American Economic Review 2023, 113(3): 766–799
https://doi.org/10.1257/aer.20211207

Subjective Performance Evaluation, Influence Activities, and Bureaucratic Work Behavior: Evidence from China[†]

By Alain de Janvry, Guojun He, Elisabeth Sadoulet, Shaoda Wang, and Qiong Zhang*

Subjective performance evaluation could induce influence activities: employees might devote too much effort to pleasing their evaluator, relative to working toward the goals of the organization itself. We conduct a randomized field experiment among Chinese local civil servants to study the existence and implications of influence activities. We find that civil servants do engage in evaluator-specific influence to affect evaluation outcomes, partly in the form of reallocating work efforts toward job tasks that are more important and observable to the evaluator. Importantly, we show that introducing uncertainty about the evaluator's identity discourages evaluator-specific influence activities and improves bureaucratic work performance. (*JEL* D73, H83, J45, M54, O17, O18, P25)

For a large share of jobs in modern economies, objective performance measures are difficult to obtain, leading employers to rely heavily on supervisors' subjective evaluations to provide work incentives (Prendergast 1999; Deb, Li, and Mukherjee 2016). This is particularly ubiquitous in the public sector, due to the inherent problems of measuring individual achievements and the multiplicity of tasks for most civil service jobs (Olken and Pande 2013; Finan, Olken, and Pande 2015).

While subjective performance measures might improve contractual power (Gibbons and Murphy 1992; Baker, Gibbons, and Murphy 1994), they also open the door to influence activities: employees can take actions to affect the evaluator's

*de Janvry: University of California, Berkeley (email: alain@berkeley.edu); He: University of Hong Kong (email: gjhe@hku.hk); Sadoulet: University of California, Berkeley (email: esadoulet@berkeley.edu); Wang: University of Chicago and NBER (email: shaoda@uchicago.edu); Zhang: Renmin University of China (email: zhangqiong8@ruc.edu.cn). Rema Hanna was the coeditor for this article. We thank the three anonymous referees, Ricardo Alonso, Oriana Bandiera, Alan Benson, Ernesto Dal Bo, Ethan Bueno de Mesquita, Miguel Espinosa, Ray Fisman, Fred Finan, Bob Gibbons, Marco Gonzalez-Navarro, Chang-Tai Hsieh, Ruixue Jia, Jin Li, Weijia Li, Ethan Ligon, Zhaotian Luo, Jeremy Magruder, Aprajit Mahajan, Leslie Martin, Dan Mattingly, Margaret Meyer, Gerard Padro, Albert Park, Nancy Qian, Yingyi Qian, Gerard Roland, Michael Song, Yongxiang Wang, Yuhua Wang, Yanhui Wu, Yang Xie, Guo Xu, David Yang, Noam Yuchtman, and various seminar and conference participants for helpful comments and suggestions. Generous financial support from the J-PAL Governance Initiative and the National Natural Science Foundation of China (grant 72003188) is gratefully acknowledged. Wenwei Peng, Yuhang Pan, Haoyang Xie, Anran Tan, Ziyi Liu, Jie Li, Danfeng Cao, Weiting Miao, Yuliang Nie, Guizi Zhouwu, and a large group of survey enumerators provided outstanding research assistance. The research described in this article was approved by UC Berkeley IRB (Protocol 2017-07-10117). The experiment was registered in the American Economic Association Registry for randomized control trials under trial AEARCTR-0002621.

[†]Go to https://doi.org/10.1257/aer.20211207 to visit the article page for additional materials and author disclosure statements.

assessment in their favor, which could potentially be detrimental to the interests of the organization (Milgrom and Roberts 1988; Milgrom 1988). Specifically, as noted by Milgrom and Roberts (1988), influence activities can be categorized into two types: productive activities, such as putting extra effort into tasks that are more visible to the evaluator, and nonproductive activities, such as "buttering up" the evaluator with personal favors. While a rich theoretical literature has investigated the formation and consequences of influence activities, these theoretical arguments have rarely been confronted with rigorous empirical analysis (Oyer and Schaefer 2011; Lazear and Oyer 2012).

Empirically studying influence activities is challenging for at least three reasons. First, spending extra effort on tasks that are more visible to the evaluator, or trying to personally benefit the evaluator, can be regarded unfavorably by others, which means that the agent might try to hide such behaviors. Second, even if such behaviors are observed, it is difficult to infer that they are driven by intentions of improving evaluation outcomes (rather than simply being hardworking or friendly), making it difficult to classify them exclusively as influence activities. Third, even if the existence of influence activities is established, quantifying their effects on work performance still requires exogenous variation in such behavior across agents.

In this paper, we conduct a large-scale field experiment in two Chinese provinces, which aims at addressing these three challenges and providing empirical evidence on the existence and consequences of influence activities in the workplace. Our experiment focuses on China's "3+1 Supports" program, a large national "human capital reallocation" initiative that hires more than 30,000 college graduates annually to work as entry-level state employees in rural townships on two-year contracts. These individuals are referred to in this paper as College Graduate Civil Servants (CGCSs).

A distinct institutional feature of the Chinese governance system is its dual-leadership arrangement (Shirk 1993), whereby every government organization/subsidiary has two leaders: a "party leader" (i.e., party secretaries at various levels) and an "administrative leader" (i.e., the head in a village, the mayor in a city).[1] As a result of this dual system, every CGCS reports to two supervisors who both assign her job tasks and provide performance feedback on a regular basis. Under the status quo, every CGCS is evaluated by one of her two supervisors every year.[2] The evaluation outcome will determine whether the CGCS can be awarded a permanent contract upon completing her two-year term, a highly sought-after outcome for most CGCSs due to the prestige of permanent civil service jobs in China. Under the current arrangement, rich anecdotal evidence suggests that many CGCSs exert substantial efforts trying to please their specific evaluating supervisor, in both productive and nonproductive ways, in the expectation of better evaluation outcomes. The prevalence of influence activities is concerning to the government, because such efforts to please specific evaluators might crowd out efforts on productive tasks that are deemed more important by the organization.

[1] The two leaders have large overlaps in their responsibilities, introducing de facto checks and balances in employee supervision. See Li (2019) for information on the institutional details of the dual system.

[2] Sixty-two percent of the CGCSs are female, so we use the female pronouns (she/her/hers) for the CGCS throughout this paper. In contrast, the majority of the CGCSs' supervisors and colleagues are male, so we use the male pronouns (he/him/his) when referring to them throughout this paper.

To examine the existence of influence activities in this environment, and to understand the impacts of such activities on CGCS work behavior, we collaborated with two provincial governments in China and randomized two performance evaluation schemes among their 3,785 CGCSs working in 788 townships. In both schemes, we randomly selected one of the two supervisors to be the evaluator. The only difference is that, in the "revealed" scheme, we announced the identity of the evaluator to the CGCS at the beginning of the evaluation cycle, so that, throughout the year, the CGCS knew whose opinion would influence her promotion case. In the "masked" scheme, we kept the identity of the evaluator secret until the end of the evaluation cycle, so that, throughout the year, the CGCS perceived each supervisor as having a 50 percent chance of influencing her promotion. We did not inform the supervisors about who was the chosen evaluator in either scheme.

We find that, in the revealed scheme, the evaluating supervisor gave significantly more positive assessments of CGCS performance than his nonevaluating counterpart. This result is consistent with a scenario where the agent engages in evaluator-specific influence activities—either productive or nonproductive – to improve evaluation outcomes.[3] In comparison, we find no such asymmetry in supervisor assessments in the masked scheme. Exploiting the random assignment of the two evaluation schemes, we find that masking the evaluator's identity incentivizes the CGCSs to reallocate their efforts from evaluator-specific influence activities to productive tasks that are valued by both supervisors, which can significantly improve CGCS work achievements according to a series of performance indicators.

Following the classification of Milgrom and Roberts (1988), we attempt to better understand the nature of potential influence activities in our setting. Regarding productive influence activities, we find that, under the revealed scheme, the CGCS devotes more efforts to the job tasks assigned by her evaluator, and deems the assignments from the evaluator as more important; in addition, her work performance improves more in areas that are valued more highly by the evaluator. Further analysis suggests that these patterns are driven by the behavior of the CGCS, rather than the behavior of the evaluator. We interpret these findings as indicating the existence of productive influence activities in this environment. As for nonproductive influence activities, we document suggestive empirical patterns that are consistent with such behaviors. However, because we cannot directly observe and measure nonproductive influence activities, we discuss these patterns with caution, and do not take a strong stance on the prevalence of such activities in our context.

We conduct a battery of additional tests to rule out alternative interpretations of our findings. For example, we find that the assessment asymmetry under the revealed scheme is not driven by the evaluator potentially finding out about his role and thus changing his behavior, nor by any additional information about CGCS performance being presented to him. We also find that the improved performance under the masked scheme cannot be explained by the CGCS engaging in even more influence activities, either productive or nonproductive, directed toward her supervisors and colleagues.

[3] Analysis of administrative data suggests that these evaluation outcomes indeed have significant influence on the CGCSs' subsequent promotions to permanent civil service positions.

This paper speaks to three strands of literature. First and foremost, it provides the first rigorous empirical test for the existence and implications of influence activities in the workplace. As pointed out by Lazear and Oyer (2012), while a large theoretical literature has studied how agents try to engage in influence activities in the workplace (e.g., Milgrom and Roberts 1988; Milgrom 1988; Meyer, Milgrom, and Roberts 1992; Schaefer 1998, Alonso, Dessein, and Matouschek. 2008; Powell 2015), there is a lack of rigorous empirical evidence, going beyond anecdotes and case studies, to verify these arguments.[4] Our paper fills this gap by providing field experimental evidence, as well as quantifying the causal impact of reducing influence activities on job performance.[5] More broadly, while subjective performance evaluation has been investigated extensively by a large body of theoretical work (Gibbons and Murphy 1992; Baker, Gibbons, and Murphy 1994; Prendergast and Topel 1996; MacLeod 2003, Maestri 2012; Deb, Li, and Mukherjee 2016), empirical evidence on the effectiveness and limitations of subjective evaluation is still largely missing, with only a handful of exceptions (Chevalier and Ellison 1999; Hayes and Schaefer 2000). Our paper contributes to this literature by showing how influence activities can undermine the effectiveness of subjective performance evaluations.[6]

Second, this paper adds to a growing experimental literature on the personnel economics of the developing state, specifically on incentivizing public employees (Finan, Olken, and Pande 2015). Most of the existing field experiments on this topic focus on the role of financial incentives,[7] with only a few exceptions studying nonpecuniary incentives, such as transfers and postings (Banerjee et al. 2012), social incentives (Ashraf and Bandiera 2018), and intrinsic motivation (Ashraf, Bandiera, and Jack 2014). Our paper adds to this line of work by exogenously varying the (implicit) career incentive involved in performance evaluations; this is a prevalent form of motivation in the public sector due to an often compressed wage structure, but has rarely been studied in the literature until recently (Deserranno and Ciliotta 2021).[8] In addition, we show that, holding the career reward fixed, a slight refinement of the performance evaluation practice can lead to a substantial improvement in bureaucratic performance, indicating a highly cost-effective way to enhance state effectiveness.

[4] Rasul and Rogger (2016) find a negative correlation between incentives/monitoring practices and public project completion in Nigeria, which is stronger for more experienced bureaucrats. This empirical pattern is consistent with bureaucrats learning to engage in influence activities over time. Our paper complements Rasul and Rogger (2016) by experimentally altering the bureaucrats' incentives to engage in influence activities, which allows us to causally evaluate the existence and consequences of these activities in the public sector.

[5] A related paper is Wu (2017), which shows that, in a newspaper context, when both mid-level editors and top editors make editorial decisions, the bottom-level reporters have improved work performance. Our paper complements Wu (2017) by not only randomizing the authority for evaluation between two supervisors at the same level, but also cross-randomizing the employee's knowledge of the randomized evaluator's identity. This design allows us to better understand the underlying mechanisms through which the allocation of authority within an organizations affects work performance, and to highlight the role of influence activities.

