Received: 17 August 2021 | Revised: 14 April 2022 | Accepted: 25 April 2022 | Published on: 4 June 2022

DOI: 10.1002/smj.3433

RESEARCH ARTICLE

STRATEGIC MANAGEMENT JOURNAL

WILEY

How do managers evaluate individual contributions to team production? A theory and empirical test

Jose Uribe[1] | **Seth Carnahan**[2] | **John Meluso**[3] | **Jesse Austin-Breneman**[4]

[1]Ross School of Business, University of Michigan, Ann Arbor, Michigan, USA

[2]Olin Business School, Washington University, St. Louis, Missouri, USA

[3]Vermont Complex Systems Center, University of Vermont, Burlington, Vermont, USA

[4]Department of Mechanical Engineering, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Jose Uribe, Ross School of Business, University of Michigan, 701 Tappan Avenue, Ann Arbor, MI 48109, USA.
Email: jnu@umich.edu

Abstract

Research Summary: Organizations rely on subjective evaluations to reward employees for team-based performance. However, it is unclear how supervisors determine individuals' contributions to collective output. We theorize that supervisors rely on the covariance between employees' presence and their teams' productivity. If teams are more productive when an employee is present, the supervisor may infer a greater contribution from the employee. Using data from a manufacturing firm, we find that covariation between an employee's presence and her team's output has a positive effect on her evaluation. This relationship is stronger when supervisors have more opportunities to observe an employee across various teams and when the employee has more authority to direct team production, supporting counterfactual information as an important component of evaluations for individuals engaged in team production.

Managerial Summary: It is notoriously difficult to evaluate the individual performance of employees when the only available metric is team-based output. We suggest that supervisors help solve this problem by observing how team output correlates with changes in

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2022 The Authors. *Strategic Management Journal* published by John Wiley & Sons Ltd.

team membership. We construct a measure of the covariance between an employee's presence in a team and the team's productivity, and find a positive relationship between this measure and the employee's annual subjective performance evaluation. Our results indicate that subjective evaluations reflect individual contributions to team production fairly well for employees who (a) have sufficient authority to direct team production and (b) are frequently rotated beyond a single team. We discuss what kinds of organizations might benefit from this measure as an input to their performance evaluation processes.

KEYWORDS

subjective evaluations, team performance

1 | INTRODUCTION

Strategy researchers increasingly seek to understand how firms can successfully aggregate the knowledge, skills, and abilities of individual employees into team- and firm-level resources that contribute to competitive advantage (Kryscynski, Coff, & Campbell, 2021; Ployhart & Moliterno, 2011). This line of work describes many mechanisms that managers can use to convert individual employees' talents into an integrated whole that is highly productive (Hamilton, Nickerson, & Owan, 2003) and difficult for rivals to copy (Rumelt, 1984). An underlying theme of this work is that firms can benefit from creating more interdependencies between an employee, her coworkers (Bermiss & Murmann, 2015; Cooper & Gubler, 2020; Di Stefano, Grohsjean, & Gutierrez, 2020; Raffiee & Byun, 2020), and other organizational resources (Kehoe & Bentley, 2019). Not only might these interdependencies generate complementarities that make an employee more productive, these interdependencies might also make it more difficult for rivals to discern how much the employee contributes to the firm's performance (e.g., Acemoglu & Pischke, 1999), thus increasing the amount of value that the firm can capture from the employee's productivity (Ethiraj & Garg, 2012).

As Coff (1997) emphasizes, the difficulty of discerning employees' individual contribution to group-level output extends beyond rivals, to the focal firm itself (see also Groysberg, Lee, & Nanda, 2008; Huckman & Pisano, 2006). Alchian and Demsetz (1972) refer to the challenge of measuring individual contributions to group-level output as the "performance non-separability problem." They famously describe how it is difficult for a manager to evaluate the individual productivity of two people moving cargo, a seemingly simple team production task, because there is no clear measure of individual output. Firms usually try to solve the performance non-separability problem by employing subjective performance evaluations (Baker, Gibbons, & Murphy, 1994), whereby an employee receives a numerical rating based on her supervisor's perception of her performance.

However, the existing literature contains few theories about how managers, in practice, might discern their employees' individual contribution to team-level output. In fact, the strategy and economics literatures mostly focus on policies that firms may implement in order to avoid or simplify the challenge of discerning individual contributions to team production, including rewarding only extreme individual performance (Holmstrom, 1982; Zenger, 1992), creating smaller organizational units (Zenger, 1994; Zenger & Hesterly, 1997), and, if available, relying solely on a worker's individual outputs (Sarsons, 2017). The vast literature in organizational psychology about performance evaluations explicitly avoids theorizing about how supervisors might arrive at performance ratings in a team production context. For example, Levy and Williams' (2004, p. 894) influential review states that "[u]nfortunately, while a good bit of work has been conducted on the metrics involved in team performance measurement, very little empirical work has been conducted regarding how to develop and implement appraisals in this context despite the great influx of team-based work environments."[1] Thus, despite the fact that performance evaluations often exist in order to match employees' rewards to their performance when measures of individual productivity are not available, the process by which supervisors solve Alchian and Demsetz's (1972) performance non-separability problem remains undertheorized.

This problem is vitally important. In its current state, the literature that concerns the strategic management of teams encourages managers to develop groups of employees with high levels of interdependence, while providing limited guidance on the ways that managers might cope with the practical challenge of evaluating individuals who operate interdependently. If firms are unable to match employees' rewards to their contributions, they risk demotivating or losing valuable employees (e.g. Gartenberg & Wulf, 2017; Larkin, Pierce, & Gino, 2012), which means that it is critical to shed light on how managers might understand employee's individual contributions to team production.

In this paper, we seek to make initial progress toward an understanding of how managers perform subjective performance evaluations when they lack clear measures of individual productivity because employees produce output in teams. We propose that supervisors use counterfactual information from their observation of fluctuations in team membership, stemming from workers' days off, absences, and team rotations, to inform their beliefs about workers' individual contributions to team output. We argue that fluctuations in team membership facilitate supervisors' subjective performance evaluations, because supervisors accumulate information about the covariance between a worker's presence on the team and the team's output. For example, if a team performs better when a worker is present and worse when she is absent, the supervisor may conclude that the worker makes a valuable contribution to the team's production. But if the team performs *worse* when the worker is present and *better* when she is absent, the supervisor may conclude that the worker drags down the team's productivity. This kind of counterfactual logic has a strong tradition in social psychology (e.g. Kahneman & Miller, 1986; Roese, 1997), and it guides empirical research designs that seeks to understand how star employees affect the performance of their peers (Azoulay, Graff Zivin, & Wang, 2010; Chen & Garg, 2018; Oettl, 2012; Stuart, 2017). To our knowledge, however, it has not been used to explain how supervisors discern workers' contributions to team production.

We test the implications of this theory using detailed data from a manufacturing firm where teams of five to seven employees carry out production. We inform our quantitative analyses with 20 interviews with company personnel and dozens of hours of site visits, in order to be

[1]Arvey and Murphy's influential review (1998, p. 142) notes that "we explicitly deal with individual job performance and do not consider group or team performance or individual performance within teams." DeNisi and Smith (2014) provide an updated review, with a similar absence of evaluation processes in teams.

certain that our empirical tests and interpretations are consistent with the reality in the company. We use the term "inferred productivity" to refer to the information about individual worker's productivity that supervisors derive from fluctuations in team membership. We operationalize inferred productivity using a standard econometric technique, namely the calculation of yearly individual worker fixed effects from team-level productivity data (for a similar approach see Bertrand & Schoar, 2003, among others). We find that higher inferred productivity has a positive correlation with subjective evaluations. The effect is robust to time-invariant worker and supervisor fixed effects and to various controls known to shape evaluations and replicates using an alternative dependent variable. We carry out two mechanism/channel checks to support the argument that supervisors incorporate counterfactual information from team membership fluctuations into their evaluations. We find that the relationship between inferred productivity and evaluations is stronger when the supervisor has more opportunities to observe counterfactual information and when the worker has sufficient authority to credibly influence team production.

