Descriptive Statistics with Means and Standard Deviations

[illegible]

Table 3. (Continued)

	Mean	SD	Min	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	
Evaluated peer (N = 2,883)																											
(1) Edited pages (log.)	6.468	1.833	0.693	11.036	1																						
(2) Activity	0.027	0.069	0	0.757	0.27	1																					
(3) Edits (log.)	7.248	1.744	0.693	11.102	0.97	0.26	1																				
(4) Edited pages (log.)	6.468	1.833	0.693	11.036	1	0.27	0.97	1																			
(5) Engagement	-0.059	1.101	-0.875	11.743	0.52	0.13	0.51	0.52	1																		
(6) Share of edits	0.281	0.122	0.002	0.822	0.3	0.06	0.3	0.3	0.22	1																	
(7) Tenure	17.465	16.959	0.033	119.567	0.2	0.11	0.19	0.2	0	0.21	1																
(8) Edits commented	0.823	0.225	0	1	0.53	0.2	0.5	0.53	0.09	0.09	0.27	1															
(9) Edits reverted (total)	60.422	107.098	1	1,715	0.49	0.21	0.48	0.49	0.44	0.08	0.06	0.21	1														
(10) Edits reverted (share)	0.066	0.114	0	0.695	0.31	0.14	0.25	0.31	0.04	-0.02	0.09	0.3	0.46	1													
(11) Self-nomination	0.467	0.499	0	1	-0.48	-0.14	-0.49	-0.48	-0.3	-0.1	-0.13	-0.31	-0.18	-0.05	1												
(12) Previous failed candidacy	0.023	0.151	0	1	0.06	0.02	0.06	0.06	0.03	-0.03	0.03	0.04	0.07	0.04	0	1											
(13) Edits, Content pages	6.507	1.851	0	10.548	0.94	0.2	0.96	0.94	0.53	0.25	0.18	0.45	0.45	0.23	-0.47	0.05	1										
(14) Edits, Discussion pages	4.314	1.929	0	10.473	0.78	0.21	0.83	0.78	0.39	0.25	0.22	0.51	0.33	0.08	-0.47	0.06	0.78	1									
(15) Edits, Template pages	2.229	1.778	0	8.157	0.63	0.16	0.65	0.63	0.34	0.29	0.18	0.32	0.33	0.08	-0.3	0.03	0.62	0.59	1								
(16) Edits, Categories pages	1.352	1.619	0	9.286	0.56	0.08	0.53	0.56	0.43	0.28	0.09	0.22	0.27	-0.02	-0.3	0.02	0.54	0.47	0.57	1							
(17) Edits, User pages	3.811	1.663	0	9.026	0.64	0.24	0.69	0.64	0.3	0.28	0.09	0.33	0.32	0.22	-0.34	0.05	0.6	0.55	0.49	0.35	1						
(18) Edits, User discussion	4.941	2.032	0	10.208	0.82	0.33	0.83	0.82	0.3	0.26	0.16	0.56	0.5	0.45	-0.41	0.06	0.72	0.7	0.52	0.33	0.69	1					
(19) Edits, Wikipedia pages	4.696	2.069	0	9.138	0.83	0.33	0.84	0.83	0.39	0.17	0.14	0.56	0.36	0.21	-0.48	0.08	0.74	0.75	0.56	0.44	0.63	0.81	1				
(20) Edits, Files pages	1.28	1.433	0	8.915	0.43	0.15	0.45	0.43	0.24	0.22	0.13	0.19	0.25	0.06	-0.21	0.03	0.44	0.39	0.45	0.34	0.35	0.38	0.37	1			
(21) Activity gaps	39.099	46.5	1	179	-0.63	-0.15	-0.65	-0.63	-0.22	-0.12	-0.35	-0.41	-0.18	-0.1	0.36	-0.06	-0.66	-0.62	-0.44	-0.32	-0.41	-0.48	-0.54	-0.31	1		
(22) Created pages (log.)	4.652	1.893	0	9.498	0.92	0.27	0.9	0.92	0.48	0.52	0.16	0.48	0.47	0.34	-0.43	0.04	0.85	0.72	0.62	0.53	0.65	0.83	0.79	0.44	-0.53	1	
(23) Average similarity	0.006	0.004	-0.001	0.035	0.35	0.34	0.39	0.35	0.23	-0.24	-0.12	0.29	0.2	0.23	-0.33	0.1	0.32	0.37	0.17	0.02	0.31	0.46	0.55	0.07	-0.22	0.3	

Table 4. Multinomial Logistic Regression; DV: Decision How to Evaluate (Base Case: Abstention)

	Model 1.1 Negative	Model 1.2 Positive	Model 2.1 Negative	Model 2.2 Positive
<i>About-to-be candidate (Hypothesis 1)</i>			1.086*** (0.091)	1.074*** (0.083)
Evaluating member				
<i>Activity</i>	−0.167 (0.237)	−0.253 (0.203)	−0.145 (0.245)	−0.209 (0.213)
<i>Edited pages</i>	−0.201** (0.063)	−0.133* (0.058)	−0.243*** (0.064)	−0.168** (0.059)
<i>Edits</i>	0.146* (0.062)	−0.032 (0.057)	0.147* (0.063)	−0.039 (0.057)
<i>Engagement</i>	−0.011 (0.026)	−0.076** (0.024)	0.013 (0.027)	−0.048* (0.024)
<i>Share of edits</i>	−0.414 (0.278)	0.572* (0.265)	−0.380 (0.278)	0.599* (0.270)
<i>Tenure</i>	0.012*** (0.002)	0.006** (0.002)	0.011*** (0.002)	0.006** (0.002)
<i>No. of elections participated</i>	−0.001* (0.000)	−0.000 (0.000)	−0.001* (0.000)	−0.000 (0.000)
<i>Admin</i>	−0.062 (0.065)	−0.227*** (0.057)	0.106 (0.068)	−0.061 (0.057)
<i>Relative time to vote</i>	1.805*** (0.155)	2.480*** (0.139)	1.818*** (0.153)	2.481*** (0.138)
<i>Previous failed candidacy</i>	0.601*** (0.116)	0.638*** (0.103)	0.084 (0.113)	0.131 (0.103)
<i>Reversed evaluations</i>	0.001 (0.001)	0.002*** (0.000)	0.000 (0.001)	0.002*** (0.000)
<i>Deviation from majority</i>	3.600*** (0.168)	−2.237*** (0.154)	3.718*** (0.170)	−2.123*** (0.155)
Evaluated peer				
<i>Edited pages (log.)</i>	−0.023 (0.028)	−0.016 (0.023)	−0.028 (0.028)	−0.019 (0.023)
<i>Activity</i>	−0.952*** (0.188)	−0.815*** (0.147)	−0.955*** (0.189)	−0.813*** (0.149)
<i>Edits (log.)</i>	0.198*** (0.041)	0.205*** (0.041)	0.205*** (0.042)	0.211*** (0.041)
<i>Tendency to evaluate neg.</i>	4.500*** (0.667)	1.993*** (0.524)	4.506*** (0.667)	2.013*** (0.524)
<i>Engagement</i>	−0.144*** (0.020)	−0.184*** (0.016)	−0.144*** (0.020)	−0.187*** (0.016)
<i>Share of edits</i>	−1.594*** (0.133)	−1.393*** (0.124)	−1.578*** (0.134)	−1.383*** (0.125)
<i>Tenure</i>	−0.002** (0.001)	0.002* (0.001)	−0.002** (0.001)	0.002** (0.001)
<i>Edits commented</i>	−1.142*** (0.069)	−0.569*** (0.090)	−1.138*** (0.070)	−0.567*** (0.091)
<i>Edits reverted (total)</i>	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)
<i>Edits reverted (share)</i>	−0.510*** (0.102)	−0.195* (0.082)	−0.515*** (0.103)	−0.186* (0.084)
<i>Self-nomination</i>	0.013 (0.024)	−0.159*** (0.017)	0.013 (0.024)	−0.158*** (0.017)
<i>Previous failed candidacy</i>	0.455*** (0.022)	0.348*** (0.018)	0.452*** (0.023)	0.346*** (0.018)
<i>Edits, Content pages</i>	−0.115*** (0.021)	−0.061** (0.019)	−0.117*** (0.021)	−0.061** (0.019)
<i>Edits, Discussion pages</i>	0.053*** (0.010)	0.072*** (0.008)	0.052*** (0.010)	0.073*** (0.008)
<i>Edits, Template pages</i>	0.016* (0.006)	−0.003 (0.005)	0.016* (0.007)	−0.002 (0.006)
<i>Edits, Categories pages</i>	−0.036*** (0.007)	−0.012* (0.005)	−0.036*** (0.007)	−0.012* (0.005)
<i>Edits, User pages</i>	0.009 (0.008)	−0.006 (0.006)	0.008 (0.008)	−0.007 (0.006)