[6] Our intervention of "masking evaluator identity to discourage influence activities" also relates to a large theoretical literature on using "strategic opacity" and "ex ante randomization" of incentive parameters to combat moral hazard issues (Gjesdal 1982; Stiglitz 1982; Grossman and Hart 1983; Lazear 2006; Jehiel 2015; and Ederer, Holden, and Meyer 2018).

[7] See Finan, Olken, and Pande (2015) for a summary.

[8] Previous research has focused on the selection effect of career incentives; see Ashraf, Bandiera, and Jack (2020), for example. Our paper complements this line of work by investigating the "intensive margin" impact of career incentives, while holding selection fixed.

Third, our paper relates to the research agenda on Chinese political meritocracy. Since Li and Zhou (2005), a large number of empirical studies have tried to investigate how the design of various performance indicators, such as fiscal revenue (Lü and Landry 2014), environmental standards (He, Wang, and Zhang 2020), policy experimentation (Wang and Yang 2022), and population control (Serrato et al. 2019), can affect the behaviors of provincial and prefectural leaders in China. However, existing evidence has focused almost exclusively on high-level government officials, leaving incentives and constraints for the vast majority of local bureaucrats under-researched, even though they could differ substantially from those for high-level leaders.[9] Our paper sheds light on incentive schemes for grassroots bureaucrats in China, who are the building blocks of state capacity and play key roles in public service delivery. More broadly, this paper adds to an emerging literature on bureaucratic performance in developing countries (He and Wang 2017; Bertrand et al. 2020; Martinez-Bravo et al. 2020).

The remainder of this paper is organized as follows. In Section I, we introduce the institutional background, design, and implementation of our field experiment. In Section II, we lay out a simple conceptual framework to help rationalize the empirical setting and experimental design. In Section III, we present the empirical results. In Section IV, we discuss potential alternative interpretations of our findings. Section V concludes.

I. Background and the Experiment

A. *Institutional Background*

Since the early 2000s, the Chinese government has launched several large-scale public employee assignment programs, which have hired more than one million college graduates to work with local governments in rural areas, in the hope that their human capital and independence from local interest groups could improve state effectiveness at the grassroots level. For example, in the College Graduate Village Officials (CGVO) program, new college graduates were hired as village officials on a contractual basis, and the arrival of CGVOs in rural villages has been shown to improve policy implementation and reduce leakages in poverty subsidy distribution (He and Wang 2017).

In this paper, we focus on the "3+1 Supports" initiative—a human capital building program for local governments launched in 2006 by the Ministry of Human Resources and Social Security.[10] Through this program, college graduates are hired to work as temporary civil servants in rural townships. They assume four types of positions: township government clerks focusing on poverty alleviation, township government clerks focusing on agricultural support, teachers in

[9] For instance, a key distinction is that job tasks for low-level bureaucrats are much more difficult to quantify with objective measures such as GDP growth and environmental quality. As a result, most grass-root bureaucrats are rewarded based only on subjective evaluations by their supervisors.

[10] In Chinese, the initiative corresponds to the "*San Zhi Yi Fu* (三支一扶)" program. Six other ministries and departments cosponsored the program, including the Ministry of Education, the Ministry of Finance, the Ministry of Agriculture, the National Health Commission, the State Council Leading Group Office of Poverty Alleviation and Development, and the Communist Youth League Central Committee.

township primary schools, and nurses in township clinics. By the end of 2018, more than 350,000 college graduates had been hired as "College Graduate Civil Servants" (CGCSs) through this program.

The CGCSs are recruited nationwide on a yearly basis. In May, before the end of the school year, each provincial government announces vacancies on its website and invites college graduates to apply. In most provinces, the procedure for CGCS recruitment is similar to that for recruiting regular state employees. Applicants first take a comprehensive written exam, which is similar to the Administrative Aptitude and Essay Writing Tests on the National Civil Service Exam. High-scoring applicants are then interviewed, and top-ranked candidates (based on combined scores) are recruited. Some provinces forgo tests and interviews, and screen applicants simply based on their application materials.

The selection of CGCSs is highly competitive. In most provinces, the acceptance rate for the "3+1 Supports" program is consistently below 10 percent. For example, Shandong province had around 1,500 positions in 2017 and attracted over 31,000 applicants (acceptance rate $<$ 5 percent); in Guangxi province, the government planned to hire 800 CGCSs in 2017 and the total number of applicants exceeded 13,600 (acceptance rate $<$ 6 percent). Such intense competition ensures the high quality of selected CGCSs.

The job tasks of a CGCS are similar to those of a regular entry-level township civil servant. Specifically, for CGCSs in clerical positions—as in the case of regular rural civil servants—job tasks tend to be a combination of routine paperwork, visits to villages, interactions with villagers, meeting attendance, and other case-based assignments from supervisors. Sometimes they are also responsible for policy propaganda, policy enforcement, and identifying and screening beneficiaries for various social assistance programs.

For CGCSs in more specialized positions, such as township clinic nurses or primary school teachers, job tasks are also similar to those of their colleagues who are formal public employees. CGCS teachers work in township public schools; they typically teach multiple courses, help with administrative work, and assist the regular teachers in various ad hoc tasks. CGCS nurses work in township clinics; their daily tasks involve assisting with diagnosis and treatment, visiting villages to provide health consultations and check-ups, managing patients with chronic diseases, and providing health education. While some dimensions of these teaching and nursing jobs are better defined than those of clerical jobs, objective performance evaluation remains difficult. For example, due to the nonpermanent nature of the CGCS positions, CGCS teachers are often assigned to teach noncore courses (such as art or music) or lower grades (first to third grades), where there are no school-wide exams to test student performance, and thus student scores cannot be used to objectively evaluate the performance of these teachers.

Since the multi-dimensional and vaguely defined nature of CGCS job tasks makes it infeasible to objectively compare job performance across individuals, the evaluation of a CGCS relies solely on the evaluating supervisor's subjective assessment. This is also the norm for the vast majority of regular civil service jobs in China and across the world.

The only major difference between a CGCS position and a regular civil servant position is that the former is based on a two-year contract while the latter is

"tenured."[11] The majority of CGCSs are eager to be promoted to tenured positions upon finishing their two-year terms, which can only be approved by the government if the supervisor's evaluation is satisfactory.[12] As a result, CGCSs have exceptionally strong incentives to impress their evaluators. A potential byproduct of such high-powered incentives is the CGCSs' engagement in influence activities—either productive or nonproductive ones—directed toward their evaluators. Simple examples of productive influence activities include strategically allocating more efforts to job tasks assigned by the evaluator, working harder on job dimensions that are more observable or valuable to the evaluator, trying to get more involved in projects initiated by the evaluator, etc. Nonproductive influence activities consist of behaviors such as "buttering up" the evaluator, picking up the evaluator's kids from school, making tea for the evaluator, doing personal chores for the evaluator, or even directly bribing the evaluator.[13]

Under the dual-leadership governance structure, every CGCS reports to both a party leader and an administrative leader. In principle, the administrative leader is in charge of the day-to-day operation of the government entity, while the party leader oversees the process and has the final say in the most high-stakes decisions. These two leaders have the same official ranking, but the party leader is normally perceived to have an edge in authority. At the grassroots level, such as a township (which is the lowest layer of formal bureaucracy), the division of labor between the two leaders often becomes less clear, and there tends to be substantial overlap in their roles. This dual arrangement provides de facto checks and balances in local governance, including employee supervision (Li 2019). It is prevalent in many levels of administrative units, ranging from the central ministries to village committees. It is also implemented in public institutions such as schools, hospitals, and state-owned enterprises, as long as there are more than three Communist Party members among the employees.

Under the current evaluation scheme, when a CGCS is first assigned to a township by the provincial Department of Human Resources, she is explicitly told that the Department of Human Resources has designated one of the two leaders as the "evaluator" who is responsible for evaluating her performance at the end of the year.[14] The CGCS, therefore, knows whose opinion matters for her career development, starting at the beginning of her appointment. Nevertheless, the CGCS is hired to work for the entire organization rather than the specific evaluator, which means that she is expected to respond to the job tasks assigned by both leaders, even though only one of them will matter for her evaluation outcomes.

[11] In this setting, "tenure" corresponds to "*Bian Zhi* (编制)," which is essentially a permanent contract provided by the government.

[12] In the provinces where our study took place, about 40 percent of CGCSs subsequently become permanent civil servants.

[13] In contrast, productive activities valued by the entire organization are typically routine job tasks that can be observed and appreciated by both leaders. For example, for the CGCSs working as township government clerks, this category typically includes tasks like hosting visiting villagers and helping them benefit from existing social assistance programs, preparing policy documents to be submitted to upper-level governments by the organization, helping the organization adopt e-governance systems, attending meetings and discussions for the organization, etc.

[14] In our field interviews, we learned that the government decided to choose only one of the two supervisors to evaluate the CGCS in order to avoid potentially sensitive cases where the two supervisors give drastically different evaluations regarding the same CGCS, which might cause political or legal trouble for the government itself.

B. *Experimental Design*

In this section, we explain the experimental design and discuss the intuitions for our main hypotheses. A formal rationalization of the experiment is presented with a conceptual framework in Section II.

In collaboration with two provincial governments in China, in 2017, we randomly assigned the "revealed" and "masked" subjective performance evaluation schemes across all 3,785 CGCSs whom they employed in that year. For every CGCS in our sample, one of her two supervisors was randomly selected to be the evaluator, meaning that this supervisor's assessment was given all the weight in the final evaluation outcome. We also collected the nonevaluating supervisor's assessment of each CGCS's performance, but this assessment was given no weight in the actual evaluation. In both schemes, we never directly informed a supervisor whether or not he was chosen as the evaluator, nor did we inform the CGCS's colleagues.

Two-thirds of the CGCSs in our sample were assigned to the "revealed" scheme. In this scheme, we informed each CGCS about the identity of her evaluating supervisor at the beginning of the evaluation cycle. This mimics the current system of CGCS performance evaluation, where the agent is informed ex ante about the evaluating supervisor's identity. The key difference is that, in the current system, the evaluator is endogenously chosen from the two supervisors, typically through an opaque process combining supervisor opinions, division of evaluation duties between supervisors, and other idiosyncratic factors. Because our "revealed" scheme randomly selected the evaluator, endogeneity in evaluator selection was eliminated.

We exploit the revealed scheme to test whether knowing the evaluator's identity generates asymmetry in supervisor assessments. Since the evaluator was randomly selected, both supervisors should give similar assessments of CGCS performance on average, in the absence of any evaluator-specific influence activities. However, if the CGCS indeed engaged in evaluator-specific influence activities, we would expect to observe asymmetry in the two supervisors' assessments of the same CGCS.

The remaining one-third of the CGCSs were assigned to the "masked" scheme. In this scheme, while we still randomly selected one of the two supervisors as the evaluator, we did not inform the CGCS about the identity of the evaluator until the end of the evaluation cycle. Therefore, from the CGCS's perspective, each supervisor had a 50 percent chance of determining her evaluation outcome. Compared to the revealed scheme, the masked scheme reduced the relative return to supervisor-specific influence activities. If the CGCS put effort into influencing a specific supervisor, there was a 50 percent chance that this supervisor would not end up evaluating her performance, significantly reducing the expected benefit from engaging in influence activities. As a result, under the masked scheme, a CGCS had incentives to reallocate her efforts from influence activities toward productive activities that would be appreciated by both supervisors, which could improve overall work performance.