This paper provides a critical contribution to the strategy literature. Scholars have suggested that firms might respond to the performance non-separability problem by simply avoiding or downplaying the measurement of individual performance in teams. Instead, firms might only reward or punish individuals based on team output (Bandiera, Barankay, & Rasul, 2010; Holmstrom, 1982; McAfee & McMillan, 1991), or firms might only try to discern extremely high performers and extremely low performers (Lazear & Rosen, 1981; Zenger, 1992). Yet subjective performance appraisals continue to be a major managerial tool in team-based settings, and the process by which supervisors ultimately arrive at their subjective evaluations of their subordinates' contributions to team production remains undertheorized.

We introduce and test a theory of how supervisors arrive at subjective performance evaluations when performance is team-based, relying on the idea that supervisors can use fluctuations in team membership to ascertain the covariance between a worker's presence and the productivity of her team. We complement the growing body of work that exhorts managers to convert individual-level knowledge, skills, and abilities into complex, group-level resources by explaining how managers might subsequently *disaggregate* group-level output into individual-level contributions. With this contribution in mind, it is important to emphasize that the goal of this paper is to advance a simple theory and to provide robust, albeit correlational, empirical tests of the theory. We want to be clear that our data and empirical analyses cannot support claims of causality (see Sevcenko & Ethiraj, 2018, for a recent example of a similar approach).

2 | WHY COVARIANCE BETWEEN A WORKER'S PRESENCE AND HER TEAM'S OUTPUT MIGHT AID SUBJECTIVE PERFORMANCE EVALUATIONS

We suggest that supervisors might use a simple causal model to help them solve the performance non-separability problem when evaluating subordinates' contributions to team output. We posit that supervisors may track, perhaps even subconsciously, the way that teams' output correlates with the presence or absence of the focal subordinate. If supervisors observe that a team's output is higher when a subordinate is present and lower when the subordinate is absent, these data points might help to convince a supervisor that a subordinate is an important contributor to group productivity. If, on the other hand, the supervisor observes that team productivity is *higher* when the subordinate is absent and *lower* when the subordinate is present, these data points might help to convince the supervisor that the subordinate is a drag on group productivity. The subordinate's presence or absence from a group might stem from a variety of sources, including planned group rotations, sick leave, vacation, and more. This kind of counterfactual, correlational logic, whereby an observer is more likely to believe that X causes Y when X and Y co-occur has a long history (e.g. David Hume, 1739). Many theories—such as attribution theory (Kelley, 1973), associative learning theory (Shanks, 1995), contingency theory (Ward & Jenkins, 1965), and causal power theory (Cheng, 1997)—rely on it.

Anecdotal evidence of the use of counterfactuals to evaluate individual members of work groups abounds, in part because of the well-documented human tendency, emphasized by Ross (1977), to attribute outcomes to people (in this case, the focal worker) instead of other stimuli (e.g., other work conditions when production is observed). The most famous example might be Steve Jobs. Supporters of Jobs often note how Apple improved when he returned to the company in 1996 (e.g. Weinberger & Hartmans, 2020), while detractors note that the company performed poorly under his leadership in the early 1980s (e.g. Uttal, 1985). Athletes sometimes hurry back from injury if their team succeeds during their absence, for fear that they will be deemed expendable.[2] Employees who lose their jobs sometimes remark that they hope the organization's performance declines upon their departure, so that their former colleagues and bosses will appreciate their contributions.

To illustrate the idea, we extend the classic example introduced by Alchian and Demsetz' (1972). The original example in Alchian and Demsetz (1972) presupposes a supervisor evaluating a hypothetical two-worker team lifting cargo. In their stylized example, the only observation made by the supervisor is the total amount of cargo jointly loaded by the two workers per day. In our extended, perhaps more realistic, example, the firm employs more than two workers, and the supervisor observes each worker working in different two-person teams throughout the year. These repeated observations with different outcomes and with different coworkers provide the supervisor with a wide array of data points. Our central claim is that, by comparing, again subconsciously or implicitly, the amount of cargo that is loaded by teams that include the focal worker to the amount of cargo loaded by teams that do not include the focal worker, the supervisor can begin to isolate each worker's contribution to joint production.

When a focal worker is absent from their usual team, we make no assumptions about the relative quality of replacement workers.[3] This is because the variance in ability of the replacement worker is part of what allows the counterfactual inference of the supervisor. When a better employee steps in, the supervisor can implicitly tell that the focal employee is less skilled than the replacement, and the opposite is true if the replacement is less skilled. In the company that we study, one supervisor shared that "when you need five [employees] for an overtime shift but only four [employees] can stay, you need to pull an extra from another line. You can quickly tell if that person is better quality or not [than the original team member]." Over time a

[2]A clear example of this logic is provided by sports journalist Jared Woodcox when analyzing the contributions of NBA player Derrick Favors to the 2017 performance of the Utah Jazz basketball team, "...some may wonder, is Derrick Favors truly vital to this team's success? Or is he proving to be an expendable piece? There's certainly an argument to be made about Derrick's expendability. The Jazz are posting a net rating of 5.8 with him off the court as compared to 3.6 with him on. Favors' detriment has been most noticeable on offense where the Jazz post an offensive rating of 105.1 with him on and 108.1 with him off." Retrieved from https://thejnotes.com/2017/03/13/has-derrick-favors-become-expendable-or-is-he-the-utah-jazzs-missing-link/.

[3]While the relative quality of any individual replacement worker is unknown a priori, the full sample of replacement workers is neither systematically better nor worse than the sample of workers being replaced. This is because both samples are drawn from the same population and have similar training.

supervisor can infer individual productivity by observing the correlation between the presence and absence of a worker from the team, along with changes in team productivity. We propose that employees with higher inferred productivity values will tend to receive higher performance evaluations from their supervisors.

Hypothesis (H1). *An employee's inferred productivity has a positive relationship with the yearly performance evaluation score that the employee receives from their supervisor.*

3 | RESEARCH SETTING

We test this core hypothesis using data from a medium sized manufacturer of healthcare products in the Midwest region of the United States. The quantitative data include annual performance evaluations and daily data on payroll and team production for the period 2012–2019. We complement these data with information gathered through multiple site visits and 20 interviews with company personnel in order to understand the relevant processes thoroughly, to inform our modeling of the team production and evaluation processes, and to help us interpret the empirical results. Table A1 contains representative data from our interviews and fieldwork that shed light on our theory and empirical approach.

3.1 | Team production at the company

The company organized production via teams consisting of 5–7 workers who process raw materials into finished goods using highly specialized machinery. The company used six production lines to manufacture dozens of products, and due to increasing demand for the company's goods, many of the production lines ran continuously in three shifts of 8 h.

The employees in a production team included an operator, an assistant operator, a helper, and two to four crew members. Operators and assistant operators had the most authority and skill among the members of the team. They directed their teammates' activities and solved complex problems related to the machinery. Helpers were middle-skilled workers who conducted some routine technical tasks, such as fixing problems with the machinery and re-configuring the machinery so that it could produce a different product. Crew members' main tasks consisted primarily of routine manual labor, such as placing raw materials into the machinery and removing finished goods from the machinery.

While the company typically assigned a worker to a primary production team, customer demand, absences and other pressures often required reassigning workers to different lines or different shifts. For example, a worker's standard routine might be to work the first shift (7 a.m.–3 p.m.) on the first line, but she might work a different line and/or a different shift, depending on company needs. As a consequence, over 80% of the worker-years in our sample worked in at least one additional team beyond their primary team at some point during the year.

The teams operate with a mixture of sequential and reciprocal interdependence (Van de Ven, Delbecq, & Koenig Jr, 1976). Crew members usually engage in sequentially interdependent work because they place finished products in boxes, move boxes of finished products from the machine to palettes, and wrap palettes for transport to a warehouse, with relatively little coordination with other team members. Helpers, assistant operators, and operators engage in reciprocally interdependent work because they have well-defined roles, but their tasks do not unfold in well-defined steps. They collaborate to monitor and load raw materials; monitor product quality; adjust machinery; transition the machines from producing one product to another; and troubleshoot any problems that arise.

The company measures output at the "production event" level of analysis, a line × shift × product level combination. The company is unable to track output at the individual worker level of analysis. Therefore, the company faces the classic performance non-separability problem, whereby it cannot discern individual worker performance from output data. Like many other organizations with this problem, it relies on yearly subjective performance evaluations to determine individual worker rewards.