Table 4. (Continued)

	Model 1.1 Negative	Model 1.2 Positive	Model 2.1 Negative	Model 2.2 Positive
<i>Edits, User discussion</i>	0.019 (0.014)	0.029* (0.012)	0.021 (0.014)	0.030** (0.012)
<i>Edits, Wikipedia pages</i>	−0.012 (0.012)	0.072*** (0.010)	−0.014 (0.012)	0.071*** (0.010)
<i>Edits, Files pages</i>	−0.024*** (0.006)	−0.021*** (0.005)	−0.024*** (0.006)	−0.022*** (0.005)
<i>Activity gaps</i>	0.004*** (0.000)	0.003*** (0.000)	0.004*** (0.000)	0.003*** (0.000)
<i>Created pages (log.)</i>	−0.021 (0.019)	−0.037* (0.016)	−0.020 (0.019)	−0.037* (0.016)
<i>Average similarity</i>	−52.265*** (4.358)	−75.595*** (3.997)	−51.440*** (4.360)	−75.340*** (4.003)
Dyad				
<i>Preceding peer approval</i>	2.518*** (0.147)	3.080*** (0.134)	2.498*** (0.148)	3.064*** (0.134)
<i>Preceding peer criticism</i>	2.790*** (0.400)	2.180*** (0.361)	2.778*** (0.402)	2.170*** (0.364)
<i>Work overlap</i>	−0.175*** (0.014)	−0.032** (0.011)	−0.174*** (0.014)	−0.031** (0.011)
<i>Edits of candidate deleted</i>	0.091* (0.045)	0.462*** (0.034)	0.103* (0.045)	0.472*** (0.034)
<i>Edits by candidate deleted</i>	0.208*** (0.056)	0.333*** (0.051)	0.215*** (0.058)	0.338*** (0.054)
<i>Edits of candidate deleted × edits by candidate deleted</i>	0.026 (0.028)	−0.149*** (0.024)	0.025 (0.028)	−0.149*** (0.025)
Election-level controls				
<i>Activity first day of election</i>	0.044*** (0.001)	0.047*** (0.001)	0.044*** (0.001)	0.047*** (0.001)
<i>Current no. evaluations</i>	−0.058*** (0.001)	−0.061*** (0.001)	−0.058*** (0.001)	−0.060*** (0.001)
<i>Current approval rating</i>	−0.202** (0.076)	4.550*** (0.091)	−0.178* (0.076)	4.561*** (0.091)
<i>Current number of comments</i>	−0.004*** (0.001)	−0.009*** (0.001)	−0.004*** (0.001)	−0.009*** (0.001)
<i>No. elections same month</i>	0.006*** (0.001)	0.004*** (0.001)	0.006*** (0.001)	0.004*** (0.001)
<i>Current pivotality</i>	1.618*** (0.050)	0.595*** (0.043)	1.624*** (0.050)	0.602*** (0.043)
Constant	−1.487*** (0.227)	−2.891*** (0.320)	−1.420*** (0.229)	−2.777*** (0.317)
Year fixed effects	Yes	Yes	Yes	Yes
Pseudo R ²		0.467		0.471
Log-likelihood		−199,331		−197,676
N		278,822		278,822

Notes. Robust standard errors in parentheses. We do not report the results for the neutral category. They can be found in the online appendix.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

the same time, deletions by and of the candidate are seen as a positive signal; only when the evaluating member and the candidate delete each other's edits (interaction of both variables) is the likelihood of positive evaluations reduced. We next turn to test the hypotheses.

Testing Hypothesis 1: Increased Participation. About-to-be candidates are more likely to participate in evaluations by evaluating candidates negatively or positively (Table 4, Model 2, $\beta_{\text{neg}} = 1.086$ and $\beta_{\text{pos}} = 1.074$, both $p < 0.001$). The marginal coefficients for negative and positive evaluations are positive and significant; the

probability for negative evaluations changes from 10.6% to 14.4% ($p = 0.005$) and, for positive evaluations, from 37% to 47% ($p = 0.008$). This shows that about-to-be candidates do not reduce their negative evaluations. This result is consistent with the idea that candidates increase their positive and negative evaluations of candidates when they are about to be evaluated.

Testing Hypothesis 2: Targeting Nonpivotal Evaluations. We next test whether about-to-be candidates reduce their pivotal evaluations (Hypothesis 2). The marginal effects are significant for negative (Hypothesis

Table 5. Multinomial Logistic Regression; DV: Decision How to Evaluate (Base Case: Abstention); Moderators for Hypotheses 1 and 2 Included

	Model 1.1	Model 1.2	Model 2.1	Model 2.2	Model 3.1	Model 3.2
	Hypothesis 2		Hypothesis 3		Combined	
	Negative	Positive	Negative	Positive	Negative	Positive
About-to-be candidate (ATBC)	1.279*** (0.099)	1.163*** (0.089)	1.150*** (0.094)	1.074*** (0.088)	1.305*** (0.102)	1.154*** (0.094)
Evaluation pivotality	1.687*** (0.050)	0.614*** (0.043)	1.615*** (0.049)	0.581*** (0.042)	1.683*** (0.050)	0.614*** (0.043)
Candidate activity	−0.978*** (0.192)	−0.827*** (0.149)	−0.884*** (0.195)	−0.842*** (0.151)	−0.902*** (0.194)	−0.847*** (0.150)
ATBC × evaluation pivotality (Hypothesis 2)	−0.629*** (0.105)	−0.372*** (0.091)			−0.602*** (0.103)	−0.370*** (0.090)
ATBC × candidate activity (Hypothesis 3)			−1.164** (0.380)	0.023 (0.320)	−0.889* (0.373)	0.135 (0.319)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Evaluating member controls	Yes	Yes	Yes	Yes	Yes	Yes
Evaluated peer controls	Yes	Yes	Yes	Yes	Yes	Yes
Dyad-level controls	Yes	Yes	Yes	Yes	Yes	Yes
Election-level controls	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R ²		0.345		0.345		0.345
Log-likelihood		−239,508		−239,543		−239,497
N		274,672		274,672		274,672

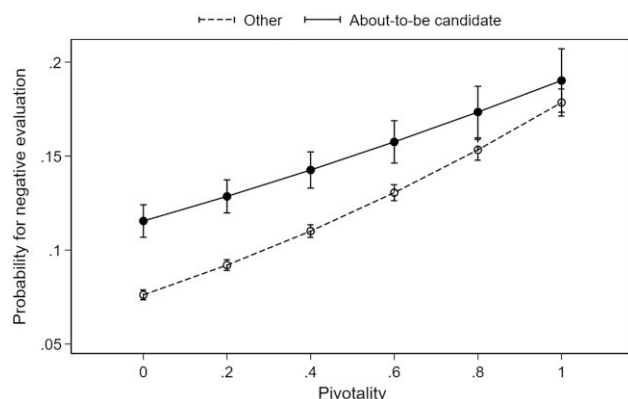
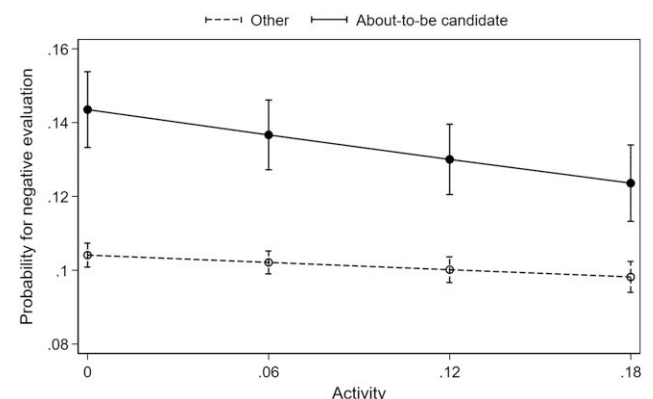
Notes. Robust standard errors in parentheses. We do not report the results for the neutral category. They can be found in the online appendix.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

2b) and positive (Hypothesis 2a) evaluations (Table 5, Model 3, $\beta_{\text{neg}} = -0.602$, $\beta_{\text{pos}} = -0.370$, both $p < 0.001$). Figure 2 illustrates this for negative evaluations: the likelihood that they will negatively evaluate a peer in pivotal evaluation processes is no different from other evaluators (diff. = -0.006 , $p = 0.461$). Still, about-to-be candidates are far more likely to evaluate peers negatively in cases where the outcome is obvious (diff. = 0.027 , $p < 0.001$). Comparing the coefficients, we find that negative evaluations' reduction is larger than that of positive ones, supporting Hypothesis 2c ($p < 0.001$).