Exploiting the randomization of CGCSs into the "revealed" versus "masked" schemes, we can test whether introducing uncertainty about the evaluator's identity improves CGCS performance. Our benchmark performance indicator is the average assessment given by other colleagues. We define "colleagues" as coworkers in the same office as the CGCS, who were not hired through the "3+1 Supports" program. We consider the colleagues' assessments an informative performance measure in

this context for three reasons. First, the colleagues were randomly chosen from the same office where the CGCSs work. They worked closely with the CGCSs and could thus accurately observe the CGCSs' performance. Second, there is no obvious conflict of interest between the CGCSs and their colleagues. Most colleagues already have tenure and have worked in the office for many years. As a result, the CGCSs and their colleagues do not directly compete with each other for career advancement. Finally, the CGCSs did not have obvious incentives to influence their colleagues for evaluation purposes; at the beginning of the experiment, the provincial governments explicitly told each CGCS that only the evaluating supervisor's opinion would count for promotion.[15]

In addition to colleagues' assessments, we also measured CGCS performance in two other ways. First, we elicited performance assessments from both the evaluating supervisor and the nonevaluating supervisor. Using administrative data obtained from the provincial governments on the eventual career outcomes of the CGCSs, we verified that the evaluator's assessment is indeed important in determining the CGCS's promotion to a permanent position. Second, we tried to benchmark performance objectively using the actual salaries received by the CGCSs. While it is difficult to measure performance objectively due to the multi-dimensional nature of most CGCS jobs, a modest amount of "monthly bonus" is explicitly linked to certain well-defined performance indicators for some CGCS positions.[16] Therefore, we can compare the actual salaries received by CGCSs between the two schemes, and infer the differences in objective performance measures based on the bonus pay algorithms.

C. *Implementation*

Our experiment was conducted in collaboration with the governments of two large provinces in China, with a combined population of more than 150 million. Province A is coastal and more developed, while Province B is inland with a lower average income. Our sample covers all 3,785 CGCSs employed by these two provinces as of September 2017 (cohorts admitted in 2016 and 2017). Our research team was appointed by the two provincial Human Resources Departments as the third-party evaluator for their "3+1 Supports" programs to help pilot new performance evaluation schemes. The provincial governments officially informed all the CGCSs of this pilot. This high-level endorsement helped ensure that the vast majority of CGCSs were well aware of the high stakes involved in the evaluation outcomes under the newly introduced evaluation schemes.

The baseline survey was carried out in September 2017, one month after the 2017 CGCS cohort finished job training and received their assignments to positions. Every CGCS was then randomized into one of the two evaluation schemes. The randomization was conducted at the work unit level instead of the individual level.[17] Different CGCSs working in the same unit (i.e., an organization branch led by the

[15] Most CGCSs, in fact, did not even expect that we would survey their colleagues until the enumerators were sent to their workplaces at the end of the experiment.

[16] For example, CGCSs who serve as nurses receive bonuses based on the number of night shifts they work.

[17] In Chinese, a work unit corresponds to a "*Gong Zuo Dan Wei* (工作单位)."

same set of supervisors) were assigned to the same scheme. This was at the request of our government partners to ensure that the evaluation outcomes of CGCSs working in the same unit could be fairly compared to each other. Because 83.9 percent of the work units had only one CGCS assigned, randomizing at the work unit level instead of the individual level did not make any substantial difference statistically.

In September 2017, we informed every CGCS about the evaluation scheme to which she had been assigned. If a CGCS was randomized into the revealed scheme, we notified her that "among your two supervisors A and B, we randomly selected supervisor A to be your evaluator, whose opinion will be collected at the end of this evaluation cycle and provided to the provincial Human Resources Department for their review." If a CGCS was randomized into the masked scheme, we notified her that "among your two supervisors A and B, we will randomly select one of them to be your evaluator. The randomization will be determined at the end of this evaluation cycle, at which time the evaluator's opinion will be collected and provided to the provincial Human Resources Department for their review." The individualized notification letters are translated in online Appendix B.

To ensure the credibility of our intervention, the two provincial governments sent formal notifications with official stamps to every CGCS. The government notifications emphasized the importance of this "third-party" performance evaluation and confirmed the design of the evaluation schemes that we sent to the CGCSs. We reminded the CGCSs about their evaluation schemes in January 2018.

The end-line survey was carried out in June 2018, which consisted of three parts: colleague assessment, supervisor assessment, and self-assessment. When the enumerators visited the office where a CGCS worked, if there were fewer than five colleagues in the office, all of them were invited to fill in the colleague questionnaire; if there were more than five colleagues, the surveyor randomly sampled five of them to fill in the colleague questionnaire, using a random number generator.[18] To protect the privacy of colleagues and encourage truth-telling, colleague questionnaires were strictly anonymous, and CGCSs were not allowed to communicate with colleagues during the entire process. The CGCS survey was also conducted on-site, but independently from the colleague survey to avoid interference. Supervisor assessment was completed online, with an individual-specific link for each supervisor, listing all the CGCSs in his unit.

In the colleague and supervisor surveys, we collected information on the main characteristics of the colleague/supervisor, their interactions and familiarity with the CGCS, the job tasks of the CGCS, and their assessments of the CGCS along various dimensions. Specifically, we asked for an overall assessment of CGCS performance, as well as a "revealed preference" measure asking each colleague/supervisor whether he recommended that the CGCS be promoted to a permanent civil servant position in the current work unit.

The end-line CGCS survey followed a similar structure by asking about interactions with supervisors/colleagues and self-assessments along multiple dimensions.

[18] If a colleague was not at the office when the enumerator visited, his contact information was collected and he was surveyed over the phone the following day. To ensure data accuracy, the leader of the surveying team randomly called some of the surveyed colleagues on the following days to verify the sampling procedure and the answers collected.

Table 1—Balance Check: CGCS Characteristics

	Revealed scheme (1)	Masked scheme (2)	Difference (3)
Age	24.868	24.928	0.039
	(1.630)	(1.604)	(0.061)
Female	0.592	0.600	0.009
	(0.492)	(0.490)	(0.019)
Social science major	0.555	0.545	−0.015
	(0.497)	(0.498)	(0.020)
Four-year college or above	0.723	0.724	−0.004
	(0.448)	(0.447)	(0.017)
STEM students in high school	0.347	0.342	−0.006
	(0.476)	(0.475)	(0.020)
Party member	0.217	0.218	−0.002
	(0.412)	(0.413)	(0.017)
Parent completing college	0.288	0.285	−0.005
	(0.453)	(0.452)	(0.019)
Work in village	0.160	0.150	−0.012
	(0.366)	(0.357)	(0.015)
CEE score (100 points)	4.803	4.832	0.045
	(0.715)	(0.702)	(0.035)
Risk averse	0.471	0.477	−0.000
	(0.499)	(0.500)	(0.021)
Locally born	0.684	0.678	0.002
	(0.465)	(0.468)	(0.016)
Joint test p-value	–	–	0.54
Observations	1,935	919	2,854

Notes: The first two columns summarize the mean and standard deviation of CGCS characteristics. Column 1 uses the sample of CGCSs in the revealed scheme, column 2 uses the sample of CGCSs in the masked scheme. Column 3 checks the covariate balance between the revealed group and the masked group, controlling for county FE, CGCS type FE, and cohort FE, with standard errors clustered at the work unit level. A joint significance test of all variables presented in the table yields an F-statistic of 0.90 with the corresponding p-value of 0.54.

In addition, we also asked a series of questions related to future career plans and satisfaction with the "3+1 Supports" program. After the experiment, we collected administrative data on salaries and promotion outcomes for the CGCSs in our sample through our government partners. We list the key variables used in this paper and their sources in online Appendix Table A1.

D. *Balance and Attrition Tests*

To ensure that the randomization was well executed, we conduct a series of balance tests. Table 1 reports the summary statistics of the CGCSs' characteristics and the differences in these variables between the revealed and masked schemes. All the characteristics are balanced across the two schemes, suggesting our randomization was well executed. In online Appendix Tables A2 and A3, we also report balance tests for supervisor characteristics and colleague characteristics, and in online Appendix Table A4, we further test whether supervisor characteristics in the revealed scheme are balanced between the evaluating and nonevaluating supervisors.

Between the baseline and end-line surveys, we lost 929 (24.5 percent) CGCSs in the sample. The main cause for attrition was that some CGCSs or their supervisors were reassigned to different job posts during our study period (14.9 percent). For example, a CGCS could be relocated from one township to another because of changes in government priorities. The supervisors could retire or be promoted or rotated to other institutions. Such job changes would break the supervisor-subordinate relationship defined by our intervention and thus invalidate the experimental design. In addition, some CGCSs passed the formal civil service exams or were admitted to graduate schools and thus decided to quit their jobs during our experiment (7.4 percent).

To test whether our experiment suffers from potential attrition bias, we regress the attrition status on the treatment status in online Appendix Table A5. We find that the masked scheme does not increase overall attrition, nor does it predict any specific type of attrition. To further investigate the potential impacts of CGCS attrition on our findings, in online Appendix Table A6, we regress a CGCS's attrition status on her baseline characteristics and their interaction terms with our treatment variable for the masked scheme. As can be seen, while several covariates are correlated with attrition (college type, College Entrance Examination (CEE) score, age, parental education, and social science (SOSC) major), these types of attrition do not systematically differ between treatment and control groups.

In addition, as will be elaborated in Section III, we try to correct for potential nonrandom sample selection by applying Lee bounds to our baseline analyses (Lee 2009) and find that the baseline results hold. We therefore conclude that, while CGCS attrition is common in this institutional context, it has limited impact on the empirical analyses presented in this paper.

II. Conceptual Framework

In this section, we conceptualize our empirical setting and experimental design, and derive the main hypotheses that will guide the empirical analysis.

Assume that a CGCS's work performance can be (at least partially) observed by her supervisors and coworkers but cannot be verified quantitatively. The organization therefore relies on a subjective performance evaluation scheme, where the agent's reward depends on the assessment given by her evaluator. To mimic our empirical setting, we assume that there are two supervisors, $j \in \{1, 2\}$. The CGCS allocates her efforts across three dimensions. First, she can work on the "common productive dimensions" of the job (X), which can be observed and appreciated by both supervisors. Second, she can work on "supervisor-specific productive tasks" (x_j), which are assigned or observed solely by supervisor j. Finally, she can exert nonproductive efforts to personally flatter a supervisor (u_j). Following Milgrom and Roberts (1988), we categorize x_j as "productive influence activities" and u_j as "nonproductive influence activities."[19]

[19] The symmetry between the two supervisors in the model corresponds to the fact that we have randomly assigned one of them to be the evaluator in the experiment. Therefore, in the data, they are on average balanced along different dimensions.

From the point of view of the organization, only productive activities contribute to the overall performance of the CGCS,

$$P = X + x_1 + x_2.$$

In contrast, a supervisor j values all three types of activities—common productive activities (X), productive influence activities directed toward him (x_j), and nonproductive influence activities directed toward him (u_j). The assessment score of supervisor j is thus given by

$$Y_j = \alpha X + x_j + u_j, \quad j = 1,2,$$

where $\alpha > 0$ measures the relative weight that a supervisor places on the common productive activities over the supervisor-specific influence activities.

Each CGCS maximizes her utility subject to a time constraint:

$$\max_{X,x,u} V = \alpha X + \sum_{j=1,2} s_j(x_j + u_j) - G(X) - g\Big(\sum x_j\Big) - h\Big(\sum u_j\Big),$$

$$\text{subject to} \quad X + \sum x_j + \sum u_j = T, \quad X, x_j, u_j \in [0,T],$$

where s_j is the probability of each supervisor j's assessment being used to determine the CGCS's reward in the performance evaluation scheme $(\sum_{j\in\{1,2\}} s_j = 1)$. The costs of working on different activities follow strictly convex functions $G(\mathrm{X})$, $g\big(\sum x_j\big)$, and $h\big(\sum u_j\big)$. T is the total time budget for an individual.

When the CGCS is informed about the identity of her evaluator (revealed scheme), she knows exactly whose opinion matters for her career development: $s_1 = 1$, $s_2 = 0$; or $s_1 = 0, s_2 = 1$. When the CGCS is not informed about the evaluator's identity until the end of the evaluation cycle (masked scheme), she perceives each supervisor as equally likely to determine her career development: $s_1 = s_2 = 1/2$.