3.2 | Subjective performance evaluations at the company

Each worker received, on a yearly basis, a subjective performance evaluation from his or her production supervisor. In the period under study, supervisors attended a short training each year provided by HR to understand the purpose the performance evaluations. The firm did not use any production numbers in the evaluation process.[4] Rather, supervisors developed workers' performance scores via an unstructured, qualitative process informed by supervisors' direct observation of workers' individual effort and ability on the shop floor. In line with similar firms, supervisors came up with a numerical score based on their subjective assessment of the employee's performance, and communicated this score to the employee through a short, in-person meeting. According to company management and our analysis of company data, the performance score was highly influential in promotion and bonus decisions.

Supervisors reported that they primarily relied on direct observation of production workers' performance to arrive at their evaluation scores. Supervisors spent about two-thirds of their time walking the production floor, interacting with production employees and resolving technical and human resource matters that came up. Most of the supervisors started with the company as workers on the production floor, which gave them first-hand knowledge of their subordinates' various tasks.

4 | DATA AND ANALYSIS

We obtained detailed data for 21,888 production events from 2012 to 2019. A production event is defined by a production date (n = 1,588), an item type (n = 182) produced on a specific shift in one of the six lines. We also obtained payroll data with daily entries for the 178 workers for whom we have annual performance evaluation scores. Crucially, the company updates payroll data daily to reflect in which production events the worker actually participated on a specific date.

4.1 | Estimating the covariance between a worker's presence and team output

The company measured output using the total number of units of a specific item passing quality control, called "good units." Figure A1 shows a histogram of the *Good units (log)* dependent

[4]According to our interviews, the only "hard data" supervisors regularly used in their evaluations were attendance records.

variable. To estimate a worker's contribution to the number of good units produced during the year, net of teammate quality and the challenge of different types of production, we estimate yearly individual worker fixed effects. We argue that supervisors' process of inferring workers' productivity from counterfactuals can be captured, at least directionally, through this well-established econometric technique.[5]

The data contain 121,842 daily worker × production unit observations. We construct these data using the 21,887 production events, and we expand that data to include an observation for each worker involved in the production event. This corresponds to the approach used in Bertrand and Schoar (2003) and Pierce, Wang, and Zhang (2020), (see Table A1, Panel B in that work). We estimate annual worker fixed effects as follows:

$$Y_{(ij)kdw} = \beta_0 + WIP_{wt} + \eta_i + \delta_j + \omega_k + \gamma_{(ij)k} + \phi_{\text{month}} + \varepsilon_{(ij)kdp} \tag{1}$$

Y is the log of the number of good units of item k produced on date d by the team working on line i on shift j and which included worker w. β_0 is an overall intercept.

WIP_{wt} is the Worker's *Inferred Productivity*, our quantity of interest; it is a worker $w \times$ year t fixed effect that we will correlate with the worker's performance evaluation score. WIP_{wt} captures the average increase or decrease in productivity that accompanies a worker's participation in a production event in a given year, net of the covariates in Equation 1. Figure A2 shows the distribution of WIP_{wt}.

Given that each production line has different machinery, the model includes fixed effects for each of the six specific lines (η_i). We include three shift-level fixed effects (δ_j) because the most productive workers generally prefer and get assigned to the first work shift (7 a.m.–3 p.m.), and less productive workers get assigned to the later shifts (3 p.m.–11 p.m. and 11 p.m.–7 a.m.). We include 182 item-type fixed effects (ω_k) to control for heterogeneity in the difficulty of manufacturing the different items. We also included $\phi_{\text{month}\times\text{year}}$, which are time dummies for each month-year pair to control for secular trends and idiosyncratic company events in different months. $\gamma_{(ij)kd}$ contains covariates listed in Section A1.1 of the Appendix. Table A2 shows descriptive statistics and bivariate correlations between all control variables in the production data. Table A3 contains estimates of the model from Equation 1.

4.2 | Correlating workers' inferred productivity with performance evaluation scores

The average worker participated in 144 production events during the year. We dropped a small number of workers from the payroll data who had worked for fewer than 10 days during the year.[6] The final sample of annual performance evaluations comprised 165 workers for the period 2012–2019 (N = 565 worker-year observations). In order to examine the relationship

[5]This approach follows the spirit of John Rust's (1987) empirical model, where the author demonstrates how decisions about the replacement of bus engines made by the superintendent of maintenance at the Madison Metropolitan Bus Company (Harold Zurcher) could be approximated using a dynamic programming optimization algorithm.

[6]Results are robust to a wide range of choices around this cutoff, such as including all workers, or excluding workers with fewer than 30 working days.

between workers' inferred productivity and their performance evaluation scores, we estimate the following equation using OLS:

$$\text{Evaluation score}_{wst} = \beta_0 + WIP_{wt} + \gamma_{wt} + \alpha_s + \phi_t + \varepsilon_{wst} \quad (2)$$

where w indexes workers who earn a score, s indexes supervisors who provide scores, and t indexes years. α_s is a supervisor fixed effect, ϕ_t is a vector of year dummies, and ε_{wst} is the error term, which we cluster by supervisor in most regressions, in order to account for the non-independence of observations from the same supervisors (e.g., Castilla, 2011; Frederiksen, Kahn, & Lange, 2020). Clustering standard errors by worker produced similar results. WIP_{wt} is the worker inferred productivity described in Equation 1. The data allow us to include control variables in γ_{wt} which, based on our interviews and review of the literature, might correlate both with a worker's inferred productivity and evaluation score, allowing us to reduce the importance of alternative explanations.

4.2.1 | Performance evaluation covariates

Length of relationship records the number of years of co-tenure at the company between the supervisor and the employee and helps account for favoritism, which might cause favorable work assignments (thus higher WIP_{wt}) and favorable performance evaluations (e.g., Sundvik & Lindeman, 1998).

Avg. joint experience records the average number of previous production runs (in hundreds) in which the focal individual participated with her current teammates. This variable captures a worker's embeddedness, interdependent routines, and social capital.

Overtime work records the percentage of total hours worked composed of overtime by worker w in year t. Managers at the focal firm reported that overtime opportunities depend on seniority, so this variable may also capture some favoritism by supervisors who offer more overtime to workers that they like.

Female worker records whether the employee being rated was a woman (coded as 1) or a man (coded as zero). All the supervisors in the company are men; all but one are White.

Minority group member measures whether a worker belonged to a minority group (Black, Hispanic or Asian; coded as 1), or not (0).[7] Workers who are ascriptively dissimilar from their supervisor may receive worse job assignments and worse performance ratings (e.g., Elvira & Town, 2001).

Absenteeism (%) records the percent of days on payroll in year t where worker w was absent from their production unit for any reason. This variable helps account for supervisors' perceptions of the commitment of workers (e.g. Somers, 1995).

Time on rotation (%) is a continuous variable measuring the proportion of hours that a worker spent outsider of her home team during the year. Rotations provide supervisors with more counterfactual information because workers are observed under a greater variety of production conditions and team configurations.

[7]Results of interest do not change when analyzing the three reported minority categories separately (Black, Asian, or Hispanic), or when omitting evaluations filled out by the single non-White supervisor (a Black man). These results are available from the authors upon request.

TABLE 1 Summary statistics and correlations for performance evaluation data

		Mean	SD	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1.	Performance evaluation	3.33	0.55									
2.	Inferred productivity	0.01	0.15	−.10								
3.	Length of relationship	5.69	5.04	.21	.05							
4.	Avg. joint experience	1.26	1.36	.20	.06	.46						
5.	% Absenteeism	0.07	0.04	−.00	−.02	.33	.25					
6.	Female worker	0.30	0.46	.06	.05	.15	−.01	.09				
7.	Minority group member	0.25	0.43	−.07	.03	−.10	−.02	−.01	.20			
8.	% Overtime	0.05	0.06	−.02	.05	−.11	−.01	−.29	−.29	−.15		
9.	Job level	1.06	1.05	.08	−.04	.10	.19	.10	−.49	−.23	.33	
10.	Time on rotation (%)	0.20	0.22	.06	−.00	−.15	−.40	−.09	.05	.03	.04	−.22

Note: Correlation values over .085 are statistically significant at the .05 level. Based on workers who participated in at least 10 production events $N = 165$ workers; 541 worker-year evaluations.