Overall, this supports the logic that about-to-be candidates target their evaluations, specifically negative ones, to those that are unlikely to swing the evaluation outcome in either direction. One interviewee mentioned

that many “don’t want to go against the majority. So, you tend to get herd behavior.” Another, reflecting on the period before his RfA, explained that he often waited until he could get a better idea of what others were thinking: “[It’s] helpful to vote later, [...] you already see other people’s rationales.” In general, our interviews confirmed that members were wary of pivotal evaluations. As one remarked, “I only will put myself in a position that I’m confident of and my reasoning would be sound when I made that final decision, especially a pivotal decision that requires the highest levels of impartiality, balance, fairness, and objectivity.” Another interviewee explained, “And so we thought there were definitely people who were shying away from some candidates, particularly because while you can vote neutral,

Figure 2. Predicted Probability for Negative Evaluation by Members Moderated by Pivotality (Hypothesis 2b)**Figure 3.** Predicted Probability for Negative Evaluation by Members Moderated by Active Peers (Hypothesis 3b)

it’s probably more common, if you can’t make up your mind, to just not vote.” These findings point in the direction that about-to-be candidates, while becoming more rather than less active, begin to target their evaluations more on people where their evaluation is not pivotal. This is consistent with candidates changing their evaluation pattern when concerned about their candidacy.

Testing Hypothesis 3: Targeting (In-)Active Peers.

Hypothesis 3 posits that members who are about to be evaluated are more likely to make positive evaluations (Hypothesis 3a) about active members and less likely to make negative evaluations (Hypothesis 3b) about them. The marginal effects are significant for negative but not positive evaluations (Table 5, Model 3, $\beta_{\text{neg}} = -0.889$, $p = 0.017$; $\beta_{\text{pos}} = 0.135$, $p = 0.673$). This indicates that about-to-be candidates avoid negative evaluations on active peers but do not increase their positive evaluations of them.

Possibly, about-to-be candidates avoid fostering positive reciprocity as this might be seen as strategic by other peers. About-to-be candidates tend to negatively evaluate candidates who are not active (diff. = 0.039, $p < 0.001$), but they are marginally less likely to negatively evaluate active candidates (diff. = 0.025, $p = 0.001$). While the negative coefficient is significant and the positive is not, this suggests that negative evaluations have a stronger effect (Hypothesis 3c). However, a direct comparison of coefficients does not show a significant difference ($p = 0.240$). Figure 3 shows the likelihood of negative evaluations—with a downward slope for about-to-be candidates concerning a current candidate’s activity—which, as expected, is smaller for members that do not

anticipate their evaluation. Our interviews support the notion that going against active members of the community can be dangerous: “[...] like at some point if you’re active enough you just know people, you just know what they’re like [...]. So, like, either you just know you can trust them and, like, then the big question is always, like, why were you not an admin already?” When asked whether candidates would evaluate active members negatively, one interviewee responded, “I think they avoid conflict. I think they avoid pissing anyone off who might be influential.”

Additional Analyses and Robustness Checks

As an additional analysis, we investigate when evaluators make their evaluations. Hypothesis 2 argues that about-to-be candidates, fearing retaliation, will be especially reluctant to make a pivotal negative evaluation. By the same token, we can extend the argument that, in such cases, about-to-be candidates would wait longer to contribute an evaluation (a) to see how others evaluate the candidate, thereby reducing the perceived uncertainty of the outcome, and (b) because negative evaluations are asymmetrically perceived as being more significant.

To test this, we ran an additional regression with *Relative Time to Evaluate* as the dependent variable, measured as the number of days since the evaluation process started. We used a proportional hazards model, arguing that members tend (not) to wait, depending on various factors (Table T14 in Online Appendix T6 shows the marginal effects of that regression). Coefficients larger than zero imply that members evaluate more quickly—it takes them less time to evaluate—whereas negative

Table 6. OLS of Approval for Request to Be Promoted to Admin; DV: Approval Percentage

	Model 1 Hypothesis 1	Model 2 Hypothesis 2	Model 3 Hypothesis 3
Positive evaluations (Hypothesis 1)	0.001*** (0.000)	0.001* (0.000)	0.000 (0.000)
Negative evaluations (Hypothesis 1)	−0.002** (0.001)	−0.002* (0.001)	−0.002** (0.001)
Negative evaluations (any) (Hypothesis 1)	0.045*** (0.011)	0.057 (0.059)	0.068 (0.062)
Negative evaluations in pivotal RfAs (Hypothesis 2b)		−0.111** (0.057)	
Positive evaluations in pivotal RfAs (Hypothesis 2a)		−0.137** (0.060)	
Negative evaluations of active candidates (Hypothesis 3b)			−0.112* (0.062)
Positive evaluations of active candidates (Hypothesis 3a)			0.015 (0.063)
Year fixed effects	Yes	Yes	Yes
Candidate-level control	Yes	Yes	Yes
R ²	0.445	0.339	0.336
N	2,743	1,100	1,100

Note. Robust standard errors in parentheses.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. Two-sided tests.

coefficients indicate that they wait longer. About-to-be candidates are slower to evaluate candidates in situations with high pivotality ($\beta = -0.156, p < 0.001$) and slower to offer negative evaluations ($\beta = -0.672, p < 0.001$). Thus, we find evidence that about-to-be candidates adjust whether and how they evaluate and when.

We also run a series of robustness checks. First, we change the duration for which we assumed evaluators anticipate their own evaluation. We use a window of three and nine months and found comparable results (Online Appendix T3). Second, we use alternative model specifications to better account for user fixed-effects and selection effects (Online Appendix T4). Third, we vary which groups the members anticipating their evaluation (“About-to-be candidate”) are compared with: members who have been promoted (admins), members who are not going to be evaluated, and members who have been candidates but have not been promoted. To rule out that our results are driven by comparing them to a group that could distort them, we compare about-to-be candidates to each subgroup (Online Appendix T4.1). Fourth, we employ several alternative matching techniques to rule out that our results are driven by a specific matching technique (Online Appendix T5). Fifth, we exclude two alternative explanations, selection effects and candidate quality. For the latter we exploit a facet of the RfA process, that is, the possibility to self-nominate, to strengthen the argument for strategic behavior (Online Appendix T7). Sixth, we determine that the dependent variable categories differ and that no categories should be added or subtracted (Online Appendix T9).

Post Hoc Analysis: Linking Hypothesized Strategic Behaviors to Individual Outcomes

Our overarching argument is that the dual role structure leads members to use their evaluations to influence how they will be evaluated. A major strength of our study is that it allows us to analyze, as an additional step, whether the behaviors we study resulted in the desired outcome of being positively evaluated. Thus, we create a regression that tests the effect of following the strategic targeting behavior on the eventual outcome of a member’s own RfA. We quantify each strategy: how often about-to-be candidates evaluated other candidates positively and negatively in general (Hypothesis 1) and how often in the situations described in Hypotheses 2 and 3. To construct the variables, we use the six-month look-back window. We determine all RfAs in this window and all positive and negative evaluations of the member. We operationalize Hypothesis 1 as the number of positive and negative evaluations in that window. Next, for Hypotheses 2 and 3, we calculate the correlation between the evaluation, a dummy that takes the value one when the evaluation is negative (positive), and the pivotality of evaluation (Hypothesis 2) or the candidate’s activity (Hypothesis 3). Table 6 shows an ordinary least squares

(OLS) regression with the dependent variable being an approval percentage measured as the number of positive evaluations divided by the number of positive and negative evaluations.¹²

Results indicate that, on average, positive (negative) evaluations of other candidates increase (decrease) the likelihood of their own RfA’s approval. However, when including a dummy to indicate whether members evaluated somebody negatively at least once, the dummy has a positive effect (Table 6, Model 1). This suggests that some negative evaluations positively affect approval, but the marginal effect of more negative evaluations is negative. There is some evidence that increasing pivotal negative and positive evaluations is detrimental to a future candidate (Table 6, Model 2, $\beta_{\text{neg}} = -0.111, p = 0.050$; $\beta_{\text{pos}} = -0.137, p = 0.021$). Moreover, negative evaluations in RfAs with active candidates reduce the likelihood of being approved (Model 3, $\beta = -0.112, p = 0.069$), but positively evaluating has no benefit (Model 3, $\beta = 0.015, p = 0.816$). According to our analysis, a typical candidate increases her expected approval rating by 5% compared with a candidate not behaving strategically, at all. This relatively small percentage change can have a big effect on the success rate of the RfA, as an approval rating above 75% will be automatically successful and below 65% will not and the average approval rating is 70%. In contrast, a fully strategic member could increase their expected approval rating by about 8%. Specifically, a strategic actor could achieve this by voting positively in about 40 RfAs (a high but realistic value), voting negatively in 1 RfA, and not voting against an active candidate or on pivotal evaluations. Overall, the results suggest that acting strategically is at least partially effective in being granted adminship in the future.