Solving the CGCS's maximization problem in the two schemes, we can derive the main hypotheses that will guide the empirical investigations. We summarize the main propositions and briefly discuss their intuitions below. We provide more detailed proofs and model extensions in online Appendix C.[20]

PROPOSITION 1: *Under the revealed scheme, the agent engages in evaluator-specific influence activities* (x_j, u_j)*, and the evaluating supervisor gives a higher assessment* (Y_j) *than the nonevaluating supervisor.*

[20] In online Appendix D, we evaluate the robustness of our model predictions with respect to alternative specifications. Specifically, under the current baseline model setup, we explore the more general case where the evaluation score of supervisor j is a nonlinear function of the three types of actions: $E_j = F(X + x_j + u_j)$. We find that all the model predictions remain unchanged under simple regularity conditions. While our model assumptions are fairly general, they are meant to rationalize our specific institutional setting, and there could exist different organizational environments where our propositions no longer hold. The empirical results should thus be viewed as results in this setting that are consistent with the model, rather than a more generalized test of the model. In online Appendix E, we use a generic model to demonstrate that our model predictions hold as long as there is sufficient substitutability between the common productive activities and influence activities.

DISCUSSION: *When the agent knows the identity of the evaluator, she has incentives to exert influence activities toward the evaluator (but not toward the nonevaluator), which leads to the evaluator giving more a positive assessment than the nonevaluator.*

PROPOSITION 2: *Compared to the revealed scheme, the masked scheme increases common productive efforts (X) and improves work performance (P). The masked scheme increases the nonevaluating supervisor's assessment score, but the assessment change is ambiguous for the evaluating supervisor.*

DISCUSSION: *When the agent does not know who the evaluator is, her expected return to supervisor-specific influence activities (either productive or nonproductive) is reduced by one-half, while her expected return to common productive activities remains unchanged. Therefore, she has incentives to reallocate her efforts from influence activities to common productive activities, thereby improving performance.*[21] *Under the masked scheme, both supervisors benefit from the increased common productive efforts, but the evaluator also suffers from reduced evaluator-specific influence activities; as a result, the masked scheme leads to an unambiguous increase in nonevaluator assessment, and an ambiguous change in evaluator assessment.*

III. Baseline Results

In this section, we present the experimental results. In the revealed scheme, we find that the assessment given by the (randomized) evaluating supervisor is substantially higher than that given by the (randomized) nonevaluating supervisor, which is consistent with Proposition 1 of our conceptual framework. When switching from the revealed scheme to the masked scheme, the asymmetry in supervisor assessments no longer exists. Instead, we find significant improvements in colleague assessments, nonevaluator's assessments, and performance payment, and no significant change in evaluator's assessments. Taken together, these findings are consistent with Proposition 2 of our conceptual framework.

Analysis of data on the eventual allocation of permanent contracts among CGCSs indicates that performance evaluation in this setting is not just a formality, instead, the evaluators' assessments are indeed given substantial weights in CGCS promotions. Further, under the revealed scheme, we find clear evidence for the existence of productive influence activities, as well as suggestive evidence for nonproductive influence activities. We discuss these findings in more detail below.

A. *Proposition 1: Asymmetry in Supervisor Assessments under Revealed Scheme*

First, we investigate *Proposition 1*: whether revealing the identity of the evaluator to the CGCS will cause the evaluator to be more positive about the CGCS than

[21] In our model, productive and nonproductive influence activities should comove with each other. So, a reduction in total influence activities indicates reductions in both types of influence activities. Given the fixed time budget, a reduction in nonproductive activities means an increase in total productive activities ($X + x$), which means better overall performance.

the nonevaluator. For each CGCS's two supervisors, we randomly label them as "Supervisor 1" and "Supervisor 2," and then use the subsample of CGCSs in the revealed scheme to estimate the following econometric model:

$$(1) \qquad Sup1\_Edge_{icst} = \alpha \times Sup1\_Eval_i + \gamma_c + \lambda_s + \phi_t + \epsilon_{icst},$$

where the outcome variable $Sup1\_Edge_{icsy}$ is defined as "Supervisor 1's assessment score minus Supervisor 2's assessment score for CGCS i," who is in county c, cohort t, and serves as CGCS type s.[22] $Sup1\_Eval_i$ is a dummy variable indicating whether CGCS i is being evaluated by supervisor 1 (instead of supervisor 2). γ_c, λ_s, and ϕ_t represent county FE, CGCS type FE, and cohort FE, respectively. Standard errors are clustered at the work unit level. Under this specification, since the evaluator is randomly chosen among the two supervisors for each CGCS, α causally identifies the additional positiveness of the evaluation due to being assigned as the evaluator.[23] Throughout all baseline specifications, we always include the same set of fixed effects and no control variables, to keep things consistent. Our results are highly robust to alternative specifications, as we show below.

As shown in column 1 of Table 2, for CGCSs in the revealed scheme, if a supervisor was chosen as the evaluator at the baseline, he indeed gave a more positive assessment at the end line (relative to his nonevaluating counterpart), and the magnitude of this "evaluator edge" in assessment scores is as large as 0.24 standard deviation. This asymmetry in supervisor assessments is consistent with the agent engaging in evaluator-specific influence activities to improve evaluation outcomes. If the "assessment asymmetry" documented in column 1 is indeed caused by evaluator-specific influence activities, as we have argued, it should only exist when the CGCS knows who the evaluator is. Under the masked scheme, when the CGCS no longer knows the identity of the evaluator, there should be no asymmetry in supervisor assessments. In column 2, we focus on the masked scheme where the randomly chosen evaluator's identity was not announced until the end of the evaluation cycle. As we can see, in the masked scheme, being selected as the evaluator indeed no longer leads to more positive assessments compared to the other nonevaluating supervisor.

In columns 3 and 4, we focus on an alternative outcome variable: a dummy that indicates whether Supervisor 1 is strictly more positive than Supervisor 2.[24] Again, we find that the evaluating supervisor is more likely to give a more positive assessment than the nonevaluating supervisor in the revealed scheme, a phenomenon that disappears in the masked scheme.

In online Appendix Figure A1, we plot the distributions of evaluator and nonevaluator assessment scores across the two schemes.[25] Reassuringly, the assessment asymmetry documented under the revealed scheme is not driven by outliers

[22] This represents the four types of CGCS positions: township government clerks focusing on poverty alleviation, township government clerks focusing on agricultural support, teachers in township primary schools, and nurses in township clinics.

[23] Here, we define "more positive" as Supervisor 1's score being strictly larger than that of Supervisor 2, with each assessment score ranging from 1 to 7.

[24] Since in many cases both supervisors give equal assessment scores, the mean of this variable is substantially smaller than 0.5.

[25] In online Appendix Figure A2, we also plot the distribution of the "Edge" variable in the two schemes.

Table 2—Revealing Supervisor Identity Leads to Evaluation Asymmetry

	Supervisor 1's score minus supervisor 2's score		Supervisor 1 is more positive than supervisor 2	
	(1)	(2)	(3)	(4)
Supervisor 1 evaluating	0.311	−0.097	0.075	0.024
	(0.082)	(0.121)	(0.028)	(0.042)
Sample	Revealed	Masked	Revealed	Masked
DV mean	−0.03	−0.00	0.29	0.29
DV SD	1.31	1.22	0.45	0.45
Observations	1,300	580	1,300	580
R^2	0.161	0.243	0.163	0.275

Notes: This table tests whether revealing the identity of the evaluator to the CGCS affects the evaluator's assessment of the CGCS's job performance. Each column represents a separate regression. County fixed effects, CGCS type fixed effects and cohort effects are included in all the regressions. Columns 1 and 3 use data from the revealed scheme only; columns 2 and 4 use data from the masked scheme only. A joint significance test of outcome variables in the revealed scheme (columns 1 and 3) yields an F-statistic of 7.63 with the corresponding p-value of 0. A joint significance test of outcome variables in the masked scheme (columns 2 and 4) yields an F-statistic of 1.74 with the corresponding p-value of 0.18. The p-value for a χ^2 test of coefficient equality between column 1 and column 2 is 0. The p-value for a χ^2 test of coefficient equality between column 3 and 4 is 0.25. Standard errors clustered at the work unit level are reported below the coefficients.

in the outcome variable (e.g., a few evaluators giving the highest scores or a few nonevaluators giving the lowest scores). Instead, we observe that the evaluators appear to be systematically more positive than the nonevaluators across the entire distribution in the revealed scheme (panel A), which no longer holds in the masked scheme (panel B).

In the online Appendix, we also provide a battery of robustness checks on Table 2. First, in online Appendix Table A7, we present the main results controlling for variables chosen by the post-double-selection method using LASSO from a large pool of predetermined covariates.[26] Our estimates remain essentially unchanged. Second, in online Appendix Table A8, we directly control for all basic CGCS characteristics, and the results again hold.[27] Third, in online Appendix Table A9, we correct for potential nonrandom sample selection by applying Lee bounds to our baseline analyses (Lee 2009), and our results are quantitatively similar. Finally, instead of assessing the evaluator's extra positiveness using the split-sample approach, we can estimate the effects using an interaction approach. Specifically, we can use the full sample and run a regression that includes three explanatory variables: "Supervisor 1 evaluating" dummy, the "masking" dummy, and their interactions. The regression

[26] Three sets of baseline covariates enter the LASSO selection. First, we include CGCS basic characteristics: age, gender, party membership, parental education, college type, and college major. Second, we include the characteristics of the evaluators: age, gender, work experience, and educational background. Third, we also include colleague characteristics: age, gender, tenure status, educational background, work experience, and relationship with CGCS.

[27] Controls include: CGCS's age, gender, college major, college type, high school track (STEM or not), party member status, parental education, work place (in village or not), risk attitude, and birth place (local or not).

Table 3—Impacts of Masking the Evaluator's Identity on Performances

	(1)	(2)	(3)	(4)
Panel A. Performances evaluated by colleagues				
	Performance (1–7)	Top 10 percent	Hardworking	Qualify for tenure
Masking	0.217	0.077	0.028	0.035
	(0.035)	(0.013)	(0.012)	(0.011)
DV mean	5.23	0.71	0.43	0.87
DV SD	0.92	0.33	0.43	0.26
Observations	2,837	2,837	2,837	2,837
Panel B. Performances evaluated by supervisors				
	Mean assessment (1–7)	Evaluator assessment	Nonevaluator assessment	Assessment Deviation
Masking	0.139	0.049	0.215	−0.100
	(0.046)	(0.055)	(0.059)	(0.050)
DV mean	5.14	5.19	5.11	0.90
DV SD	0.91	1.12	1.10	0.93
Observations	1,940	1,940	1,940	1,940
Panel C. Performance pay				
	Wage	ln(Wage)	Wage: Medical Support	ln(Wage: Medical Support)
Masking	48.81	0.02	115.54	0.05
	(22.41)	(0.01)	(61.94)	(0.03)
DV mean	2,103.73	7.61	1,851.58	7.51
DV SD	644.66	0.26	349.31	0.16
Observations	2,750	2,750	193	193

Notes: Each column represents a separate regression. County fixed effects, CGCS type fixed effects and cohort effects are included in all the regressions. A joint significance test of all the outcome variables in panel A yields an F-statistic of 11.36 with the corresponding p-value of 0. A joint significance test of all the outcome variables in panel B yields an F-statistic of 6.32 with the corresponding p-value of 0. Standard errors clustered at the work unit level are reported below the coefficients.

results from this interaction approach are reported in online Appendix Table A10, which carries similar information.

B. *Proposition 2: Improved Work Performance under the Masked Scheme*

As suggested by Proposition 2, the masked scheme could cause CGCSs to reallocate their efforts from evaluator-specific influence activities toward common productive tasks, leading to improved performance. To test this hypothesis, we evaluate the impacts of masking the evaluator's identity on a series of work performance measures. Table 3 presents the CGCSs' performance under a series of different performance indicators collected in our end-line surveys. We use the full sample of CGCSs (in both revealed and masked schemes), and estimate the following econometric model:

$$Y_{icst} = \alpha \times Mask_i + \gamma_c + \lambda_s + \phi_t + \epsilon_{icst}, \tag{2}$$

where Y_{icst} is a performance measure for CGCS i, who is in county c, cohort t, and serves as CGCS type s. $Mask_i$ is a dummy variable indicating whether CGCS i belongs to the masked scheme. γ_c, λ_s, and ϕ_t stand for county FE, CGCS type FE, and cohort FE, respectively. Because the CGCSs were randomly assigned between the two evaluation schemes, α identifies the causal effect of being assigned to the masked scheme (relative to the revealed scheme). The standard errors are clustered at the work unit level.