Job level is a continuous variable, which ranges from 0–3. Workers with greater authority have the ability to direct the activities of lower authority workers. In descending order of authority, workers functioned in the role of *Operator*, *Assistant Operator*, *Helper* and *Crew member*. *Job level* is weighted by the percentage of working days in the year that a worker spent as crew, helper, assistant operator or operator.[8]

5 | RESULTS

Table 1 shows summary statistics and correlations for the 541 worker-years where the worker had at least 10 working days in the payroll data. We see that *Inferred productivity* (WIP_{wt}) has a weak, negative correlation with performance evaluations in the raw data.[9] Performance evaluations are positively correlated with longer relationships with the supervisor and occupying a higher job. Other controls are not highly correlated with evaluations.

We next test our hypothesis using OLS regressions. Model 1 in Table 2 shows a positive relationship between *Inferred productivity* and performance scores accounting for year fixed effects but in the absence of any control variables. Table 2, Model 2 includes various controls and adds supervisor fixed effects. Table 2, Model 3 adds worker fixed effects. We evaluate effect sizes

[8]For example, if a worker spent the entire year as an operator, job level would equal 3. If she spent half the year as an operator and half as an assistant operator, job level would equal 2.5.

[9]The company's productivity increased dramatically over time, but its average evaluation score declined modestly over time. This creates the noisy negative correlation between WIP_{wt} and evaluation score in the raw data. As can be seen in Table A4, Model 1, once we include year of production dummies (and no other covariates), the partial correlation between *Inferred productivity* and *Performance evaluation* is large, statistically significant, and positive.

TABLE 2 OLS estimates of relationship between inferred productivity and performance evaluations

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Length of relationship		0.004	−0.009	0.004	0.003	0.003	0.004
		(0.003)	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)
Avg. joint experience		0.072	0.034	0.073	0.066	0.073	0.073
		(0.023)	(0.016)	(0.021)	(0.021)	(0.023)	(0.024)
% Absenteeism		−0.708	−0.594	−0.727	−0.745	−0.780	−0.758
		(0.488)	(0.502)	(0.489)	(0.488)	(0.466)	(0.453)
Female worker		0.193		0.190	0.202	0.191	0.168
		(0.056)		(0.053)	(0.059)	(0.055)	(0.054)
Minority group member		−0.076		−0.071	−0.072	−0.079	−0.078
		(0.053)		(0.051)	(0.055)	(0.053)	(0.058)
% Overtime		1.325	1.130	1.338	1.307	1.276	1.388
		(0.282)	(0.362)	(0.285)	(0.265)	(0.256)	(0.249)
Job level		0.069	−0.039	0.068	0.068	0.066	
		(0.017)	(0.038)	(0.016)	(0.018)	(0.017)	
Time on rotation (%)		0.309	0.080	0.312		0.301	0.276
		(0.097)	(0.086)	(0.095)		(0.096)	(0.093)
Inferred productivity	0.264	0.242	0.246	0.110	0.063	0.037	0.021
	(0.118)	(0.087)	(0.091)	(0.120)	(0.135)	(0.142)	(0.116)
Inferred prod. x rotation (%)				0.769			
				(0.449)			
High rotation (dummy)					0.109		
					(0.045)		
Inferred prod. × high rotation					0.390		
					(0.201)		
Inferred prod. × job level						0.205	
						(0.074)	
High authority (dummy)							0.110
							(0.034)
Inferred prod. × high authority							0.611
							(0.165)
Constant	3.259	3.113	3.341	3.163	3.185	3.209	3.232
	(0.032)	(0.098)	(0.104)	(0.099)	(0.095)	(0.093)	(0.096)
Supervisor fixed effects?	No	Yes	Yes	Yes	Yes	Yes	Yes
Team fixed effects?	No	Yes	Yes	Yes	Yes	Yes	Yes
Worker fixed effects?	No	No	Yes	No	No	No	No
Adj. *R*-squared	.449	.592	.737	.593	.590	.594	.593

Note: $N = 541$ observations. All models include a dummy for the year of evaluation, between 2012 and 2019. Robust standard errors clustered by supervisor in parentheses. Sample includes all workers who have at least 10 full working days and a performance evaluation in year$_t$.

using Table 2, Model 2. The effect size is relatively modest: when WIP_{wt} moves from the sample mean (0.01) to +1 standard deviation (i.e. to 0.16), the performance score increases about 0.03, which is 5.5% of the standard deviation of performance score and 0.9% of the sample mean of performance score. These moderate effect sizes are consistent with our theoretical expectations, as the company does not have an explicit policy that recommends the use of counterfactual variation data to evaluate performance. See Figure A3 for a binned scatterplot of this relationship.

As a robustness test, we replicate the results for H1 using a different dependent variable: peer-voted awards for employee of the month. This dependent variable should not be subject to as many concerns about supervisor favoritism. Moreover, coworkers also have ample opportunity to observe dynamic correlations between team productivity and a worker's presence. Section A2.1 in the Appendix explains this exercise, and Table A4 displays the results, which are consistent with those in Table 2.

It is important to note that the impact of inferred productivity on evaluation scores may be epiphenomenal, in the sense that it captures a worker's underlying ability, not because supervisors actually use variation in team composition to inform their beliefs about workers' individual contributions to the firm. Another way to state this concern is that variation in team composition might allow an econometrician to capture ability that is "observable but not verifiable" to supervisors (Gibbons, 1998, p. 121), but that the supervisor herself does not actually use variation in team composition to (even subconsciously) to evaluate workers. We present two empirical checks that address this alternative explanation.

5.1 | Supervisors' opportunities to observe counterfactual information

If the logic underpinning H1 is correct, we should observe that the positive relationship between an employee's inferred productivity and her performance evaluation score is even stronger when the employee spends more time working with different teams and different teammates. When an employee moves across teams, it provides the supervisor with more counterfactual information to evaluate the focal worker. The supervisor can observe how the performance of the new team changes with the focal employee's presence, and the supervisor can observe how the employee's former team performs in the employee's absence. By contrast, a supervisor will have less counterfactual information when observing workers producing with the same set of coworkers. Thus, the use of counterfactual information by supervisors implies that *Time on rotation (%)* should positively moderate the relationship between a worker's *Inferred productivity* and their evaluation score.

Model 4 in Table 2 adds an interaction term *Inferred productivity* × *rotation (%)*. Model 5 in Table 2 replaces the *Time on rotation (%)* variable with an indicator capturing observations above the median for this variable, which captures workers who spent more than 10% of their time on rotation beyond their home team. Both interaction effects in Models 4 and 5 are positive, with *p*-values of .104 and .068. The effect size is large: Model 5 suggests that the positive relationship between *Inferred productivity* and the evaluation score is about six times larger for workers where *Time on rotation* is above the sample median.

5.2 | Supervisors' attribution of individual contribution to team production

If inferred productivity were simply a proxy for individual ability, it would impact the evaluations of workers regardless of job level. By contrast, if supervisors incorporate counterfactual information to generate subjective evaluations, we need to consider the specific ways in which human judges use hierarchy when making attributions. First, a wide span of literature documents how higher authority agents tend to receive the credit—and the blame—for group performance (Hamilton, 1978), because higher authority agents have more control over how the group conducts its work. Second, the work of lower authority workers tends to be more routinized, so supervisors may not attribute variation in team performance to the presence or absence of a lower authority worker, but rather to the work routine that the lower authority worker is following, the quality of which might be the responsibility of the higher authority worker. Therefore, supervisors might attribute more of the variation in a group's performance to the higher authority members of the group. Thus, the use of counterfactual information by supervisors implies that *Job Level* should positively moderate the relationship between a worker's *Inferred productivity* and their evaluation score.

Model 6 in Table 2 adds an interaction term *Inferred productivity* × *Job level*. The interaction effect is positive ($p = .013$). Model 7 in Table 2 replaces the continuous *Job level* variable with an indicator equal to 1 for high authority workers (i.e. operators and assistant operators) and zero otherwise. The interaction term *Inferred prod.* × *High authority* is also positive with $p = .002$. The effect size is large: Model 7 suggests that the positive relationship between *Inferred productivity* and the evaluation score is many times larger for operators and assistant operators than for other employees.