Discussion

A fundamental challenge for any organization is how to evaluate its members. Such evaluations are important as the outcome often determines who gains authority within the organization, which guides resource allocation and strategic decision making (e.g., Bower 1986, Rivera 2012, Klingebiel and Rammer 2014, Boudreau et al. 2016, Fernandez-Mateo and Fernandez 2016, Bode et al. 2022). The evaluation process is important because if organizational members or other stakeholders perceive it as unfair, it can cause frustration and reduced motivation (Ambrose and Cropanzano 2003, Franke et al. 2013, Vough and Caza 2017). The evaluation criteria are important, as they reward and guide organizational members’ behavior (e.g., Castilla 2008, Rider et al. 2023). Traditionally, research has focused on managerial evaluations (Pierce et al. 2015, Castilla and Ranganathan 2020, Black and Vance 2021, Uribe et al. 2022). We extend this research by studying a common, consequential, but neglected type of evaluation: peer evaluations.

Organization Design of (Peer) Evaluations

Our research adds to the understanding of organization design by investigating the impact of peer evaluations and specific design choices on the behavior of members within an organization. Peer evaluations are particularly interesting for new forms of organizing that do not rely on a traditional hierarchy (Tushman et al. 2012, Puranam et al. 2014, Joseph et al. 2018, Zhao et al. 2021, Schweisfurth et al. 2023). Previous research has shown that these organizations develop new coordination mechanisms, such as granting members autonomy over tasks without a manager dictating what needs to be done (Fleming and Waguespack 2007, O'Mahony and Ferraro 2007, Dahlander and O'Mahony 2011). Using data from Wikipedia, we examine the process by which authority is allocated through peer evaluations, which are conducted openly and can be traced over time, making the process transparent to all participants. Our study reveals two important design features of peer evaluations: *transparency* and *self-selection*.

Choosing *transparency* within an organization can have a significant impact, with both positive and negative consequences (Bernstein 2012, Schnackenberg and Tomlinson 2016, Bernstein 2017, Obloj and Zenger 2022). On the one hand, it can lead to increased engagement and allow colleagues to monitor one another, potentially detecting any dishonest behavior (Bernstein 2012, 2017). For example, on Wikipedia, members put much effort into evaluating candidates and justifying their decisions. If another member perceives an evaluation as unfair, they can confront the evaluator and discuss the issue further. However, transparency can lead to strategic evaluations and distortions from merit-based ideals. This is because members can anticipate being evaluated and adjust their evaluations to be perceived positively. However, it is difficult to argue conclusively whether transparency is detrimental, even in peer evaluations. Future research could explore transparency in greater detail, focusing on variations in transparency. For example, some journals have experimented with single-blind reviews (Blank 1991). In some European countries, all professor applications are open, and all candidates can be scrutinized with reviews that are visible for everyone to see.

One important aspect of the peer evaluation system we researched is *self-selection*. This means that members of an organization can choose which peer evaluations they want to participate in (Botelho 2017). Recent studies have investigated the impact of self-selection on organizations (Bernstein et al. 2014, Ketkar and Workiewicz 2021, Raveendran et al. 2022), and while it can have some advantages—such as allowing people to evaluate colleagues they know well—we also found that there are downsides. For instance, our research shows that people tend to participate in evaluations only if they believe it will benefit them rather than if their evaluation would be

most helpful. This means they refrain from participating in evaluations where the outcome is uncertain to avoid any potential retaliation. Incidentally, the organization is likely to miss important evaluations when they would be most valuable, that is, the outcome is not clear. This raises a larger issue about participation in peer evaluations. It is conceivable that only politically savvy people participate, rather than those who are best qualified to provide an accurate assessment.

Consequences of (Transparent) Peer Evaluations

Our analysis delves into the impact of peer evaluations on the behavior of organizational members. We argue that the dual role of evaluating and being evaluated is what sets peer evaluations apart from other forms of information aggregation (Sah and Stiglitz 1986, Knudsen and Levinthal 2007, Christensen and Knudsen 2010, Csaszar and Eggers 2013, Piezunka et al. 2022). By examining how role duality influences behavior, we also contribute to existing research on how an individual's position within the organization shapes their evaluation, for example, due to their hierarchical level (Reitzig and Maciejovsky 2015, Keum and See 2017) or their pivotal role in the evaluation (Piezunka and Schilke 2023). On the one hand, organizational members as evaluators might refrain from evaluating if this can lead to retaliation. On the other hand, the evaluated might be negatively evaluated or not evaluated at all due to their peripheral position in the organization and thus become collateral damage.

The introduction of peer evaluations significantly impacts the relationships between individuals, rendering them multiplex (Ingram and Roberts 2000, Li and Piezunka 2020, Ertug et al. 2023). Organizational members collaborate while also evaluating each other's work. Previously, multiple relationships emerged when individuals with strong relationships in one domain extended them to another domain, leveraging their trust and knowledge. However, in the case of peer evaluations, the evaluative relationship does not need mutual agreement and is not emergent (Mindruta et al. 2016, Katila et al. 2022). Any organizational member can self-select to participate in the evaluation of another. This new evaluative layer can interfere with the collaborative layer, making future collaborations more challenging. For example, suppose a member negatively evaluates a peer with whom they have collaborated. In that case, their relationship changes from a collaborative uniplex relationship of positive valence to a multiplex bivalent relationship. This kind of relationship can result in conflict and make future collaborations difficult. Even the mere prospect of being evaluated can make people strategic about whom they collaborate with and how. Our study addresses the need for research on how multiplexity can lead to conflict.

Peer Evaluation as a Stage

Our theory and findings show that (peer) evaluations are not just an aggregation process. They also allow people to portray themselves. To use Goffman’s (1959) analogy, peer evaluations can be stage-like, where the evaluating members perform to showcase who they are and position themselves to reap later benefits. This very concept of a stage on which actors present themselves stands in contrast to the desired meritocracy where the focus is on the accuracy of the evaluation that an actor receives, not on the (strategic) evaluation that an actor provides. This insight that peer evaluations constitute a stage on which organizational members perform has wide-ranging implications.

Various research studies focus on how individuals position themselves and manage how others perceive them. These studies explore different ways, means, and tools that people use to shape their image, such as language (Srivastava et al. 2018, Piezunka and Dahlander 2019). Our study highlights that individuals in organizations have another tool at their disposal—evaluations from their peers. We can interpret evaluations as social ties individuals form to position themselves in the organization. Prior research has shown that social actors need to position themselves in social structures to succeed in organizations (Burt 1992). This typically involves having the right connections, which has been the focus of previous research on advice and friendship ties (Gargiulo 1993, Burt 2004, Reagans et al. 2004, Goldberg et al. 2016). However, our study introduces a new type of tie—evaluative ties. Our theory suggests that individuals can use evaluative ties to improve their standing. Evaluative ties are effective because they act like prisms, reflecting what others think about the individual who forms the tie (Podolny 2001). Overall, our study highlights the importance of evaluative ties in shaping perceptions in organizations. By leveraging these ties, individuals can improve their positioning within the organization and increase their chances of success.

Our study has revealed that members recognize the importance of negative evaluations in portraying themselves. Prior work suggests negative ties elicit stronger behavioral responses than positive ones (Labianca et al. 1998, Labianca and Brass 2006, Everett and Borgatti 2014, Halgin et al. 2020). It is thus plausible that as organizational members are more apprehensive about negative evaluations, they tend to avoid giving such feedback to avoid any kind of retaliation. However, while strategic peer evaluations can aid in portraying oneself, they carry the risk of being detected and penalized. As peer evaluations are not legitimized through authority but by peer acceptance of the process’s norms and procedures, any form of strategic behavior that deviates from the ideal of honest and merit-based assessment undermines the legitimacy of the overall evaluation process procedures (Merton and Storer 1973, Bourdieu 1993, Zelditch and Walker 2003, Lamont 2009). To preserve that legitimacy, peers punish those who do not comply with the norm of honest and merit-based evaluations (Fehr and Schmidt 1999, Cialdini 2001). Thus, members must understand what most others view as legitimate (Ridgeway and Correll 2006, Correll et al. 2017), which can be challenging, given the diversity of organizational members and their expectations (Cattani et al. 2014, Ertug et al. 2016, Sudaby et al. 2016, Fini et al. 2018, Aadland et al. 2019)

If peer evaluations allow members to portray themselves strategically and deviate from merit-based evaluations, it adds to the skepticism about whether organizations can deliver on their meritocratic promises (Castilla and Ranganathan 2020). Our analysis provides a view complementary to previous work (Castilla 2008), which suggests that managers in meritocratic organizations can make more biased evaluations of their employees because they feel more objective. However, our study focuses on evaluations among peers instead of managers assessing employees. While we cannot determine whether peer evaluations adhere to merit-based ideals without a clear counterfactual, the transparent peer evaluation process requires members

Table 7. Peer Evaluation Designs

Empirical examples	Parameters			
	Transparent	Self-select into evaluation	Fixed no. of positions	Notable reference(s)
Wikipedia	Y	Y	N	Jemielniak (2014)
Creative competitions	Y	Y	Y	Cattani et al. (2014)
Intrafirm peer evaluation	Y/N	Y	Y/N	Brivot et al. (2014)
360-degree feedback	Y/N	N	N	Shipper et al. (2007)
Academic peer review	Y/N	Y	Y/N	Lamont (2009)
Grant reviews (e.g., NSF, ERC)	Single-blind	Y	Y	Li and Agha (2015)
Professional awards (e.g., Oscars)	Y/N	Y	Y	Rossman and Schilke (2014)
Open-source software (e.g., GNOME)	Y	Y	N	O’Mahony and Ferraro (2007)
Peer evaluation of students	Y/N	N	N	Howard et al. (2010)
Crowdsourcing of ideas	Y/N	Y	N	Piezunka and Dahlander (2015), Bayus (2013)

Note. NSF, National Science Foundation; ERC, European Research Council.

to put in effort and engage in discussions, ultimately increasing the quality of subsequent evaluations. In the case of Wikipedia, members are typically not hesitant to be critical, possibly due to the community's pseudonymous nature. This may not hold true in traditional face-to-face organizations where members may be more reluctant to criticize each other.