The multi-dimensional and subjective nature of the CGCS jobs means that we are unable to collect comprehensive objective performance indicators that are interpersonally comparable across all CGCSs, which is why the government had to use a subjective evaluation scheme for CGCS promotion in the first place. That being said, as shown in Table 3, we try our best to paint a more complete picture of the CGCSs' performance under different evaluation schemes, by investigating a series of different performance indicators collected in our end-line surveys.

Colleague Assessments.—First, in Table 3, panel A, we investigate how colleague assessment of CGCS performance varies between the revealed and masked schemes. As explained in Section IB, we consider colleague assessment to be an informative measure of a CGCS's performance in the common productive tasks that benefit the organization, which, according to Proposition 2, should improve under the masked scheme.[28]

In column 1, the dependent variable is the average colleague assessment of the CGCS's performance, which is framed relative to other civil servants employed in the same work unit. The assessment score in the questionnaire ranges from 1 to 7, representing different categories from "worse than all other colleagues" to "better than all other colleagues." Being assigned to the masked scheme led to significantly higher colleague assessment scores. To benchmark the treatment effect, in online Appendix Table A11, we show that the improvement in colleague assessment associated with masking the evaluator's identity is comparable to the performance gap between four-year regular college graduates and three-year community college graduates. This result suggests that the treatment effect of the masked scheme is economically significant.

This is further corroborated by columns 2 to 4. As shown in column 2, when we asked colleagues whether they thought the CGCS's performance ranked in the top 10 percent of the organization, CGCSs in the masked scheme were significantly more likely to be recognized as top performers. In column 3, we show that colleagues thought the CGCSs in the masked scheme were more hardworking. As column 4 shows, when we asked colleagues, hypothetically, whether they would recommend to the provincial government that the CGCS be promoted to a permanent position in this office after finishing her two-year term, more colleagues responded that the CGCS deserves "tenure" under the masked scheme.[29]

[28] Colleagues observe CGCS performance closely, but their opinions are not included in the performance evaluation scheme, so a CGCS is not incentivized to adjust her efforts to improve colleague assessments.

[29] The question is hypothetical, since only the evaluating supervisor's assessment of CGCS performance eventually gets used in determining the CGCS's promotion; neither the nonevaluating supervisor's assessment, nor the colleagues' assessments receive any weight in that decision.

Supervisor Assessments.—Second, Table 3, panel B shows how the assessments given by the supervisors change in the masked scheme. According to *Proposition 2*, we expect the nonevaluating supervisor to become more positive about CGCS performance under the masked scheme, while the change in the evaluating supervisor's assessment is theoretically ambiguous.[30]

In column 1, the outcome variable is the mean assessment of the two supervisors. We find that masking the identity of the evaluator significantly improves average supervisor assessment. In columns 2 and 3, we document that the increase in mean supervisor assessment in column 1 consists of an insignificant improvement in the evaluator's assessment and a significant improvement in the nonevaluator's assessment. This is consistent with the model's intuition that, when switching from the revealed scheme to the masked scheme, both supervisors benefit from increased CGCS efforts in the common productive dimensions, while the evaluator suffers from a reduction in evaluator-specific influence activities. As a result of these two forces, as shown in column 4, the difference between supervisor assessments decreases in the masked scheme.[31]

Performance Pay.—Third, in panel C of Table 3, we investigate the impacts of the masked scheme on monthly performance pay received by the CGCSs. Performance pay is explicitly tied to objective performance measures, such as attendance and working overtime. To the best of our knowledge, these monthly performance payments are not influenced by the yearly subjective performance evaluations, and can thus provide another independent (albeit incomplete) benchmark for CGCS performance.

Specifically, in the end-line survey, we asked each CGCS to report her total monthly remuneration, including basic wages and performance bonuses (if any), which we later verified using administrative information provided by the provincial governments. The basic wage is set by the county government and matches the entry-level permanent civil servant wage, so it should be exactly the same for all CGCSs within the same county, conditional on enrollment year and CGCS type. In addition to the basic wage, some work units have discretion over a modest amount of bonuses to reward the best performing employees (based on their own criteria). In columns 1 and 2 of panel C, we show that, on average, the CGCSs in the masked scheme earned 50 Chinese yuan (2.3 percent) higher remunerations than those in the revealed scheme. Since the basic salary for CGCSs is fixed, this income gap reflects the difference in performance bonuses.

During our field interviews, we were informed that the CGCSs who work as nurses in township clinics enjoy the most substantial performance bonuses, because these clinics have a "business" feature and can keep some profits to reward the hardest-working staff. For nurses, the number of night shifts taken each month is the main determinant of performance pay: every additional night shift is rewarded by

[30] In online Appendix C, we use a numerical example to demonstrate the ambiguous impacts of the masked scheme on the evaluator's assessment.

[31] The construction of the outcome variables in Table 3, panel B requires that neither supervisor's assessment is missing in the end line survey, hence the smaller number of observations. As shown in online Appendix Table A12, we find no evidence that the absence of the supervisor assessment variable is correlated with our experimental intervention. The results are also robust to the application of Lee bounds.

about 20 Chinese yuan (about $3). In columns 3 and 4 of panel C, when we restrict the sample to CGCSs working as nurses, we find an income gap greater than 115 Chinese yuan (6.2 percent) between the two schemes. The compensation differential between the revealed and masked groups is therefore equivalent to nearly six additional night shifts per month. This result suggests that the performance improvement caused by the masked scheme is indeed substantial when benchmarked objectively.

Taken together, as reflected by colleague assessments, supervisor assessments, and performance pay, the evidence consistently suggests that CGCS performance improved in common productive dimensions, and the magnitude of this improvement is economically significant. These findings thus support Proposition 2 of our model.

Furthermore, Figure 1 shows the distributions of the performance measures between the two evaluating schemes. In parallel with Table 3, for each of the three main performance indicators (average colleague assessment, average supervisor assessment, average salary), we plot its distributions under the revealed and masked schemes, respectively.[32] As we can see, the estimated impacts of the masked scheme are not driven by outliers, such as a few colleagues giving out minimum assessment scores under the revealed scheme. Instead, for each performance measure, the changes induced by the masked scheme appear to be spread out across the entire distribution. This pattern indicates that the masked scheme led to performance improvements for a wide range of CGCSs, rather than just a few of them concentrated in the tails of the performance distribution.

We check the robustness of the results in Table 3 in several ways. First, in online Appendix Table A13, we present the main results controlling for variables chosen by the post-double-selection method using LASSO. Second, in online Appendix Table A14, we control for all basic CGCS characteristics.[33] In both exercises, the results are similar to the baseline findings. We also correct for potential nonrandom sample selection by applying Lee bounds (Lee 2009) and report the results in online Appendix Table A15. We find that most of the bounds estimates are close to the baselines, although the estimates are noisier for performance pay measures.[34]

C. *CGCS Promotion Outcomes*

In principle, as long as the CGCSs perceive their evaluators' assessments to be important, such perceptions would generate incentives to engage in influence activities, regardless of whether or not the provincial governments eventually followed the evaluators' assessments. However, in a repeated game, it is important that the provincial governments live up to their promised evaluation schemes, in order to keep incentivizing future CGCSs. Therefore, we try to verify the extent to which evaluator assessments eventually affected the CGCSs' promotion chances, using

[32] In online Appendix Figure A3, we also separately plot the distributions of evaluator and nonevaluator assessment scores under the two schemes.

[33] Controls include the CGCS's age, gender, college major, college type, high school track (STEM or not), party member status, parental education, work place (in village or not), risk attitude, and birth place (local or not).

[34] The larger standard errors in the Lee bounds estimates could be driven by the fact that we are unable to control for the full sets of fixed effects in this estimation procedure.

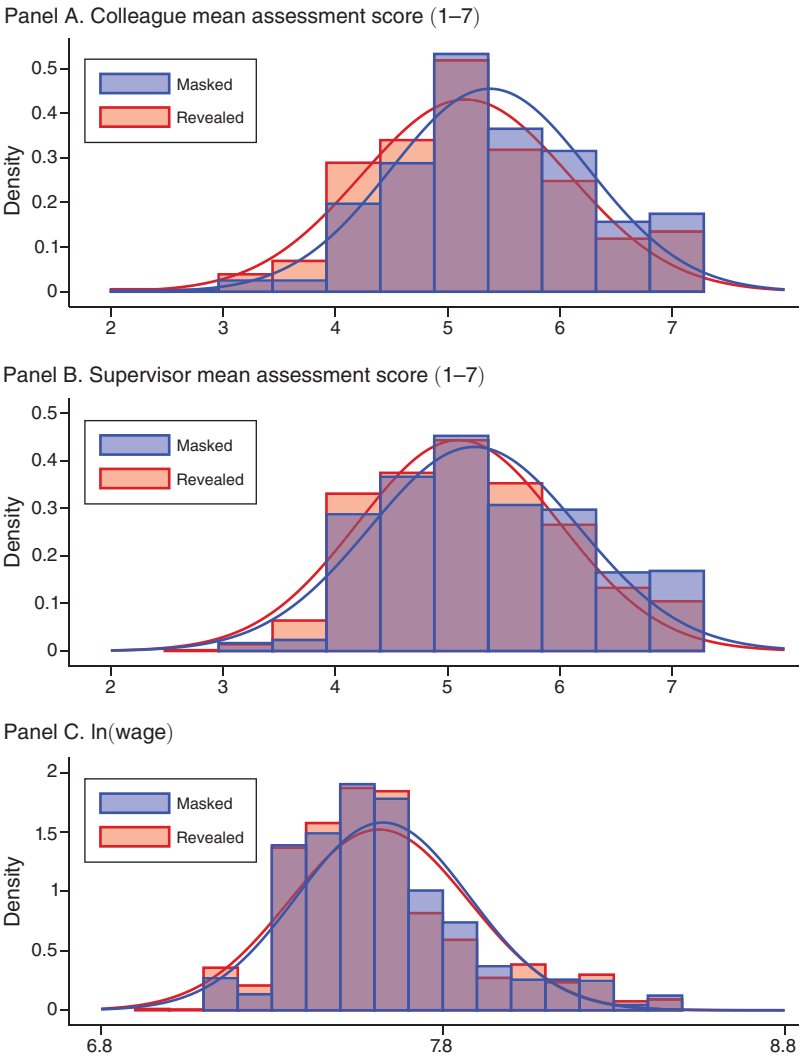

FIGURE 1. PERFORMANCE DIFFERENCES BETWEEN THE TWO EVALUATION SCHEMES

Notes: This figure plots the distributions of three performance measures separately for the masked (light blue) and revealed group (light red). Normal distribution curves are overlaid onto the associated histograms, i.e., the blue line for the blue histogram and the red line for the red histogram.

administrative records on the assignment of permanent civil service positions among the CGCSs in our sample, upon the completion of their two-year contracts.

Table 4 summarizes our findings. We have three observations. First, there is a strong positive correlation between the evaluating supervisors' assessment and eventual promotions to permanent civil service positions. Our point estimate indicates that a one-point increase in evaluator assessment score (on a scale of 1 to 7) increases the CGCS's chance of promotion by 7.5 percent, confirming that the evaluating supervisors' opinions carry significant weights in promotion decisions. Second, conditional on the evaluators' assessments, the nonevaluators' opinions

Table 4—Performance Evaluation and Tenure Decisions

	CGCS tenured			
	(1)	(2)	(3)	(4)
Evaluating sup's score	0.073	0.074	0.073	
	(0.011)	(0.013)	(0.022)	
Nonevaluating sup's score	0.015	0.020	0.006	
	(0.011)	(0.014)	(0.022)	
Masking				0.014
				(0.024)
DV mean	0.45	0.44	0.47	0.45
DV SD	0.50	0.50	0.50	0.50
Observations	1,940	1,300	580	1,940
Sample	All	Revealed	Masked	All

Notes: Each column represents a separate regression. County fixed effects, CGCS type fixed effects and cohort effects are included in all the regressions. The outcome variable is whether a CGCS becomes "tenured" after his two-year contract. The *p*-value for a χ^2 test of coefficient equality between column 2 and column 3 for "*Evaluating sup's score*" is 0.95 and the corresponding *p*-value for "*Nonevaluating sup's score*" is 0.54. Standard errors clustered at the work unit level are reported below the coefficients.

have no predictive power on the eventual promotion decisions. These patterns are salient in the full sample (column 1), as well as in both the revealed (column 2) and masked (column 3) scheme subsamples. Third, the masking treatment does not have a strong impact on promotion, as shown in column 4. While the estimated treatment effect is positive, it is small in magnitude and statistically insignificant. This result is consistent with our baseline finding: masking the identity of the evaluator mainly increases the nonevaluating supervisor's assessment, rather than the evaluating supervisor's assessment (panel B of Table 3). Given that only the evaluating supervisor's assessment is taken into account for promotion decision, it is not surprising that there is no significant difference in promotion rates between the two evaluation schemes.[35]

Taken together, these findings confirm the premise that the provincial governments rely heavily on the evaluators' assessments when making promotion decisions. This helps explain why revealing/masking the evaluators' identities has such salient impacts on the CGCSs' work behaviors.