In Table A5, we re-test the two aforementioned empirical checks by splitting the sample at the median of % Time on Rotation and Low/High Job Authority. Comparison between Models 1 and 2 and between Models 3 and 4 in Table A5 provides further empirical support. The coefficient on *Inferred Productivity* for workers above the median amount of time on rotation (Model 1, p-value = .035) is about twice the size of the coefficient for workers at or below the median (Model 2, p-value = .271). The coefficient on *Inferred Productivity* for high authority workers (Model 3, p-value <.001) is many times larger than the coefficient for low authority workers (Model 4, p-value = .791). Binned scatterplots in Figure 1 illustrate results for workers with High and Low levels of time on rotation and authority level. Note however, that the highly conservative Wald test across the split samples in Table A5 does not support a difference in the coefficient of *Inferred Productivity* between Models 1 and 2 (p-value = .40), but does support a difference between Models 3 and 4 (p-value = .02).

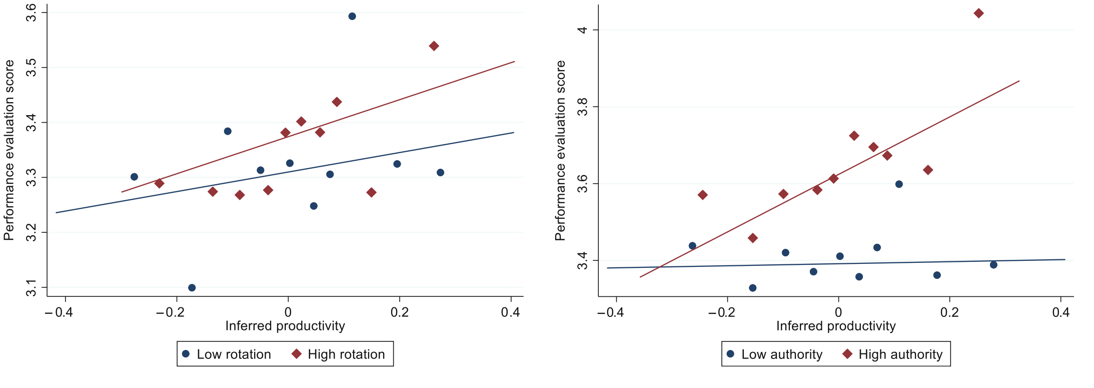

FIGURE 1 Binned scatterplots of relationship between workers' inferred productivity and performance evaluation score, split by time on rotation and job authority. This figure comes from Table A5, left pane compares Models 1 and 2; right pane compares Models 3 and 4

6 | DISCUSSION AND CONCLUSION

Scholars have emphasized possible managerial solutions to performance non-separability (Bandiera et al., 2010; Holmstrom, 1982; McAfee & McMillan, 1991), including team incentive schemes that allow workers to sort into teams based on productivity (e.g., Bandiera, Barankay, & Rasul, 2013). Yet these solutions do not provide guidance to understand the process by which supervisors ultimately arrive at their subjective evaluations of their subordinates' contributions to team production. Despite the prevalence of team production and subjective performance evaluation in modern firms, we have few theories about how managers decide their subordinates' subjective performance scores when output is team-based. We suggest that managers might glean information from fluctuations in team membership to inform their opinions about individual workers' contributions to team production. We model supervisors' inferences using standard techniques for estimating individual productivity from data on collectively generated output (Azoulay et al., 2010; Chen & Garg, 2018; Oettl, 2012; Stuart, 2017). In total, the theory and results suggest that fluctuations in team membership and output may generate useful information for supervisors who must evaluate workers' individual performance.

To our knowledge, this work is the first to argue and test that supervisors may use counterfactual information in their evaluations, especially when evaluating workers who rotate across different teams and who have sufficient authority to direct team production. Our focus on the capacity of supervisors to overcome performance non-separability enriches the extant performance management literature, which has primarily focused on supervisors' biases in the evaluation of their subordinates, including biases from ascriptive characteristics (Castilla, 2011; Sarsons, 2017; Tsui & O'Reilly, 1989) or from supervisors' own experiences receiving evaluations (Castilla & Ranganathan, 2020). We also add to that body of work by documenting how workers' inferred productivity can be captured from a team production context using well-established econometric techniques. Performance appraisal research has largely focused on the individual production context, and our simple analytic approach opens the door for deeper exploration of how subjective evaluations are carried out in a team production context.

Our work also contributes to the growing literature on worker complementarities and interdependence in team production. This work has shown that team level outcomes depend not only on individual human capital, but also on hard-to-assess complementarities embedded in specific assets and routines and interdependencies between team members (e.g., Groysberg et al., 2008; Huckman & Pisano, 2006). In this present work, supervisors evaluate an individual's contribution as a unitary whole, and this contribution encompasses human capital, social capital, and various complementarities. Future research can examine how supervisors parse out the unique complementarities that exist between a specific worker and the different teams in which they operate. A natural question is whether important differences exist in supervisors' ability to infer contributions coming from workers with different levels of complementarity, social capital, and human capital. Addressing this question empirically will require access to disaggregated data about supervisors' assessment of different components of an individual's contribution, as well as random allocation of members into teams.

Finally, we want to highlight the managerial implications of this paper. First, the paper suggests that particular firm policies and structures might help the firm to better assess individual contributions to team production. For example, General Electric famously emphasized two management practices: management rotation programs, where new hires move across several job functions in their first years in the company, and "rank and yank," where supervisors create a forced ranking of subordinates and fire the lowest performers. The theory in this paper suggests that these two management practices may have complementarities: the rotation program may help to create variation, which increases the accuracy of the forced rankings of employees.

Second, the paper suggests that managers might incur influence costs and agency costs from employees who seek to manage perceptions of their inferred productivity. The theory in this paper treats the timing of employee absences and job rotations as exogenous. However, in many situations, employees (especially executives) have some control over the timing of their absences, job rotations, and/or departures from the organization (e.g. McDonnell & Cobb, 2020). If the theory in this paper is correct, then employees and executives have incentives to game these events in order to cast themselves in the best possible light (e.g., Graffin, Carpenter, & Boivie, 2011) and perhaps enhance their opportunities for value capture. For example, the leader of a firm concerned about her legacy might choose to leave the organization when she perceives its performance to be on a downward trajectory, or she might even sabotage the firm on the way out the door, in order to maximize the contrast between organizational performance during her leadership and her successor's leadership. Employees might also try to avoid difficult project assignments in order to avoid tarnishing their reputations if those projects do not succeed.

6.1 | Limitations

We wish to be very clear about the limitations of this paper. First, while our theory treats worker assignments to teams as exogenous, our empirical setting does not allow us to achieve this ideal. While our informants suggest that the firm does not typically adjust production based on worker availability, we cannot rule out the possibility that the results are driven by unobserved variables (such as worker quality, demand shocks, production problems), which might drive productivity and worker assignments into teams. Readers should regard the results as correlational evidence of the theory that we put forward in the paper, and we hope future researchers will be able to identify data where workers are randomly assigned into teams.

Second, we draw on data from one organization. This insider approach has numerous benefits for the internal validity of our work. However, there is a meaningful sacrifice of external validity. While the organization should be broadly representative of team-based, manufacturing-oriented organizations, our results call for replication in other settings, and future work will be required to test the validity of this theory in organizations with different team incentives, structures, and task interdependencies. We think the effect sizes uncovered in this paper may be much larger in other settings, such as those where supervisors have less ability to observe the production process directly and must rely on data regarding team membership and group output more heavily. In other organizations, though, especially those where work is less interdependent, supervisors may be able to infer employees' productivity by simply observing how hard they are working, and information gleaned from variation in team membership may be less important. We hope that our work inspires future research that can shed further light on how supervisors may implicitly infer individual contributions to team output and when the explicit use of such information can benefit workers, teams, and organizations.