Boundary Conditions

Wikipedia has features that make it essential to elaborate on the boundary conditions of our theorizing. Peer evaluations can differ along multiple dimensions, including transparency, self-selection into evaluation, and competitiveness (e.g., if the number of positions is limited). Table 7 outlines how peer evaluation settings explored by previous work differ along these dimensions. This table should not be interpreted by its frequencies of yes and no to suggest how common a feature is. Rather, it highlights to what degree we would expect our theory to hold in other settings.

As discussed, transparency is the most important dimension. In the complete absence of it, organizational members could not deploy peer evaluations to advance their cases. While Wikipedia provides complete transparency, as everyone can see how everyone is evaluated, other forms of transparency may have similar effects. Even if the evaluation process is double-blind, through a middleman people may engage in strategic behavior. In many organizations, informal networks allow important information to reach many members (Noon and Delbridge 1993, Michelson and Mouly 2004). As long as there is some degree of transparency, whether intended or not, role duality can affect evaluating behavior. Another important feature is self-selection: organizational members need the ability to choose whether to evaluate to avoid evaluating certain peers strategically and have to decide between positive and negative evaluations. In this case, the ability to be strategic is limited, and we would expect members to use more positive evaluations, as the potential retaliations from negative evaluations are asymmetrically higher. Another feature of Wikipedia is that evaluation is not directly competitive (i.e., the evaluated organizational members do not compete for a fixed and thus a limited number of positions, so that only a subset of the nominated could be promoted). One member's success is not anyone else's lost opportunity. However, strategic behavior is likely exacerbated by direct competition for a fixed number of positions, particularly negative evaluations. Inside a company with higher stakes and fiercer competition, we expect the theory to hold with a higher baseline rate of being strategic.

Limitations

There are multiple limitations to our work. First, our theory development suggests strategic behavior by

organizational members. We measure this indirectly by examining how members change their evaluations when they are about to be evaluated in comparison with when they are not. Our research design, however, *cannot* measure strategic intent. The research design we deploy (i.e., archival rather than experimental) precludes us from causally identifying individual motivations, a limitation we share with most previous work. While we consider our theory about strategic behavior to be the most plausible, we cannot fully rule out alternative explanations.

Second, we illustrate that strategic behavior distorts organizational members' evaluations. To test our theory, we examine people's behavior in a period in which strategic behavior can be expected to be particularly salient. However, we do not have a counterfactual of how members would evaluate sincerely. It is likely that organizational members also show strategic behavior in other situations. Given that we only measure the effect of a temporary increase in strategic behavior, we cannot specify the degree to which strategic evaluations differ from sincere evaluations.

We face potential endogeneity concerns as we investigate an organization's peer evaluations in real life without a real experiment. Specifically, the selection to become a candidate is not random. We addressed this concern by running a fixed effects regression (Online Appendix T5) and employing different comparison groups. There are remaining endogeneity concerns that only an experiment could address, such as reverse causality—members who act strategically see the opportunity to become a candidate.

In closing, research on peer evaluation has been held back by secrecy. We hope that researchers gain access to a wider range of peer evaluations and uncover what shapes how actors evaluate one another in these settings.

Acknowledgments

The authors thank John Joseph and three anonymous reviewers. Their feedback and guidance were tremendously helpful. The authors also thank Gianluca Carnabuci, Rudy Durand, Javier Gimeno, Joe Labianca, Matteo Prato, David Reetz, Eric Quintane, Brian Silverman, and seminar participants at Aarhus University, Rotterdam School of Management, Imperial College London, St. Gallen, and Technical University Hamburg-Harburg for their invaluable feedback.

Endnotes

¹ In line with research in political science (Margolis 2001), we use the term strategic to mean evaluating peers based on aspects other than their pure merits (cp. Piezunka and Schilke 2023).

² We discuss to what degree this atypical degree of transparency constrains our theory in "Boundary Conditions" in the discussion section.

³ For more details on RfA, please see Online Appendix T10.

⁴ See https://en.wikipedia.org/wiki/Wikipedia:Requests_for_adminship.

⁵ Jimmy Wales, 'Sysop status', posting to Wikien-I mailing list, February 11, 2003.

⁶ Wikipedians call this policy SNOW, as in the candidate would have “a snowball’s chance in hell.” We thank a reviewer for bringing this to our attention.

⁷ Sixteen of them either were currently striving to become an administrator (and were thus running an RfA) or had already tried to become administrators (i.e., they had run an RfA in the past, though not necessarily successfully). The remaining eight had participated in the RfA process for others and in some cases were considering running their own RfA in the future.

⁸ Members also have the option to vote “neutral.” In our main analysis we do not report this category for clarity reasons. Results do not change if the category is included or omitted. We do not report the results because the norms around using neutral evaluations on Wikipedia are complex and, in fact, do not resemble abstentions, as we would have expected. Neutral evaluations are often used when members comment on other evaluations, thereby expressing opinions, but want to signal that their opinion is not strong enough to evaluate a candidate. We report the results of the neutral category in Online Appendix T1.

⁹ We examined in two ways whether the six-month window captures the adequate duration for which a member anticipates her own evaluation. First, we investigated quantitatively how long a member’s anticipation can be traced (see Online Appendix T3). The results indicate that potential candidates increased their evaluation activity at least six months before an RfA. Second, we asked our interviewees how long they had thought about running for administrator before actually declaring it; the range was from a few months to approximately a year. In our robustness checks, we tested two alternative time windows—three months and nine months—and found consistent results. Any other variable that considers past activity uses the same six-month lookback window as described.

¹⁰ See <https://en.wikipedia.org/wiki/Wikipedia:Namespace>.

¹¹ We selected propensity score matching as our main matching algorithm, as coarsened exact matching tends to eliminate many treatment cases (Black et al. 2020).

¹² Descriptive statistics of the post hoc analysis can be found in Online Appendix T8.