We also investigate how introducing the masked scheme affects the selection into the public sector, by testing whether different types of CGCSs have different likelihoods of getting permanent positions under the revealed versus masked schemes. As shown in online Appendix Table A17, overall, introducing the masked scheme has a rather limited impact on political selection. That said, in column 2, we do find suggestive evidence that the masked scheme is marginally more likely to result in the

[35] In addition, in the end-line survey, 92 percent of the CGCSs reported to us that they were planning to apply for permanent civil service positions. As shown in online Appendix Table A16, their intention to apply for the permanent positions is not correlated with the assessment scores provided by the two supervisors, nor is it affected by the introduction of the masked scheme. In other words, our experiment does not change the CGCSs' willingness to work in civil service positions.

promotion of CGCSs who graduated from four-year regular colleges (as opposed to three-year community colleges). This is consistent with the masked scheme inducing positive political selection in terms of human capital.

IV. Mechanisms

In this section, we discuss the mechanisms of our findings and investigate several alternative interpretations. Specifically, in Section IVA, we discuss evidence on productive versus nonproductive influence activities; in Section IVB, we discuss how influence activities are affected by the matching between CGCSs and supervisors; in Section IVC, we discuss alternative interpretations of asymmetric supervisor assessments under the revealed scheme; and, in Section IVD, we discuss alternative interpretations of the improved colleague and supervisor assessments under the masked scheme.

A. *Types of Influence Activities*

In this subsection, we discuss the relevance of each type of influence activity—productive and nonproductive—in our setting. In Table 5, we investigate the existence of productive influence activities under the revealed scheme. In the end-line survey, we asked each CGCS, "among all the job tasks you need to do, what is the proportion that is assigned by each supervisor?" In column 1, we find that, if Supervisor 1 is chosen as the evaluator, the CGCS reports that she has more job tasks assigned by Supervisor 1. In addition, we asked each CGCS, "is your most important job task assigned by Supervisor 1 or Supervisor 2?" As shown in column 3, when Supervisor 1 has been revealed as the evaluator, the GCGS is more likely to think her most important job task is assigned by Supervisor 1. Finally, we asked the CGCS, "among all the job dimensions, in which dimension do you think you improved most in the past year?" We find that, when a job dimension is deemed important by the revealed evaluator, the CGCS is more likely to improve along this specific dimension (column 5). In contrast, none of these patterns exist under the masked scheme (columns 2, 4, and 6).

The results in Table 5 point to the prevalence of productive influence activities in our context. In contrast, for nonproductive influence activities, while we cannot directly observe such behaviors, we try to gauge their existence through indirect evidence.

In Table 6, we infer the importance of nonproductive influence activities by examining a series of value questions that we elicited from the CGCSs in our end-line survey. First, we asked the CGCSs, "what was the most challenging part of your CGCS experience?"[36] As shown in Table 6, column 1, the CGCSs under the revealed scheme were significantly more likely to report that "handling personal relationships with the supervisors" was the most challenging part of their experience, as compared to

[36] The choices included "familiarizing myself with the local governance system," "handling the personal relationship with my supervisor," "handling personal relationships with my colleagues," "adjusting to life in a rural area," "working on tasks unrelated to my college major," "adjusting to unfamiliar work and life conditions," "getting useful work feedback," and "other challenges."

TABLE 5—EVIDENCE ON THE EXISTENCE OF PRODUCTIVE INFLUENCE ACTIVITIES UNDER THE REVEALED SCHEME

	Ratio of job tasks assigned by supervisor 1 (reported by CGCS)		The most important job task is guided by supervisor 1 (reported by CGCS)		CGCS improved more in areas deemed important by sup 1 (relative to sup 2)	
	(1)	(2)	(3)	(4)	(5)	(6)
Supervisor 1 evaluating	0.031	−0.015	0.072	−0.015	0.132	−0.006
	(0.014)	(0.023)	(0.032)	(0.050)	(0.058)	(0.095)
Sample	Revealed	Masked	Revealed	Masked	Revealed	Masked
DV mean	0.48	0.49	0.44	0.46	0.04	−0.03
DV SD	0.24	0.24	0.50	0.50	1.06	1.04
Observations	1,482	659	1,134	529	1,482	659

Notes: Each column represents a separate regression. County fixed effects, CGCS type fixed effects and cohort effects are included in all the regressions. A joint significance test of outcome variables in the revealed scheme (columns 1, 3, and 5) yields an F-statistic of 1.93 with the corresponding p-value of 0.10. A joint significance test of outcome variables in the masked scheme (columns 2, 4, and 6) yields an F-statistic of 0.39 with the corresponding p-value of 0.81. The p-value for a χ^2 test of coefficient equality between column 1 and column 2 is 0.03. The p-value for a χ^2 test of coefficient equality between column 3 and column 4 is 0.05. The p-value for a χ^2 test of coefficient equality between column 5 and column 6 is 0.06. Standard errors clustered at the work unit level are reported below the coefficients.

TABLE 6—TREATMENT EFFECTS ON INFLUENCE ACTIVITIES AND WORK EFFORTS

	CGCS challenge: supervisor relationship	CGCS challenge: colleague relationship	CGCS belief: civil service is meritocratic	CGCS belief: hard work pays off
	(1)	(2)	(3)	(4)
Masking	−0.030	−0.003	0.017	0.024
	(0.014)	(0.009)	(0.009)	(0.012)
DV mean	0.15	0.05	0.94	0.90
DV SD	0.36	0.22	0.23	0.30
Observations	2,839	2,839	2,839	2,839

Notes: Each column represents a separate regression. County fixed effects, CGCS type fixed effects and cohort effects are included in all the regressions. A joint significance test of outcome variables in columns 1, 3, and 4 yields an F-statistic of 2.42 with the corresponding p-value of 0.06. The p-value for a χ^2 test of coefficient equality between column 1 and column 2 is 0.09. Standard errors clustered at the work unit level are reported below the coefficients.

their peers in the masked scheme. In contrast, as shown in column 2, the proportion of CGCSs identifying "handling personal relationships with colleagues" as the most challenging part of the experience is the same across the two schemes. These two results are consistent with our model, in which the CGCS engaged in more nonproductive influence activities under the revealed scheme than under the masked scheme, and did not have incentives to influence her colleagues under either scheme. We then asked each CGCS, "do you think the civil service system is meritocratic?" In column 3, we find that the CGCSs under the masked scheme were significantly more likely to give a positive answer to this question. In column 4, we also asked the CGCS, "do you think hard work pays off in your position?" Similarly, we find that

the masked scheme made the CGCS more likely to believe that hard work pays off. These results are consistent with our interpretation that the CGCS reallocates her efforts from nonproductive influence activities to common productive tasks under the masked scheme.

Admittedly, the suggestive evidence on nonproductive influence activities is indirect and could potentially be reconciled with other interpretations. Therefore, one should interpret these findings with caution. That said, it is worth noting that, in our conceptual framework, productive and nonproductive influence activities would comove with each other, and thus have the same qualitative implications.

B. *Evaluator Characteristics and Influence Activities*

We also investigate how influence activities are affected by the characteristics of the randomized evaluators. One particularly important dimension of heterogeneity is whether a CGCS and her evaluator come from the same county, since "hometown favoritism" is well-documented as playing an important role in China's bureaucratic system (Fisman and Wang 2015; Fisman et al. 2020). Specifically, hometown favoritism can be viewed as the gap in assessment scores between a "same-hometown evaluator" and a "different-hometown evaluator," holding constant the actual performance of the CGCS. Such favoritism can be decomposed into two parts: (1) the "top-down" preference, meaning that an evaluator would spontaneously assess a same-hometown CGCS more positively; and (2) the "bottom-up" influence activities, meaning that a CGCS would find it easier to influence an evaluator who is from the same hometown.

In the revealed scheme, the evaluator and the CGCS are made aware of each other's identity, so both "top-down preference" and "bottom-up influence" could be at work. In the masked scheme, however, the evaluator is aware of the identity of the CGCS, but not the other way around, which keeps the "top-down preference" while alleviating "bottom-up influence." Therefore, by comparing the magnitude of "hometown favoritism" across the revealed versus masked schemes, we can infer the relative importance of influence activities.

The results are presented in panel A of Table 7.[37] In column 1, we find that, if the CGCS shares a hometown with the evaluator, the evaluator indeed gives a higher performance assessment, confirming the existence of hometown favoritism in this setting. Then, we examine this favoritism separately for the revealed-scheme sample and the masked-scheme sample in columns 2 and 3. We find that the hometown favoritism can only be observed under the revealed scheme, but not under the masked scheme.

We can draw two conclusions from these results. First, because the "top-down" preference should remain the same across different evaluation schemes, the results suggest that the "bottom-up" influence activities are likely driving the observed hometown favoritism in our data. Second, because the assessment scores of both supervisors are uncorrelated with the "hometown tie" in the masked scheme, we can infer that, without influence activities from the CGCS, the hometown tie alone could

[37] The sample size is smaller due to missing values for supervisors' hometowns.

Table 7—Evaluation Characteristics and Influence Activities

	Evaluator assessment score			Nonevaluator assessment score		
	Full sample (1)	Revealed sample (2)	Masked sample (3)	Full sample (4)	Revealed sample (5)	Masked sample (6)
Panel A. Hometown favoritism and influence activities						
Same hometown	0.102	0.189	−0.067	−0.016	−0.045	0.046
	(0.051)	(0.067)	(0.088)	(0.049)	(0.060)	(0.100)
Observations	2,307	1,548	700	2,274	1,542	676
Panel B. Party-leader specific impacts						
Party leader evaluator	0.041	0.023	0.045	−0.007	−0.063	0.076
	(0.046)	(0.058)	(0.091)	(0.048)	(0.058)	(0.098)
Observations	2,307	1,548	700	2,274	1,542	676
Panel C. Gender-specific impacts						
Same gender evaluator	−0.024	0.003	−0.030	0.000	0.008	−0.063
	(0.050)	(0.064)	(0.094)	(0.051)	(0.061)	(0.106)
Observations	2,307	1,548	700	2,274	1,542	676
Panel D. Same education impacts						
Same education evaluator	−0.023	0.055	−0.195	0.016	−0.006	0.080
	(0.070)	(0.087)	(0.130)	(0.065)	(0.079)	(0.126)
Observations	2,179	1,454	670	2,149	1,449	642

Notes: Each column represents a separate regression. County fixed effects, CGCS type fixed effects, and cohort effects are included in all the regressions. Columns 1 and 4 use the full sample of CGCSs, columns 2 and 5 use the sample of CGCSs in the revealed scheme, columns 3 and 6 use the sample of CGCSs in the masked scheme. For panel A, the p-value for a χ^2 test of coefficient equality between column 2 and column 3 for the "*Same hometown*" variable is 0.01, and the p-value for a χ^2 test of coefficient equality between column 5 and column 6 is 0.39. For other panels, we cannot reject the null hypothesis that the estimated coefficient is equal between the revealed scheme and the masked scheme. Standard errors clustered at the work unit level are reported below the coefficients.

not generate meaningful favoritism in this setting. These findings further testify to the importance of influence activities in China's bureaucratic system.