ACKNOWLEDGEMENTS

The authors would like to thank Editor Sendil Ethiraj and two anonymous reviewers for their remarkable suggestions and guidance. The authors would also like to thank seminar participants at the University of Minnesota, the University of Michigan, the 2021 Academy of Management conference, Washington University, Purdue University, the Wharton People and Organizations conference, and the Strategy, Innovation, and Entrepreneurship Workshop. The authors are indebted to Victor Bennett, David Kryscynski, Daniel Malter, Oren Reshef, Mike Ryall, Jim Walsh and the Maryland Reading Group for their detailed comments on various versions of this project. The authors would also like to thank the members of the partner company for their time and generosity. All errors are ours.

DATA AVAILABILITY STATEMENT

The underlying data are owned by our partner company. The data contain valuable and sensitive information about productivity and personnel issues. We are not permitted to share the data, but we will share all of the relevant code if the paper is published.

REFERENCES

Acemoglu, D., & Pischke, J. S. (1999). Beyond Becker: Training in imperfect labour markets. *The Economic Journal*, *109*(453), 112–142.

Alchian, A. A., & Demsetz, H. (1972). Production, information costs, and economic organization. *The American Economic Review*, *62*(5), 777–795.

Arvey, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology*, *49*(1), 141–168.

Azoulay, P., Graff Zivin, J. S., & Wang, J. (2010). Superstar extinction. *The Quarterly Journal of Economics*, *125*(2), 549–589.

Baker, G., Gibbons, R., & Murphy, K. J. (1994). Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics*, *109*(4), 1125–1156.

Bandiera, O., Barankay, I., & Rasul, I. (2010). Social incentives in the workplace. *The Review of Economic Studies*, *77*(2), 417–458.

Bandiera, O., Barankay, I., & Rasul, I. (2013). Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association*, *11*(5), 1079–1114.

Bermiss, Y. S., & Murmann, J. P. (2015). Who matters more? The impact of functional background and top executive mobility on firm survival. *Strategic Management Journal*, *36*(11), 1697–1716.

Bertrand, M., & Schoar, A. (2003). Managing with style: The effect of managers on firm policies. *The Quarterly Journal of Economics*, *118*(4), 1169–1208.

Castilla, E. J. (2011). Bringing managers back in: Managerial influences on workplace inequality. *American Sociological Review*, *76*(5), 667–694.

Castilla, E. J., & Ranganathan, A. (2020). The production of merit: How managers understand and apply merit in the workplace. *Organization Science*, *31*(4), 909–935.

Chen, J. S., & Garg, P. (2018). Dancing with the stars: Benefits of a star employee's temporary absence for organizational performance. *Strategic Management Journal*, *39*(5), 1239–1267.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Coff, R. W. (1997). Human assets and management dilemmas: Coping with hazards on the road to resource-based theory. *Academy of Management Review*, *22*(2), 374–402.

Cooper, R., & Gubler, T. (2020). Combining Individual and Firm-level Human Capital Resources: The Creation and Performance Benefits of Human Capital Alignment. Available at *SSRN 3800449*.

DeNisi, A., & Smith, C. E. (2014). Performance appraisal, performance management, and firm-level performance: A review, a proposed model, and new directions for future research. *Academy of Management Annals*, *8*(1), 127–179.

Di Stefano G., Grohsjean, T., & Gutierrez, C. (2020) The genesis of team familiarity: An experimental investigation. *Working paper*, Bocconi University

Elvira, M., & Town, R. (2001). The effects of race and worker productivity on performance evaluations. *Industrial Relations: A Journal of Economy and Society*, *40*(4), 571–590.

Ethiraj, S. K., & Garg, P. (2012). The division of gains from complementarities in human-capital-intensive activity. *Organization Science*, *23*(3), 725–742.

Frederiksen, A., Kahn, L. B., & Lange, F. (2020). Supervisors and performance management systems. *Journal of Political Economy*, *128*(6), 2123–2187.

Gartenberg, C., & Wulf, J. (2017). Pay harmony? Social comparison and performance compensation in multi-business firms. *Organization Science*, *28*(1), 39–55.

Gibbons, R. (1998). Incentives in organizations. *Journal of Economic Perspectives.*, *12*(4), 115–132.

Graffin, S. D., Carpenter, M. A., & Boivie, S. (2011). What's all that (strategic) noise? Anticipatory impression management in CEO succession. *Strategic Management Journal*, *32*(7), 748–770.

Groysberg, B., Lee, L. E., & Nanda, A. (2008). Can they take it with them? The portability of star knowledge workers' performance. *Management Science*, *54*(7), 1213–1230.

Hamilton, B. H., Nickerson, J. A., & Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy*, *111*(3), 465–497.

Hamilton, V. L. (1978). Who is responsible? Toward a social psychology of responsibility attribution. *Social Psychology*, *41*, 316–328.

Holmstrom, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, 13, 324–340.

Huckman, R. S., & Pisano, G. P. (2006). The firm specificity of individual performance: Evidence from cardiac surgery. *Management Science*, *52*(4), 473–488.

Hume, D. (1739). *A treatise on human nature*. Oxford, UK: Clarendon Press, Oxford.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136.

Kehoe, R. R., & Bentley, F. S. (2019). Building human capital resources: Hiring and acquiring. In *Handbook of research on strategic human capital resources*. Massachusetts, USA: Edward Elgar Publishing.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107.

Kryscynski, D., Coff, R., & Campbell, B. (2021). Charting a path between firm-specific incentives and human capital-based competitive advantage. *Strategic Management Journal*, *42*(2), 386–412.

Lazear, E. P., & Rosen, S. (1981). Rank-ordertournaments as optimum labor contracts. *Journal of Political Economy*, *89*(5), 841–864.

Larkin, I., Pierce, L., & Gino, F. (2012). The psychological costs of pay-for-performance: Implications for the strategic compensation of employees. *Strategic Management Journal*, *33*(10), 1194–1214.

Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management*, *30*(6), 881–905.

McAfee, R. P., & McMillan, J. (1991). Optimal contracts for teams. *International Economic Review*, *32*, 561–577.

McDonnell, M. H., & Cobb, J. A. (2020). Take a stand or keep your seat: Board turnover after social movement boycotts. *Academy of Management Journal*, *63*(4), 1028–1053.

Oettl, A. (2012). Reconceptualizing stars: Scientist helpfulness and peer performance. *Management Science*, *58*(6), 1122–1140.

Pierce, L., Wang, L. W., & Zhang, D. J. (2020). Peer bargaining and productivity in teams: Gender and the inequitable division of pay. Manufacturing & Service Operations Management.

Ployhart, R. E., & Moliterno, T. P. (2011). Emergence ofthe human capital resource: A multilevel model. *Academy of Management Review*, *36*(1), 127–150.

Raffiee, J., & Byun, H. (2020). Revisiting the portability of performance paradox: Employee mobility and the utilization of human and social capital resources. *Academy of Management Journal*, *63*(1), 34–63.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, *121*(1), 133.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology*, *10*, 173–220.

Rumelt, R. P. (1984). Towards a strategic theory of thefirm. *Competitive Strategic Management*, *26*(3), 556–570.

Rust, J. (1987). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica: Journal of the Econometric Society*, *55*, 999–1033.

Sarsons, H. (2017). Recognition for group work: Gender differences in academia. *American Economic Review*, *107*(5), 141–145.

Sevcenko, V., & Ethiraj, S. (2018). How do firms appropriate value from employees with transferable skills? A study of the appropriation puzzle in actively managed mutual funds. *Organization Science*, *29*(5), 775–795.
Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge, England: Cambridge University Press.
Somers, M. J. (1995). Organizational commitment, turnoverand absenteeism: An examination of direct and interaction effects. *Journal of Organizational Behavior*, *16*(1), 49–58.
Stuart, H. C. (2017). Structural disruption, relational experimentation, and performance in professional hockey teams: A network perspective on member change. *Organization Science*, *28*(2), 283–300.
Sundvik, L., & Lindeman, M. (1998). Performance rating accuracy: Convergence between supervisor assessment and sales productivity. *International Journal of Selection and Assessment*, *6*(1), 9–15.
Tsui, A. S., & O'reilly III, C. A. (1989). Beyond simpledemographic effects: The importance of relational demography insuperior-subordinate dyads. *Academy of Management Journal*, *32*(2), 402–423.
Uttal, B. (1985) Behind the fall of Steve jobs. Fortune. Retrieved from https://archive.fortune.com/magazines/fortune/fortune_archive/1985/08/05/66254/index.htm.
Van de Ven, A. H., Delbecq, A. L., & Koenig, R., Jr. (1976). Determinants of coordination modes within organizations. *American Sociological Review*, *41*, 322–338.
Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *19*(3), 231–241.
Weinberger, M., & Hartmans, A. (2020). How Steve Jobs saved Apple: His career from 1996 to 2011. Business Insider. Retrieved from https://www.businessinsider.com/steve-jobs-apple-photos-2017-1.
Zenger, T. R. (1992). Why do employers only reward extreme performance? Examining the relationships among performance, pay, and turnover. *Administrative Science Quarterly*, *37*, 198–219.
Zenger, T. R. (1994). Explaining organizational diseconomies of scale in R&D: Agency problems and the allocation of engineering talent, ideas, and effort by firm size. *Management Science*, *40*(6), 708–729.
Zenger, T. R., & Hesterly, W. S. (1997). The disaggregation of corporations: Selective intervention, high-powered incentives, and molecular units. *Organization Science*, *8*(3), 209–222.