References

- Aadland E, Cattani G, Ferriani S (2019) Friends, gifts, and cliques: social proximity and recognition in peer-based tournament rituals. *Acad. Management J.* 62(3):883–917.
- Aaltonen A, Seiler S (2016) Cumulative growth in user-generated content production: evidence from Wikipedia. *Management Sci.* 62(7):2054–2069.
- Aghion P, Tirole J (1997) Formal and real authority in organizations. *J. Political Econom.* 105(1):1–29.
- Alexy O, Poetz K, Puranam P, Reitzig M (2021) Adaptation or persistence? Emergence and revision of organization designs in new ventures. *Organ. Sci.* 32(6):1439–1472.
- Ambrose ML, Cropanzano R (2003) A longitudinal analysis of organizational fairness: An examination of reactions to tenure and promotion decisions. *J. Appl. Psych.* 88(2):266–275.
- Askin N, Bothner MS (2016) Status-aspirational pricing: The “Chivas Regal” strategy in U.S. higher education, 2006–2012. *Admin. Sci. Quart.* 61(2):217–253.
- Bacharach SB, Lawler EJ (1980) *Power and Politics in Organizations* (Jossey-Bass Inc., San Francisco).
- Bagozzi BE, Marchetti K (2017) Distinguishing occasional abstention from routine indifference in models of vote choice. *Political Sci. Res. Methods* 5(2):277–294.
- Barnard C (1968) *The Functions of the Executive* (Harvard University Press, Cambridge, MA).
- Bartling B, Fischbacher U, Schudy S (2015) Pivotality and responsibility attribution in sequential voting. *J. Public Econom.* 128:133–139.
- Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD (2001) Bad is stronger than good. *Rev. Gen. Psych.* 5(4):323–370.
- Bayus BL (2013) Crowdsourcing new product ideas over time: An analysis of the Dell Ideastorm community. *Management Sci.* 59(1):226–244.
- Becker J, Brackbill D, Centola D (2017) Network dynamics of social influence in the wisdom of crowds. *Proc. Natl. Acad. Sci. USA* 114(26):E5070–E5076.
- Benson A, Li D, Shue K (2019) Promotions and the Peter principle. *Quart. J. Econom.* 134(4):2085–2134.
- Bergeron DM, Shipp AJ, Rosen B, Furst SA (2013) Organizational citizenship behavior and career outcomes: The cost of being a good citizen. *J. Management* 39(4):958–984.
- Bernstein E, Gino F, Staats B. (2014) Opening the valve: From software to hardware (A). Harvard Business School Case 415-015. Harvard Business School, Boston.
- Bernstein ES (2012) The transparency paradox: A role for privacy in organizational learning and operational control. *Admin. Sci. Quart.* 57(2):181–216.
- Bernstein ES (2017) Making transparency transparent: The evolution of observation in management theory. *Acad. Management Ann.* 11(1):217–266.
- Bitektine A, Haack P (2015) The “macro” and the “micro” of legitimacy: Toward a multilevel theory of the legitimacy process. *Acad. Management Rev.* 40(1):49–75.
- Black DE, Vance MD (2021) Do first impressions last? The impact of initial assessments and subsequent performance on promotion decisions. *Management Sci.* 67(7):4556–4576.
- Black BS, Lalkiya P, Lerner JY (2020) The trouble with coarsened exact matching. Preprint, submitted October 6, <https://dx.doi.org/10.2139/ssrn.3694749>.
- Blank G (2007) *Critics, Ratings, and Society: The Sociology of Reviews* (Rowman & Littlefield, Plymouth, UK).
- Blank RM (1991) The effects of double-blind versus single-blind reviewing: Experimental evidence from the American economic review. *Amer. Econom. Rev.* 81(5):1041–1067.
- Blau MP (1964) *Exchange and Power in Social Life* (Transaction Publishers, London).
- Bode C, Rogan M, Singh J (2022) Up to no good? Gender, social impact work, and employee promotions. *Admin. Sci. Quart.* 67(1):82–130.
- Bolino MC, Kacmar KM, Turnley WH, Gilstrap JB (2008) A multi-level review of impression management motives and behaviors. *J. Management* 34(6):1080–1109.
- Botelho TL (2017) Audience to evaluator: The effect of social influence in evaluation processes. *Acad. Management Proc.* 2017(1):11342.
- Botelho TL, Abraham M (2017) Pursuing quality: How search costs and uncertainty magnify gender-based double standards in a multistage evaluation process. *Adm. Sci. Q.* 62(4):698–730.
- Boudreau KJ, Guinan EC, Lakhani KR, Riedl C (2016) Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Sci.* 62(10):2765–2783.
- Bourdieu P (1993) *The Field of Cultural Production: Essays on Art and Literature* (Columbia University Press, New York).
- Bower JL (1986) *Managing the Resource Allocation Process: A Study of Corporate Planning and Investment* (Harvard Business Press, Boston).
- Bowers A (2015) Relative comparison and category membership: The case of equity analysts. *Organ. Sci.* 26(2):571–583.
- Bowers A (2019) Balanced but not fair: Strategic balancing, rating allocations, and third-party intermediaries. *Strategic Organ.* 18(3):427–447.
- Brass DJ, Borgatti SP. (2019) *Social Networks at Work* (Taylor & Francis Group, Milton, UK).

- Brebels L, De Cremer D, Sedikides C (2008) Retaliation as a response to procedural unfairness: A self-regulatory approach. *J. Personality Social Psych.* 95(6):1511–1525.
- Brett JF, Atwater LE (2001) 360° feedback: Accuracy, reactions, and perceptions of usefulness. *J. Appl. Psych.* 86(5):930–942.
- Brivot M, Lam H, Gendron Y (2014) Digitalization and promotion: An empirical study in a large law firm. *Br. J. Management* 25(4):805–818.
- Burt RS (1992) *Structural Holes: The Social Structure of Competition*, 1st ed. (Harvard University Press, Cambridge, MA).
- Burt RS (2004) Structural holes and good ideas. *Amer. J. Sociol.* 110(2):349–399.
- Castilla EJ (2008) Gender, race, and meritocracy in organizational careers. *Amer. J. Sociol.* 113(6):1479–1526.
- Castilla EJ, Benard S (2010) The paradox of meritocracy in organizations. *Admin. Sci. Quart.* 55(4):543–676.
- Castilla EJ, Ranganathan A (2020) The production of merit: How managers understand and apply merit in the workplace. *Organ. Sci.* 31(4):909–935.
- Cattani G, Ferriani S, Allison PD (2014) Insiders, outsiders, and the struggle for consecration in cultural fields. *Amer. Sociol. Rev.* 79(2):258–281.
- Cheng W, Warren M (2000) Making a difference: Using peers to assess individual students' contributions to a group project. *Teach. High. Ed.* 5(2):243–255.
- Choudhury P, Crowston K, Dahlander L, Minervini MS, Raghuram S (2020) GitLab: Work where you want, when you want. *J. Organ. Design* 9(1):23.
- Christensen M, Knudsen T (2010) Design of decision-making organizations. *Management Sci.* 56(1):71–89.
- Cialdini RB (2001) *Influence: Science and Practice*, 4th ed. (Allyn and Bacon, Boston).
- Collier B, Burke M, Kittur N, Kraut R (2008) Retrospective vs. prospective evidence for promotion: The case of Wikipedia. *2008 Annual Meeting of the Academy of Management* (Academy of Management, Briarcliff Manor, NY).
- Condorcet J-A-NC, McLean I, Hewitt F (1994) *Condorcet: Foundations of Social Choice and Political Theory* (E. Elgar, Aldershot, UK).
- Correll SJ, Ridgeway CL, Zuckerman EW, Jank S, Jordan-Bloch S, Nakagawa S (2017) It's the conventional thought that counts: How third-order inference produces status advantage. *Amer. Sociol. Rev.* 82(2):297–327.
- Csaszar FA, Eggers JP (2013) Organizational decision making: An information aggregation view. *Management Sci.* 59(10):2257–2277.
- Dahlander L, Frederiksen L (2012) The core and cosmopolitans: A relational view of innovation in user communities. *Organ. Sci.* 23(4):988–1007.
- Dahlander L, O'Mahony S (2011) Progressing to the center: Coordinating project work. *Organ. Sci.* 22(4):961–979.
- Diugos K, Keller JR (2021) Turned down and taking off? Rejection and turnover in internal talent markets. *Acad. Management J.* 64(1):63–85.
- Dornbusch SM, Scott WR, Busching BC (1975) *Evaluation and the Exercise of Authority*, 1st ed. (Jossey-Bass, San Francisco).
- Ertug G, Brennecke J, Tasselli S (2023) Theorizing about the implications of multiplexity: An integrative typology. *Acad. Management Ann.* 17(2):626–654.
- Ertug G, Yögev T, Lee YG, Hedström P (2016) The art of representation: How audience-specific reputations affect success in the contemporary art field. *Acad. Management J.* 59(1):113–134.
- Ethiraj SK, Levinthal D (2009) Hoping for A to Z while rewarding only A: Complex organizations and multiple goals. *Organ. Sci.* 20(1):4–21.
- Everett MG, Borgatti SP (2014) Networks containing negative ties. *Social Networks* 38:111–120.
- Fairburn JA, Malcomson JM (2001) Performance, promotion, and the Peter principle. *Rev. Econom. Stud.* 68(1):45–66.
- Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. *Quart. J. Econom.* 114(3):817–868.
- Fernandez-Mateo I, Coh M (2015) Coming with baggage: Past rejections and the evolution of market relationships. *Organ. Sci.* 26(5):1381–1399.
- Fernandez-Mateo I, Fernandez RM (2016) Bending the pipeline? Executive search and gender inequality in hiring for top management jobs. *Management Sci.* 62(12):3636–3655.
- Filippas A, Horton JJ, Golden J (2018) Reputation inflation. *Proceedings of the 2018 ACM Conference on Economics and Computation* (Association for Computing Machinery, New York), 483–484.
- Fini R, Jourdan J, Perkmann M (2018) Social valuation across multiple audiences: The interplay of ability and identity judgments. *Acad. Management J.* 61(6):2230–2264.
- Fleming L, Waguespack DM (2007) Brokerage, boundary spanning, and leadership in open innovation communities. *Organ. Sci.* 18(2):165–180.
- Follett MP (1941) *Dynamic Administration: The Collected Papers of Mary Parker Follett* (Taylor & Francis, New York).
- Franke N, Keinz P, Klausberger K (2013) “Does this sound like a fair deal?”: Antecedents and consequences of fairness expectations in the individual's decision to participate in firm innovation. *Organ. Sci.* 24(5):1495–1516.
- Gagne F, Khan A, Lydon J, To M (2008) When flattery gets you nowhere: Discounting positive feedback as a relationship maintenance strategy. *Can. J. Behav. Sci.* 40(2):59–68.
- Gallus J (2017) Fostering public good contributions with symbolic awards: A large-scale natural field experiment at Wikipedia. *Management Sci.* 63(12):3999–4015.
- Gargiulo M (1993) Two-step leverage: Managing constraint in organizational politics. *Admin. Sci. Quart.* 38(1):1–19.
- Goffman E (1959) *The Presentation of Self in Everyday Life* (Doubleday, Oxford, UK).
- Goldberg A, Srivastava SB, Manian VG, Monroe W, Potts C (2016) Fitting in or standing out? The tradeoffs of structural and cultural embeddedness. *Amer. Sociol. Rev.* 81(6):1190–1222.
- Gouldner AW (1960) The norm of reciprocity: A preliminary statement. *Amer. Sociol. Rev.* 25(2):161–178.
- Grabner I, Moers F (2013) Managers' choices of performance measures in promotion decisions: An analysis of alternative job assignments. *J. Accounting Res.* 51(5):1187–1220.
- Greenstein S, Zhu F (2016) Open content, Linus' law, and neutral point of view. *Inform. Systems Res.* 27(3):618–635.
- Halgin DS, Borgatti SP, Huang Z (2020) Prismatic effects of negative ties. *Social Networks* 60:26–33.
- Hallen BL (2008) The causes and consequences of the initial network positions of new organizations: From whom do entrepreneurs receive investments? *Admin. Sci. Quart.* 53(4):685–718.
- Hastie R, Kameda T (2005) The robust beauty of majority rules in group decisions. *Psych. Rev.* 112(2):494–508.
- Hill M (2013) *Essays on Volunteer Mobilization in Peer Production* (Massachusetts Institute of Technology, Cambridge, MA).
- Howard CD, Barrett AF, Frick TW (2010) Anonymity to promote peer feedback: Pre-service teachers' comments in asynchronous computer-mediated communication. *J. Educ. Comput. Res.* 43(1):89–112.
- Ingram P, Roberts PW (2000) Friendships among competitors in the Sydney hotel industry. *Amer. J. Sociol.* 106(2):387–423.
- Jemielniak D (2014) *Common Knowledge? An Ethnography of Wikipedia* (Stanford University Press, Stanford, CA).
- Joseph J, Gaba V (2020) Organizational structure, information processing, and decision-making: A retrospective and road map for research. *Acad. Management Ann.* 14(1):267–302.
- Joseph J, Baumann O, Burton R, Srikanth K (2018) Reviewing, revisiting, and renewing the foundations of organization design. Joseph J, Baumann O, Burton R, Srikanth K, eds. *Organization Design. Advances in Strategic Management*, vol. 40 (Emerald Publishing Limited, Bingley, UK), 1–23.