Following the same specification, in panels B, C, and D of Table 7, we explore other dimensions of heterogeneity driven by CGCS-evaluator matching, such as "the evaluator being a party (rather than administrative) leader," "the evaluator and CGCS having the same gender," and "the evaluator and CGCS both graduated from college." We find that these other characteristics do not create significant heterogeneities in evaluation outcomes, in either the revealed or masked scheme. Taken together, these heterogeneity results are consistent with the conventional wisdom that "hometown ties (*Tongxiang*)" underlie the strongest connection in China's bureaucratic system.

C. *Alternative Interpretations of Asymmetric Supervisor Assessments under the Revealed Scheme*

Our interpretation of the findings in Table 2 is based on Proposition 1: in the revealed scheme, the CGCS is able to perform evaluator-specific influence activities. There are two potential confounding explanations.

1. Evaluating Supervisor Finding Out about His Role.—In the revealed scheme, while the supervisors were not directly informed by the research team about their

Table 8—Behavioral Changes of the Evaluating Supervisors

	Tasks assigned reported by supervisors (sup 1–sup 2) (1)	List the number of CGCS' main tasks (sup 1–sup 2) (2)	Number of words in describing CGCS's job tasks (sup 1–sup 2) (3)	Familiar with work (sup 1–sup 2) (4)	Familiar with life (sup1–sup2) (5)	Not responding to the survey (sup 1–sup 2) (6)
Panel A. Revealed scheme						
Supervisor 1 evaluating	−0.590	0.233	0.614	0.527	−0.678	−0.009
	(0.649)	(0.236)	(0.528)	(1.056)	(1.438)	(0.020)
DV mean	−0.91	−0.34	−0.99	−0.50	−0.22	0.21
DV SD	10.36	3.71	8.75	17.43	23.66	0.41
Observations	1,288	1,300	1,300	1,300	1,300	1,910
Panel B. Masked scheme						
Supervisor 1 evaluating	−1.187	0.081	0.362	0.041	−1.539	−0.056
	(1.062)	(0.393)	(0.736)	(1.591)	(2.108)	(0.030)
DV mean	−0.98	−0.33	−1.18	0.48	0.08	0.22
DV SD	10.49	3.83	8.22	17.56	22.69	0.41
Observations	577	580	580	580	580	869

Notes: Each column represents a separate regression. County fixed effects, CGCS type fixed effects, and cohort effects are included in all the regressions. A joint significance test of all the outcome variables in panel A yields an F-statistic of 0.58 with the corresponding p-value of 0.71. A joint significance test of all the outcome variables in panel B yields an F-statistic of 0.49 with the corresponding p-value of 0.78. We cannot reject the null hypothesis that the estimated coefficient is equal between the revealed scheme and the mask scheme for all the outcome variables. Standard errors clustered at the work unit level are reported below the coefficients.

roles in the evaluation, it is still possible that some of them might have found out about this from the CGCSs. If the evaluator found out about his role, he might have solicited personal favors from the CGCS, which could have increased the influence activities practiced by the CGCS. Another possibility is that, knowing his role as evaluator, a supervisor might change his behaviors in supervising and evaluating the CGCS, such as paying closer attention to her work, providing more frequent feedback, or being more supportive of her career, which might affect his assessment scores through channels that are independent from influence activities. This type of evaluator behavioral change could potentially confound the empirical patterns documented in Table 2 (i.e., asymmetry in supervisor assessments under the revealed scheme).

To examine this potential confounding mechanism, in Table 8, we directly investigate whether being selected as the evaluator changes a supervisor's supervising and evaluating patterns. As shown in panel A, under the revealed scheme, there is no detectable difference between the randomized evaluators and nonevaluators along any behavioral dimensions that we could measure in the end-line survey: the total number of job tasks assigned to the CGCS, the number of important job tasks assigned to the CGCS, the number of words they used to describe the job tasks of the CGCS, their familiarity with the CGCS's work and life situations, and their response rate in our end-line survey. In panel B, we see no systematic difference in supervisor behaviors under the masked scheme. Given these precisely estimated null results, it seems very unlikely that evaluator behavioral changes are driving the asymmetry in supervisor assessments that we observed in Table 2.

Furthermore, we attempt to directly measure whether the revealed-scheme evaluators figured out their roles, and whether this affected their behaviors. Specifically, in our end-line survey, for each supervisor under the revealed scheme, we directly asked him whether he was aware of his role in evaluating the CGCS.[38] It turns out that 65.5 percent of the revealed-scheme supervisors did not know whether they were chosen as evaluators until after they had finished their assessments of the CGCSs.

In online Appendix Table A18, we reestimate the specifications in Table 2 separately for the subsample in which supervisors did not know their evaluator roles, and the subsample in which supervisors did know their evaluator roles. We find that the asymmetry in supervisor assessments under the revealed scheme is almost identical in these two subsamples, suggesting that our results are not driven by some supervisors finding out about their roles in CGCS evaluation.[39] In addition, in online Appendix Table A19, we further document that the evaluators who found out about their roles did not behave differently from their nonevaluating counterparts along any of the behavioral dimensions.

Taken together, these findings indicate that "the evaluator finding out about his role" has little to do with his supervising and evaluating behaviors.

2. More Information for the Evaluating Supervisor.—Another confounding story is that the evaluating supervisor might receive more information regarding CGCS performance from various sources; the CGCS, colleagues, and the other (nonevaluating) supervisor might try to send signals to help him evaluate the CGCS. This increase in information might improve the evaluator's assessment and thus create the scoring asymmetry shown in Table 2. Again, we think this interpretation is unlikely to be of first-order importance, given that we never directly informed colleagues or supervisors about the evaluator's identity.

Nevertheless, we explicitly examine this alternative interpretation. In our end-line survey, we asked each supervisor, "how frequently did the CGCS, the colleagues of the CGCS, or the other supervisor discuss the CGCS's performance with you?" We are interested in whether the evaluating supervisor received more information than the nonevaluating supervisor from these three sources. In online Appendix Table A20 we show that, relative to the nonevaluator, the evaluator did not gain extra information from any of these sources.[40] Therefore, the asymmetry in supervisor assessments under the revealed scheme cannot be explained by differences in information between the two supervisors.

[38] This question was not asked of supervisors in the masked scheme. By construction, in the masked scheme, neither the CGCS nor her supervisors were informed about the identity of the chosen evaluator, and the supervisors couldn't possibly have found out about their roles.

[39] A related confounding mechanism is that an evaluator in the revealed scheme, after finding out about his role as evaluator, might look at the CGCS more kindly since he now felt "invested" in her career. However, this is inconsistent with our finding that, under the revealed scheme, the informed evaluators gave evaluations similar to those by the uninformed evaluators.

[40] If anything, the evaluator was 3 percent less likely to receive information regarding CGCS performance from colleagues, although the coefficient is small in magnitude and only marginally significant.

D. *Alternative Interpretations of Improved Assessments under the Masked Scheme*

Our interpretation of the "improved colleague and supervisor assessments under the masked scheme" is based on *Proposition 2*: masking the evaluator's identity makes supervisor-specific influence activities less beneficial, which incentivizes the CGCSs to work harder on the common productive dimensions that are appreciated by both supervisors, resulting in better work performance. There are five potential confounding explanations.

1. CGCS Influencing Both Supervisors More.—The first alternative interpretation is that, under the masked scheme, the CGCS did not work harder on common productive dimensions. Instead, she simply extended more influence activities toward both supervisors, which is why we see improved average supervisor assessment. However, this interpretation is inconsistent with a series of empirical results.

First, it is inconsistent with the fact that colleague assessments are substantially better under the masked scheme. As explained in Section I, a CGCS has no systematic incentive to influence her colleagues; every CGCS is clearly informed that only her evaluating supervisor's opinion will be taken into account by the provincial government, and colleague assessments will not enter into her promotion case. Therefore, if the CGCS is simply extending more influence activities toward both supervisors, rather than working harder, there should not be a significant improvement in average colleague assessment.

Second, if the CGCS is engaging in more influence activities instead of working harder, we should not observe objective performance improvements under the masked scheme. As discussed in Section III, CGCSs under the masked scheme receive substantially higher performance bonuses, which are directly linked to objective performance indicators. This, again, supports our interpretation and contradicts the competing hypothesis.

Third, as documented in Table 6, under the masked scheme, the CGCSs are less worried about handling personal relationships with supervisors, as compared to their peers under the revealed scheme. This is also consistent with a reduction in influence activities, rather than extending influence activities to both supervisors, under the masked scheme.

2. CGCS Influencing Colleagues under the Masked Scheme.—Suppose that CGCSs, for whatever reason, tried to influence their colleagues, and did so to a larger extent under the masked scheme. Could this be confounding our results on improved colleague assessments under the masked scheme? To begin with, this interpretation is inconsistent with the result in Table 6 column 2, which shows that the proportion of CGCSs worrying about "handling personal relationships with colleagues" remains the same across both schemes. It is also inconsistent with the increase in performance pay, which is linked to objective performance indicators rather than supervisor or colleague assessments.

To further rule out this confounding interpretation, we examine whether there exists hometown favoritism in colleague assessment. Recall that, in Table 7, we document the existence of hometown favoritism in supervisor assessment and show that "bottom-up influence activities" are driving such favoritism. We conduct a similar

exercise using colleague assessment and check whether the CGCSs had incentives to influence their colleagues under the masked scheme. As shown in online Appendix Table A21, a "same hometown colleague" does not show differential positiveness across the two schemes. This result further suggests that the CGCSs were unlikely to engage in additional influence activities toward colleagues under the masked scheme.

3. Higher Information Quality under the Masked Scheme.—Another possibility is that supervisors in the masked scheme get better information on CGCS performance, which might explain the increase in average supervisor assessment. To investigate this channel, in the end-line survey, we directly asked the supervisors about the sources from which they get information on CGCS performance (i.e., from CGCS or from other colleagues). In online Appendix Table A22, we examine whether supervisors received additional information on CGCS performance under the masked scheme, either from colleagues or from the CGCS herself. We find that the masked scheme did not increase the frequency of CGCSs and other colleagues reporting to either the evaluating supervisor or the nonevaluating supervisor regarding CGCS performance. This suggests that improved supervisor assessments in the masked scheme cannot be explained by changes in information quality.

4. Behavioral Changes from the Supervisors.—As explained in Section IVC, whether the evaluator knew about his role in the evaluation had limited impacts on his behavior. To further rule out the possibility that the treatment effects of the masked scheme might be confounded by some evaluators finding out about their roles under the revealed scheme (and by construction, not under the masked scheme), we compare CGCSs in the masked scheme to the subsample of revealed-scheme CGCSs whose evaluators know their roles, as well to the subsample of revealed-scheme CGCSs whose evaluators do not know their roles, respectively. As shown in online Appendix Table A23, the main findings in Table 3 remain robust when we use either subsample as the control group, suggesting that the impacts of the masked scheme are not driven by behavioral changes from supervisors who found out about their roles.

5. CGCS Gets Discouraged When Matched to "Hostile Evaluator" under Revealed Scheme.—A remaining possibility is that, under the revealed scheme, some CGCSs might be matched with an evaluator whom they perceive as hostile, in that, no matter how hard the CGCS works, efforts will not be appreciated by this evaluator. As a result, the CGCSs get discouraged and put little effort into productive tasks, which might explain why performance is higher under the masked scheme.

In our baseline survey, before the randomizations of schemes and evaluators were realized, we asked each CGCS, "among the two supervisors, whom would you prefer to be your evaluator?" Due to randomization, one-half of the CGCSs under the revealed scheme would be evaluated by their "nonpreferred" supervisor, and the other half evaluated by their "preferred" supervisor. Since the "discouragement" mechanism should operate only through those evaluated by the nonpreferred supervisor, we can compare performance differences between CGCSs facing the preferred supervisor under the revealed scheme and those under the masked scheme. If

discouragement were driving the observed improvement in CGCS performance, we should expect the performance improvement under the masked scheme to disappear in this restricted comparison. However, as shown in online Appendix Table A24, the masking effect remains similar in this subsample analysis, providing evidence against the "discouragement" interpretation.