How to cite this article: 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 Journal*, *43*(12), 2577–2601. https://doi.org/10.1002/smj.3433

APPENDIX A

A.1 | DETAILS ABOUT OUR CALCULATION OF WORKER × YEAR FIXED EFFECTS

Recall from the main paper that we estimate annual worker fixed effects as follows:

$$Y_{(ij)kdw} = \beta_0 + WIP_{wt} + \eta_i + \delta_j + \omega_k + \gamma_{(ij)k} + \phi_{\text{month}} + \varepsilon_{(ij)kdp}, \tag{A1}$$

Y is the log of the number of good units of item k produced on date d by the team working on line i on shift j and which included worker w. β_0 is an overall intercept.

WIP_{wt} is the *Worker's Inferred Productivity*, our quantity of interest; it is a worker $w \times$ year t fixed effect that we will correlate with the worker's performance evaluation score. Given that each production line has different machinery, the model includes fixed effects for each of the six specific lines (η_i). We include three shift-level fixed effects (δ_j) because the most productive workers generally prefer and get assigned to the first work shift (7 a.m.–3 p.m.), and less

TABLE A1 Representative data from interviews and field work

Source	Comment/insight	Use in theory/empirics
Interview with Supervisor # 1	When there is an absence, getting back up to speed will depend on who you can get to replace the absent worker	Supervisors' may use counterfactual information from team membership fluctuations; supports *Inferred productivity* as metric of individual contributions to observable team output
Interview with Manager #2	Evaluations are tied to workers' end of year bonus; supervisors communicate rationale for their evaluation directly to worker	
Interview with HR Director	Supervisors are incentivized to recognize high potential employees for internal promotion	
Electronic communication with Supervisor # 2	Key labor challenge is finding and retaining operators that can lead by example and help the team	Supervisors' attribution of impact on team performance will depend on worker's authority; supports *Job level* as moderator for the impact of *Inferred productivity* on evaluations
Interview with Operator #1	Operators train other team members on machine capabilities and set the standard for the team's work ethic	
Interview with Operator #5	Supervisors oversee the staffing of multiple teams across a shift	Variation in supervisors' opportunities to observe counterfactual information; supports *Rotations beyond home team* as moderator for the impact of *Inferred productivity* on evaluations
Interview with Manager #1	High customer demand and product complexity requires rotating workers across teams to plug labor shortages	

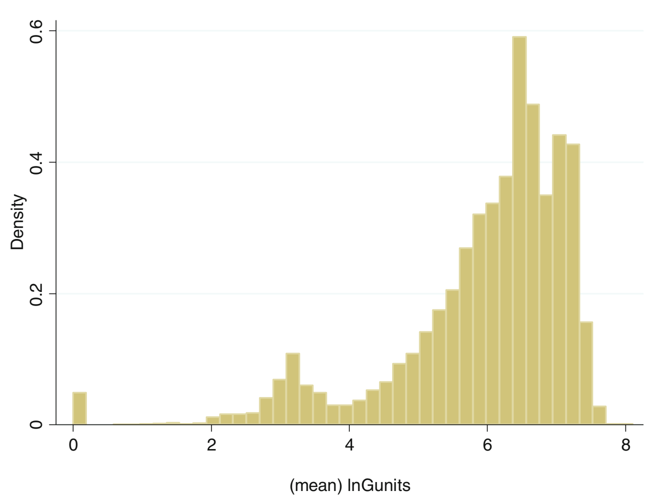

FIGURE A1 Distribution of good units. The unit of analysis in Figure 1 is a production event, defined by a production date (n = 1,588), a shift × line production unit (n = 14), and an item type (n = 182)

productive workers get assigned to the later shifts (3 p.m.–11 p.m. and 11 p.m.–7 a.m.). We include 182 item-type fixed effects (ω_k) to control for heterogeneity in the difficulty of manufacturing the different items. We also included $\phi_{\text{month}\times\text{year}}$, which are time dummies for

TABLE A2 Summary statistics and correlations for production data (N = 121,842)

		Mean	SD	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1.	Good units (log)	5.70	1.83								
2.	No. of hours on shift	7.62	1.78	.04							
3.	Planned production time	1.75	0.70	.82	.04						
4.	No. of shifts on item	0.96	0.79	.43	−.01	.37					
5.	Includes changeover	0.29	0.45	−.51	.01	−.49	−.78				
6.	Includes maintenance	0.04	0.20	−.02	−.04	−.07	.07	−.06			
7.	Avg. coworker tenure (log)	1.81	0.65	−.08	−.08	−.04	−.06	.06	−.07		
8.	Avg. coworker job level	1.94	0.30	.00	.01	−.01	.00	−.01	.02	.07	
9.	No. of coworkers (log)	1.78	0.36	.06	−.11	−.03	.07	−.02	.06	−.13	−.09

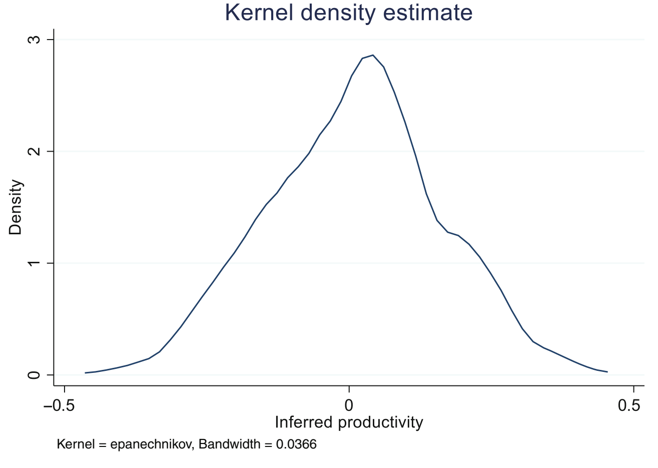

FIGURE A2 Kernel-density distribution of Workers' inferred productivity (N = 541)

each month-year pair to control for secular trends and idiosyncratic company events in different months. $\gamma_{(ij)kd}$ contains covariates listed in the next section.

A.1.1. | Control variables for the productivity model

Though supervisors do not randomly assign workers to production events, we can use our extensive interviews with the members of the planning and production departments to include a useful set of control variables in $\gamma_{(ij)kd}$. These controls should correlate with the workers' underlying quality and the expected productivity of the production event, allowing us to more accurately recover worker's incremental contributions to their team's production.

TABLE A3 Correlates of daily productivity ($N = 121{,}748$)

	Model 1
Planned production time (log)	2.027
	(0.014)
No. of shifts on item (log)	0.190
	(0.008)
Includes changeover	−0.263
	(0.022)
Includes maintenance	0.083
	(0.028)
No. of hours on shift	0.006
	(0.002)
No. coworkers present (log)	0.032
	(0.012)
Avg. coworker tenure (log)	0.032
	(0.012)
Avg. coworker job title	−0.057
	(0.014)
Constant	3.265
	(0.072)
Adj. *R*-squared	.728

Note: This is the estimation of Equation 2, used to extract *Inferred productivity* (i.e., the fixed effect for worker-year). The model contains fixed effects for item, line, shift and year-month. Robust standard errors clustered by worker in parentheses.