- Kane GC, Johnson J, Majchrzak A (2014) Emergent life cycle: The tension between knowledge change and knowledge retention in open online coproduction communities. *Management Sci.* 60(12):3026–3048.
- Katila R, Piezunka H, Reineke P, Eisenhardt KM (2022) Big fish vs. big pond? Entrepreneurs, established firms, and antecedents of tie formation. *Acad. Management J.* 65(2):427–452.
- Ketkar H, Workiewicz M (2021) Power to the people: The benefits and limits of employee self-selection in organizations. *Strategic Management J.* 43(5):935–963.
- Keum DD, See KE (2017) The influence of hierarchy on idea generation and selection in the innovation process. *Organ. Sci.* 28(4):653–669.
- Keysar B, Converse BA, Wang J, Epley N (2008) Reciprocity is not give and take: Asymmetric reciprocity to positive and negative acts. *Psych. Sci.* 19(12):1280–1286.
- King G, Zeng L (2003) Explaining rare events in international relations. *Internat. Organ.* 55(3):693–715.
- Kirsch LJ (1996) The management of complex tasks in organizations: Controlling the systems development process. *Organ. Sci.* 7(1):1–21.
- Kittur A, Suh B, Pendleton BA, Chi EH (2007) He says, she says: Conflict and coordination in Wikipedia. Rosson MB, Gilmore D, eds. *SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, New York), 453–462.
- Klapper H, Reitzig M (2018) On the effects of authority on peer motivation: Learning from Wikipedia. *Strategic Management J.* 39(8):2178–2203.
- Klingebiel R, Rammer C (2014) Resource allocation strategy for innovation portfolio management. *Strategic Management J.* 35(2):246–268.
- Knudsen T, Levinthal DA (2007) Two faces of search: Alternative generation and alternative evaluation. *Organ. Sci.* 18(1):39–54.
- Laat PB (2014) From open-source software to Wikipedia: ‘backgrounding’ trust by collective monitoring and reputation tracking. *Ethics Inform. Tech.* 16(2):157–169.
- Labianca G, Brass DJ (2006) Exploring the social ledger: Negative relationships and negative asymmetry in social networks in organizations. *Acad. Management Rev.* 31(3):596–614.
- Labianca G, Brass DJ, Gray B (1998) Social networks and perceptions of intergroup conflict: The role of negative relationships and third parties. *Acad. Management J.* 41(1):55–67.
- Lamont M (2009) *How Professors Think* (Harvard University Press, Cambridge, MA).
- Lamont M (2012) Toward a comparative sociology of valuation and evaluation. *Annu. Rev. Sociol.* 38(1):201–221.
- Lazear EP, Rosen S (1981) Rank-order tournaments as optimum labor contracts. *J. Political Econom.* 89(5):841–864.
- Lee MY, Edmondson AC (2017) Self-managing organizations: Exploring the limits of less-hierarchical organizing. *Res. Organ. Behav.* 37:35–58.
- Lerner J, Lomi A (2017) The third man: Hierarchy formation in Wikipedia. *Appl. Network Sci.* 2(1):24.
- Leskovec J, Huttenlocher DP, Kleinberg JM (2010) Governance in social media: A case study of the Wikipedia promotion process. *Proceedings of the ICWSM (AAAI, Washington, DC)*, 98–105.
- Levinthal DA, Newark DA (2023) Putting the individual in the organizational context: A Carnegie perspective on decision making. *Front. Psych.* 14:1165713.
- Li D, Agha L (2015) Big names or big ideas: Do peer-review panels select the best science proposals? *Science* 348(6233):434–438.
- Li JB, Piezunka H (2020) The uniplex third: Enabling single-domain role transitions in multiplex relationships. *Admin. Sci. Quart.* 65(2):314–358.
- Lin-Healy F, Small DA (2013) Nice guys finish last and guys in last are nice: The clash between doing well and doing good. *Social Psych. Personal. Sci.* 4(6):692–698.
- Magee JC, Galinsky AD (2008) Social hierarchy: The self-reinforcing nature of power and status. *Acad. Management Ann.* 2(1):351–398.
- Maoz Z, Russett B (1993) Normative and structural causes of democratic peace, 1946–1986. *Amer. Political Sci. Rev.* 87(3):624–638.
- March JC, March JG (1977) Almost random careers: The Wisconsin school superintendency, 1940–1972. *Admin. Sci. Quart.* 22(3):377–409.
- Margolis H (2001) Pivotal voting. *J. Theoret. Politics* 13(1):111–116.
- Merton RK, Storer NW (1973) *The Sociology of Science: Theoretical and Empirical Investigations* (University of Chicago Press, Chicago).
- Michelson G, Mouly VS (2004) Do loose lips sink ships? *Corporate Comm.* 9(3):189–201.
- Mindruta D, Moeen M, Agarwal R (2016) A two-sided matching approach for partner selection and assessing complementarities in partners’ attributes in inter-firm alliances. *Strategic Management J.* 37(1):206–231.
- Nagaraj A (2018) Does copyright affect reuse? Evidence from Google Books and Wikipedia. *Management Sci.* 64(7):3091–3107.
- Nai J, Lin Y, Kotha R, Vissa B (2022) A foot in the door: Field experiments on entrepreneurs’ network activation strategies for investor referrals. *Strategic Management J.* 43(2):323–339.
- Noon M, Delbridge R (1993) News from behind my hand: Gossip in organizations. *Organ. Stud.* 14(1):23–36.
- Obloj T, Sengul M (2020) What do multiple objectives really mean for performance? Empirical evidence from the French manufacturing sector. *Strategic Management J.* 41(13):2518–2547.
- Obloj T, Zenger T (2017) Organization design, proximity, and productivity responses to upward social comparison. *Organ. Sci.* 28(1):1–18.
- Obloj T, Zenger T (2022) The influence of pay transparency on (gender) inequity, inequality and the performance basis of pay. *Nat. Human Behav.* 6(5):646–655.
- O’Mahony S, Ferraro F (2007) The emergence of governance in an open source community. *Acad. Management J.* 50(5):1079–1106.
- Padgett JF, Ansell CK (1993) Robust action and the rise of the Medici, 1400–1434. *Amer. J. Sociol.* 98(6):1259–1319.
- Peress M (2011) Securing the base: Electoral competition under variable turnout. *Public Choice* 148(1):87–104.
- Peter LJ, Hull R (1969) *The Peter Principle* (Souvenir Press, London).
- Pierce L, Snow DC, McAfee A (2015) Cleaning house: The impact of information technology monitoring on employee theft and productivity. *Management Sci.* 61(10):2299–2319.
- Piezunka H, Dahlander L (2015) Distant search, narrow attention: How crowding alters organizations’ filtering of suggestions in crowdsourcing. *Acad. Management J.* 58(3):856–880.
- Piezunka H, Dahlander L (2019) Idea rejected, tie formed: Organizations’ feedback on crowdsourced ideas. *Acad. Management J.* 62(2):503–530.
- Piezunka H, Schilke O (2023) The dual function of organizational structure: Aggregating and shaping individuals’ votes. *Organ. Sci.* 34(5):1914–1937.
- Piezunka H, Aggarwal VA, Posen HE (2022) The aggregation–learning trade-off. *Organ. Sci.* 33(3):1094–1115.
- Piskorski MJ, Gorbatai A (2017) Testing Coleman’s social-norm enforcement mechanism: Evidence from Wikipedia. *Amer. J. Sociol.* 122(4):1183–1222.
- Podolny JM (2001) Networks as the pipes and prisms of the market. *Amer. J. Sociol.* 107(1):33–60.
- Pollack JM, Garcia R, Michaelis TL, Hanson S, Carr JC, Sheats L (2021) Pursuing B Corp certification: Exploring firms’ entrepreneurial orientation and prosocial motivation. *Acad. Management Discov.* 7(2):294–316.
- Proost K, Schreurs B, De Witte K, Derous E (2010) Ingratiation and self-promotion in the selection interview: The effects of using single tactics or a combination of tactics on interviewer judgments. *J. Appl. Social Psych.* 40(9):2155–2169.
- Puranam P, Håkansson DD (2015) Valve’s way. *J. Organ. Design* 4(2):2–4.