V. Conclusion

Subjective evaluations are widely used in both the private and public sectors, especially in contexts where job tasks are inherently multi-dimensional and vaguely defined, making it impossible to obtain sharp objective measures of employee effort and performance. A key limitation to subjective evaluation is that it may distort the employee's incentives and make her more likely to cater to the evaluator's personal tastes or private interests rather than focusing on productive tasks that benefit the whole organization. Until now, rigorous empirical evidence on the existence and implications of influence activities has remained scarce.

To shed light on this topic, we conducted a large-scale field experiment, where we helped the government randomize two subjective performance evaluation schemes among 3,785 junior state employees in China. In the "revealed" scheme, we randomly chose one of the two supervisors as the performance evaluator and informed the subordinate *ex ante* about the evaluator's identity. Under this scheme, as expected, subordinates were induced to engage in evaluator-specific influence activities to improve their evaluation outcomes, which in turn would affect their promotion to permanent civil service positions.

In the "masked" scheme, we randomly chose one of the two supervisors as the performance evaluator, but the identity of the evaluator was not disclosed to the subordinate, which reduced the expected return to supervisor-specific influence activities. We hypothesized that masking the evaluator's identity would encourage the subordinate to reallocate her efforts from influence activities toward common productive dimensions that could be appreciated by both supervisors. We find that the masked evaluation scheme indeed improved the subordinate's work performance, as measured by average colleague assessments, average supervisor assessments, and monthly bonus payments determined by objective performance indicators.

We also distinguish between two types of influence activities. On the one hand, there are productive influence activities, where a multi-tasking agent works harder on tasks that are assigned or better observed by the evaluating supervisor. On the other hand, there could also be nonproductive influence activities, where the agent will try to benefit the evaluator through personal favors that go beyond her mandated tasks. We find consistent evidence demonstrating that productive influence activities are prevalent in China's local bureaucratic system, and some suggestive evidence indicating that nonproductive influence activities might also exist.

In addition to providing rigorous empirical evidence on the existence and implications of influence activities, our findings also have important policy implications. In a setting where multiple individuals could potentially assess an employee's performance with similar information quality, introducing uncertainty about the evaluator's identity (which has minimal implementation cost) can significantly improve

the job performance of government employees.[41] Further, this uncertainty results in more state employees believing that hard work pays off and that the bureaucratic system is meritocratic. These belief changes should have far-reaching consequences for the working culture and ethics of the Chinese government. Given that the vast majority of civil service jobs rely heavily on subjective performance evaluations, and given that every level of bureaucracy in China follows a dual-leadership structure, our findings should have direct implications for the more than 50 million state employees in China.

Going beyond the context of the Chinese bureaucracy, organizations around the world have increasingly adopted and institutionalized various dual-leadership arrangements, such as pairing a chief executive officer (CEO) with a chief operating officer (COO) in private firms, and "Office of the President" arrangements in public institutions (Miles and Watkins 2007; Williams and Scott 2012). In these settings, when high-stakes rewards are linked to the subjective opinions of designated evaluators, introducing uncertainty in the subjective evaluation scheme could potentially lead to performance improvements. More generally, as pointed out by Ederer, Holden, and Meyer (2018), even in objective incentive schemes, when there exist moral hazard problems due to the agent's superior knowledge of the environment, introducing uncertainty in the payment rule could systematically reduce gaming and improve performance.

REFERENCES

Alonso, Ricardo, Wouter Dessein, and Niko Matouschek. 2008. "When Does Coordination Require Centralization?" *American Economic Review* 98 (1): 145–79.

Ashraf, Nava, and Oriana Bandiera. 2018. "Social Incentives in Organizations." *Annual Review of Economics* 10: 439–63.

Ashraf, Nava, Oriana Bandiera, Edward Davenport, and Scott S. Lee. 2020. "Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the Delivery of Public Services." *American Economic Review* 110 (5): 1355–94.

Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack. 2014. "No Margin, No Mission? A Field Experiment on Incentives for Public Service Delivery." *Journal of Public Economics* 120 (17): 1–17.

Baker, George, Robert Gibbons, and Kevin J. Murphy. 1994. "Subjective Performance Measures in Optimal Incentive Contracts." *Quarterly Journal of Economics* 109 (4): 1125–56.

Banerjee, Abhijit V., Raghabendra Chattopadhyay, Esther Duflo, Daniel Keniston, and Nina Singh. 2012. "Can Institutions be Reformed from Within? Evidence from a Randomized Experiment with the Rajasthan Police." Unpublished.

Bertrand, Marianne, Robin Burgess, Arunish Chawla, and Guo Xu. 2020. "The Glittering Prizes: Career Incentives and Bureaucrat Performance." *Review of Economic Studies* 87 (2): 626–55.

Chevalier, Judith, and Glenn Ellison. 1999. "Career Concerns of Mutual Fund Managers." *Quarterly Journal of Economics* 114 (2): 389–432.

Deb, Joyee, Jin Li, and Arijit Mukherjee. 2016. "Relational Contracts with Subjective Peer Evaluations." *RAND Journal of Economics* 47 (1): 3–28.

de Janvry, Alain, Guojun He, Elisabeth Sadoulet, Shaoda Wang, Qiong Zhang. 2023. "Replication Data for: Subjective Performance Evaluation, Influence Activities, and Bureaucratic Work Behavior: Evidence from China." American Economic Association [Publisher], Inter-university Consortium for Political and Social Research [Distributor]. https://doi.org/10.3886/E 182787 V1.

Deserranno, Erika, and Gianmarco León Ciliotta. 2021. "Promotions and Productivity: The Role of Meritocracy and Pay Progression in the Public Sector." Unpublished.

[41] It is worth noting that, if there exists one supervisor that is systematically better at observing and assessing employee performance, and the organization can accurately identify this supervisor and commit to choosing him as the evaluator, then the masked scheme might not lead to better employee performance.

Ederer, Florian, Richard Holden, and Margaret Meyer. 2018. "Gaming and Strategic Opacity in Incentive Provision." *RAND Journal of Economics* 49 (4): 819–54.

Finan, Frederico, Benjamin A. Olken, and Rohini Pande. 2015. "The Personnel Economics of the State." NBER Working Paper 21825.

Fisman, Raymond, Jing Shi, Yongxiang Wang, and Weixing Wu. 2020. "Social Ties and the Selection of China's Political Elite." *American Economic Review* 110 (6): 1752–81.

Fisman, Raymond, and Yongxiang Wang. 2015. "The Mortality Cost of Political Connections." *Review of Economic Studies* 82 (4): 1346–82.

Gibbons, Robert, and Kevin J. Murphy. 1992. "Optimal Incentive Contracts in the Presence of Career Concerns: Theory and Evidence." *Journal of Political Economy* 100 (3): 468–505.

Gjesdal, Frøystein. 1982. "Information and Incentives: The Agency Information Problem." R*eview of Economic Studies* 49 (3): 373–90.

Grossman, Sanford J., and Oliver D. Hart. 1983. "Implicit Contracts Under Asymmetric Information." *Quarterly Journal of Economics* 98 (3): 123–56.

Hayes, Rachel M., and Scott Schaefer. 2000. "Implicit Contracts and the Explanatory Power of Top Executive Compensation for Future Performance." *RAND Journal of Economics* 31 (2): 273–93.

He, Guojun, Shaoda Wang, and Bing Zhang. 2020. "Watering down Environmental Regulation in China." *Quarterly Journal of Economics* 135 (4): 2135–85.

He, Guojun, and Shaoda Wang. 2017. "Do College Graduates Serving as Village Officials Help Rural China?" *American Economic Journal: Applied Economics* 9 (4): 186–215.

Jehiel, Philippe. 2015. "On Transparency in Organizations." *Review of Economic Studies* 82 (2): 736–61.

Lazear, Edward P. 2006. "Speeding, Terrorism, and Teaching to the Test." *Quarterly Journal of Economics* 121 (3): 1029–61.

Lazear, Edward, and Paul Oyer. 2012. "12 Personnel Economics." In *Handbook of Organizational Economics,* edited by Gibbons, Robert, and John Roberts, 479–519. Princeton, NJ: Princeton University Press.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76 (3): 1071–1102.

Li, Hongbin, and Li-An Zhou. 2005. "Political Turnover and Economic Performance: The Incentive Role of Personnel Control in China." *Journal of Public Economics* 89 (9-10): 1743–62.

Li, Weijia. 2019. "Meritocracy and Dual Leadership: Historical Evidence and an Interpretation." Unpublished.

Lü, Xiaobo, and Pierre F. Landry. 2014. "Show Me the Money: Interjurisdiction Political Competition and Fiscal Extraction in China." *American Political Science Review* 108 (3): 706–22.

MacLeod, W. Bentley. 2003. "Optimal Contracting with Subjective Evaluation." *American Economic Review* 93 (1): 216–40.

Maestri, Lucas. 2012. "Bonus Payments Versus Efficiency Wages in the Repeated Principal-Agent Model with Subjective Evaluations." *American Economic Journal: Microeconomics* 4 (3): 34–56.

Martinez-Bravo, Monica, Gerard Padró i Miquel, Nancy Qian, and Yang Yao. 2020. "The Rise and Fall of Local Elections in China." Unpublished.

Meyer, Margaret, Paul Milgrom, and John Roberts. 1992. "Organizational Prospects, Influence Costs, and Ownership Changes." *Journal of Economics and Management Strategy* 1 (1): 9–35.

Miles, Stephen A., and Michael D. Watkins. 2007. "The Leadership Team." *Harvard Business Review* 85 (4): 90–98.

Milgrom, Paul R. 1988. "Employment Contracts, Influence Activities, and Efficient Organization Design." *Journal of Political Economy* 96 (1): 42–60.

Milgrom, Paul, and John Roberts. 1988. "An Economic Approach to Influence Activities in Organizations." *American Journal of Sociology* 94: 154–79.

Olken, Benjamin, and Rohini Pande. 2013. *Governance Review Paper*. Cambridge, MA: Poverty Action Lab.

Oyer, Paul, and Scott Schaefer. 2011. "Personnel Economics: Hiring and Incentives." In *Handbook of Labor Economics*, Vol. 4, edited by David Card and Orley Ashenfelter, 1769–1823. Amsterdam: Elsevier.

Powell, Michael. 2015. "An Influence-Cost Model of Organizational Practices and Firm Boundaries." *Journal of Law, Economics, and Organization* 31 (1): 104–42.

Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature* 37 (1): 7–63.

Prendergast, Canice, and Robert H. Topel. 1996. "Favoritism in Organizations." *Journal of Political Economy* 104 (5): 958–78.

Rasul, Imran, and Daniel Rogger. 2018. "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service." *Economic Journal* 128 (608): 413–46.

Schaefer, Scott. 1998. "The Dependence of Pay-Performance Sensitivity on the Size of the Firm." *Review of Economics and Statistics* 80 (3): 436–43.

Serrato, Juan Carlos Suárez, Xiao Yu Wang, and Shuang Zhang. 2019. "The Limits of Meritocracy: Screening Bureaucrats Under Imperfect Verifiability." *Journal of Development Economics* 140: 223–41.

Shirk, Susan L. 1993. *The Political Logic of Economic Reform in China*, Vol. 24. Berkeley, CA: University of California Press.

Stiglitz, Joseph E. 1982. "Self-Selection and Pareto Efficient Taxation." *Journal of Public Economics* 17 (2): 213–40.

Wang, Shaoda, and David Y. Yang. 2021. "Policy Experimentation in China: The Political Economy of Policy Learning." NBER Working Paper 29402.

Williams, David, and Mary M. Scott. 2012. "Leadership Teams: Why Two Are Better Than One." *Harvard Business Review*, April 23. https://hbr.org/2012/04/leadership-teams-why-two-are-b.

Wu, Yanhui. 2017. "Authority, Incentives, and Performance: Evidence from a Chinese Newspaper." *Review of Economics and Statistics* 99 (1): 16–31.