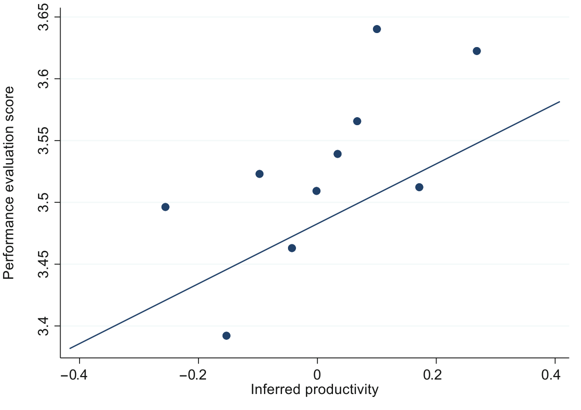

FIGURE A3 Binned scatterplot of relationship between workers' inferred productivity and performance evaluation score (H1). This figure comes from Table 2, Model 2

A.1.1.1. | Includes changeover

This is an indicator variable equal to one (zero otherwise) when in the focal production event machinery had to be reconfigured from the previous item in order to manufacture a different

TABLE A4 Logit estimates of the relationship between inferred productivity and peer award nominations (H1 robustness)

	Model 1	Model 2	Model 3[a]
Length of relationship		0.049	−0.081
		(0.074)	(0.105)
Avg. joint experience		0.222	−0.005
		(0.123)	(0.269)
% Absenteeism		−7.872	−13.779
		(5.964)	(8.549)
Female worker		0.121	
		(0.211)	
Minority group member		−0.057	
		(0.208)	
% Overtime		1.960	2.489
		(3.900)	(5.060)
Job level		0.335	0.230
		(0.080)	(0.760)
Time on rotation (%)		0.293	−1.174
		(0.320)	(1.172)
Inferred productivity	3.909	2.947	5.141
	(1.325)	(2.005)	(1.999)
Constant	−3.987	−6.229	
	(0.867)	(2.400)	
Supervisor fixed effects?	No	Yes	Yes
Team fixed effects?	No	Yes	Yes
Worker fixed effects?	No	No	Yes
Pseudo *R*-squared	.068	.202	.342
Observations	541	485	299
Log likelihood	−213.0	−173.3	−71.8

Note: All models include a dummy for the year of evaluation, between 2012 and 2019. Robust standard errors clustered by supervisor in parentheses. Sample includes all workers eligible for a peer nomination in year$_t$.

[a]The results for Model 3 use a conditional logit estimation. We also obtain a substantively similar result to *Inferred productivity* in Model 3 when using a Linear Probability Model. These results are available from the authors.

item. Machine changeovers often lead to unexpected delays when recalibrating machinery, which should negatively impact output. Higher quality workers might be assigned to these more difficult shifts.

A.1.1.2. | Planned production time (log)

This variable calculates the natural log of the number of hours that the company's planning department allotted for a specific production event. When allotting production time to a

TABLE A5 Split sample regressions of low/high time on rotation (Models 1 and 2) and low/high authority level (Models 3 and 4)

	Model 1	Model 2	Model 3	Model 4
Length of relationship	0.004	0.004	0.003	0.006
	(0.003)	(0.004)	(0.004)	(0.005)
Avg. joint experience	0.050	0.069	0.134	0.048
	(0.055)	(0.025)	(0.043)	(0.019)
% Absenteeism	−0.765	−0.602	−0.322	−1.278
	(0.592)	(0.535)	(0.641)	(0.481)
Female worker	0.164	0.234	0.641	0.121
	(0.057)	(0.085)	(0.197)	(0.044)
Racial minority	−0.023	−0.149	−0.226	−0.091
	(0.045)	(0.079)	(0.099)	(0.067)
% Overtime	1.921	0.808	2.288	0.504
	(0.418)	(0.579)	(0.469)	(0.246)
Job level	0.025	0.101		
	(0.041)	(0.026)		
Time on rotation (%)			0.019	0.33
			(0.171)	(0.113)
Inferred productivity	0.337	0.178	0.748	0.027
	(0.147)	(0.156)	(0.120)	(0.098)
Constant	3.201	3.004	3.242	3.275
	(0.097)	(0.230)	(0.068)	(0.107)
Sample?	High rotation	Low rotation	High authority	Low authority
Observations	270	271	201	340
Adj. *R*-squared	.604	.603	.591	.632

Note: All models include a dummy for the year of evaluation (between 2012 and 2019), the evaluating supervisor and the worker's main team. Robust standard errors clustered by supervisor in parentheses. Sample includes all workers who have a performance evaluation in year$_t$ and who have at least 10 working days in the payroll in year$_t$.

production event, the company's planning department deploys an algorithm that uses historical data to estimate the amount of time needed to produce a particular product on a particular machine, adjusting for any required maintenance or changeovers. Higher quality workers might be assigned to these longer production events, which will also have higher output.

A.1.1.3. | No. of shifts on item (log)

This variable calculates the log of the number of production shifts since that manufacturing line transitioned from one product to the current product. When this number is higher, productivity should be higher, because the line has been engaged in the production of the product for a longer period of time, allowing the line to become well-calibrated to the production requirements of the focal product. Lower quality workers might be assigned to these easier production events.

A.1.1.4. | No. of hours on shift

This variable is the exact number of hours, from 1 to 8, that a worker spent on a particular shift. The vast majority of workers spent the entire (8 h) shift with their team, but in some cases such as scheduling conflicts or training sessions, some workers would spend fewer hours on the production floor on particular days. This variable accounts for this source of heterogeneity.

A.1.1.5. | Includes maintenance

This is a dummy variable that indicates that the planning department has scheduled maintenance time for the machines during a production event. Higher quality workers might be assigned to these more difficult production events, which will also have lower output.

A.1.1.6. | No. coworkers present

This variable captures the number of coworkers present for the production event. Number of coworkers should positively affect productivity. Higher quality workers might be asked to work with smaller teams.

A.1.1.7. | Avg. coworker job title

This variable measures job titles among coworkers present for the production event, where 0 = crew, 1 = helpers, 2 = assistant operators, and 3 = operators. A team with greater average job title has greater average technical knowledge and authority and should be more productive.

A.1.1.8. | Avg. coworker tenure (log)

The average number of years from a production date since all co-workers in a production unit were hired. This variable proxies for coworkers' experience; less experienced coworkers should mean the production event is less successful. Higher quality workers might be given less experienced teammates, in order to compensate for their inexperience.

Table A2 shows descriptive statistics and bivariate correlations between all control variables in the production data. Table A3 contains estimates of the model from Equation A1.

A.2 | ROBUSTNESS TESTS

A.2.1. | Peer awards

As we noted above, the assignment of workers into teams is not random. Supervisors may place favored workers in teams that are more productive while also giving them inflated evaluation scores. We tried to account for favoritism by including numerous control variables. To push further, we introduce a second dependent variable into our analysis. Management routinely recognizes a small number of "employees of the month," informal awards given to employees with the largest number of nominations by coworkers. Note that coworkers will typically have as much, if not more, fine-grained knowledge as supervisors regarding the covariation between the focal worker's presence and her team's output. Crucially, nominating peers cannot directly influence the focal worker's team assignments, so the presence of favoritism should be less of an issue. If *Inferred productivity* (WIP_{wt}) correlates with a worker's nomination for this award, it would provide more evidence that this metric is not simply an endogenous process of supervisor favoritism toward the focal worker. The binary variable *Received peer nomination* has a relatively low bivariate correlation ($\rho = .21$) with supervisor-assigned performance evaluations, our main dependent variable, enhancing its value as a robustness test.

We repeated Models 1–3 from Table 2 using peer nominations instead of supervisors' performance evaluations as the dependent variable. We used logistic regression instead of OLS to accommodate the functional form of the binary dependent variable *Received peer nomination*, which takes a value of 1 for employees who received such a nomination in the focal year, and zero otherwise. Table A4 displays the results of the three models. The estimate of *Inferred productivity* in Model 2, which has all the controls but without individual worker fixed effects, is not distinguishable from zero but its directionality is in line with our theory. However, the estimates of *Inferred productivity* in Models 1 and 3 provide strong support for H1.