- Puranam P, Alexy O, Reitzig M (2014) What's "new" about new forms of organizing? *Acad. Management Rev.* 39(2):162–180.
- Raveendran M, Puranam P, Warglien M (2022) Division of labor through self-selection. *Organ. Sci.* 33(2):810–830.
- Reagans R, Zuckerman E, McEvily B (2004) How to make the team: Social networks vs. demography as criteria for designing effective teams. *Admin. Sci. Quart.* 49(1):101–133.
- Reitzig M, Maciejovsky B (2015) Corporate hierarchy and vertical information flow inside the firm: A behavioral view. *Strategic Management J.* 36(13):1979–1999.
- Richter V, Janjic R, Klapper H, Keck S, Reitzig M (2023) Managing exploration in organizations: The effect of superior monitoring on subordinate search behavior. *Strategic Management J.* 44(9):2226–2254.
- Rider CI, Wade JB, Swaminathan A, Schwab A (2023) Racial disparity in leadership: Evidence of evaluative bias in the promotions of National Football League coaches. *Amer. J. Sociol.* 129(1):227–275.
- Ridgeway CL, Correll SJ (2006) Consensus and the creation of status beliefs. *Social Forces* 85(1):431–453.
- Rivera LA (2012) Hiring as cultural matching: The case of elite professional service firms. *Amer. Sociol. Rev.* 77(6):999–1022.
- Rosenbaum PR, Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* 79(387):516–524.
- Rossman G, Schilke O (2014) Close, but no cigar: The bimodal rewards to prize-seeking. *Amer. Sociol. Rev.* 79(1):86–108.
- Saavedra R, Kwun SK (1993) Peer evaluation in self-managing work groups. *J. Appl. Psych.* 78(3):450–462.
- Safadi H, Johnson SL, Faraj S (2021) Who contributes knowledge? Core-periphery tension in online innovation communities. *Organ. Sci.* 32(3):752–775.
- Sah RK, Stiglitz JE (1986) The architecture of economic systems: Hierarchies and polyarchies. *Amer. Econom. Rev.* 76(4):716–727.
- Schnackenberg AK, Tomlinson EC (2016) Organizational transparency: A new perspective on managing trust in organization-stakeholder relationships. *J. Management* 42(7):1784–1810.
- Schweisfurth TG, Schöttl CP, Raasch C, Zaggl MA (2023) Distributed decision-making in the shadow of hierarchy: How hierarchical similarity biases idea evaluation. *Strategic Management J.* 44(9):2255–2282.
- Setia P, Rajagopalan B, Sambamurthy V, Calantone R (2012) How peripheral developers contribute to open-source software development. *Inform. Systems Res.* 23(1):144–163.
- Shah SK, Nagle F (2020) Why do user communities matter for strategy? *Strategic Management Rev.* 1(2):305–353.
- Shaw A, Barakzai A, Keysar B (2019) When and why people evaluate negative reciprocity as more fair than positive reciprocity. *Cogn. Sci.* 43(8):e12773.
- Shermon A, Moeen M (2022) Zooming in or zooming out: Entrants' product portfolios in the nascent drone industry. *Strategic Management J.* 43(11):2217–2252.
- Shipper F, Hoffman RC, Rotondo DM (2007) Does the 360 feedback process create actionable knowledge equally across cultures? *Acad. Management Learn. Ed.* 6(1):33–50.
- Simon HA (1947) *Administrative Behavior; a Study of Decision-Making Processes in Administrative Organization* (Macmillan, Oxford, England).
- Skowronski JJ, Carlston DE (1989) Negativity and extremity biases in impression formation: A review of explanations. *Psych. Bull.* 105(1):131–142.
- Sorenson O, Stuart T (2001) Syndication networks and the spatial distribution of venture capital investments. *Amer. J. Sociol.* 106(6):1546–1588.
- Srivastava SB, Goldberg A, Manian VG, Potts C (2018) Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. *Management Sci.* 64(3):1348–1364.
- Suddaby R, Bitektine A, Haack P (2016) Legitimacy. *Acad. Management Ann.* 11(1):451–478.
- Taylor SE (1991) Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psych. Bull.* 110(1):67–85.
- Tushman M, Lakhani KR, Lifshitz-Assaf H (2012) Open innovation and organization design. *J. Organ. Design* 1(1):24–27.
- Uribe J, Carnahan S, Meluso J, Austin-Breneman J (2022) How do managers evaluate individual contributions to team production? A theory and empirical test. *Strategic Management J.* 43(12):2577–2601.
- Vaast E, Levina N (2015) Speaking as one, but not speaking up: Dealing with new moral taint in an occupational online community. *Inform. Organ.* 25(2):73–98.
- Venkataramani V, Dalal RS (2007) Who helps and harms whom? Relational antecedents of interpersonal helping and harming in organizations. *J. Appl. Psych.* 92(4):952–966.
- Vough HC, Caza BB (2017) Where do I go from here? Sense-making and the construction of growth-based stories in the wake of denied promotions. *Acad. Management Rev.* 42(1):103–128.
- Waldman DA, Atwater LE, Antonioni D (1998) Has 360 degree feedback gone amok? *Acad. Management Perspect.* 12(2):86–94.
- Weaver RK (1986) The politics of blame avoidance. *J. Public Policy* 6(4):371–398.
- Willer R, Sharkey A, Frey S (2012) Reciprocity on the hardwood: Passing patterns among professional basketball players. *PLoS One* 7(12):e49807.
- Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW (2010) Evidence for a collective intelligence factor in the performance of human groups. *Science* 330(6004):686–688.
- Zelditch M, Walker H (2003) The legitimacy of regimes. *Power Status* 20:217–249.
- Zhang X, Zhu F (2011) Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *Amer. Econom. Rev.* 101(4):1601–1615.
- Zhao HH, Li N, Harris TB, Rosen CC, Zhang X (2021) Informational advantages in social networks: The core-periphery divide in peer performance ratings. *J. Appl. Psych.* 106(7):1093–1102.

Helge Klapper is an assistant professor of strategy at the Daniels School of Business at Purdue University. His research focuses on organization design and organizational decision making with a particular interest in nonhierarchical forms of organizing. He received a PhD from the University of Vienna and was an assistant professor at Rotterdam School of Management.

Henning Piezunka is an associate professor of entrepreneurship and family enterprises at INSEAD, Fontainebleau, France. His research focuses on crowdsourcing, organizational design, and competition. He received a PhD from Stanford University.

Linus Dahlander is a professor at ESMT Berlin and the holder of the Lufthansa Group Chair in Innovation. His research focuses on networks, communities, and innovation. He received a PhD from Chalmers University of Technology and was a postdoctoral fellow at Stanford University.