THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Face-to-face communication in organizations

LSE Research Online URL for this paper: http://eprints.lse.ac.uk/106580/

Version: Accepted Version

Article:

Blanes I Vidal, Jordi ORCID: 0009-0002-9237-2049 (2020) Face-to-face communication in organizations. Review of Economic Studies. ISSN 0034-6527

https://doi.org/10.1093/restud/rdaa060

Reuse

Items deposited in LSE Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the LSE Research Online record for the item.

lseresearchonline@lse.ac.uk
https://eprints.lse.ac.uk/

Face-to-Face Communication in Organisations*

Diego Battiston†, Jordi Blanes i Vidal‡ and Tom Kirchmaier§

September 14, 2020

Abstract

Communication is integral to organisations and yet field evidence on the relation between communication and worker productivity remains scarce. We argue that a core role of communication is to transmit information that helps co-workers do their job better. We build a simple model in which workers choose the amount of communication by trading off this benefit against the time cost incurred by the sender, and use it to derive a set of empirical predictions. We then exploit a natural experiment in an organisation where problems arrive and must be sequentially dealt with by two workers. For exogenous reasons, the first worker can sometimes communicate face-to-face with their colleague. Consistently with the predictions of our model we find that: (a) the second worker works faster (at the cost of the first worker having less time to deal with incoming problems) when face-to-face communication is possible, (b) this effect is stronger when the second worker is busier and for homogenous and closely-located teams, and (c) the (career) incentives of workers determine how much they communicate with their colleagues. We also find that workers partially internalise social outcomes in their communication decisions. Our findings illustrate how workers in teams adjust the amount of mutual communication to its costs and benefits.

JEL *classification*: D23, M11.

Keywords: Teamwork, Face-to-Face Communication, Help in Organisations, Queueing Theory.

*We thank Ricardo Alonso, Tore Ellingsen, Miguel Espinosa, Mitch Hoffman, Luis Garicano, Rocco Macchiavello, Alan Manning, Marc Möller, Veronica Rappoport, Yona Rubinstein, Catherine Thomas, Chris Tyson and numerous seminar and conference participants for comments and suggestions that have helped improve the paper. Special thanks also to Chief Constable Ian Hopkins QPM, and to Steven Croft, Peter Langmead-Jones, Duncan Stokes, Ian Wiggett and many others at the Greater Manchester Police for making this project possible. No financial support was received for this paper. All errors remain our own.

†School of Economics, University of Edinburgh, Edinburgh EH8 9JT, UK; and Centre for Economic Performance, London School of Economics, London WC2A 2AE, UK. Email: diego.battiston@ed.ac.uk.

‡Corresponding author, Department of Management and Center for Economic Performance, London School of Economics, Houghton Street, London WC2A 2AE, UK; and Center for Economic and Policy Research, 33 Great Sutton Street, London EC1V 0DX, UK. Email: j.blanes-i-vidal@lse.ac.uk.

§Copenhagen Business School, Solbjerg Pl. 3, 2000 Frederiksberg, Denmark; and Centre for Economic Performance, London School of Economics, London WC2A 2AE, UK. Email: t.kirchmaier@lse.ac.uk.

1 Introduction

Most production activities require the collaboration of individuals (Arrow 1974, Simon 1979). For teams and organisations to function effectively, their members must communicate with each other, for instance, to exchange technical information or to coordinate decisions (Hayek, 1945). Yet such communication is often imperfect, as it requires time and effort and may be hampered by conflicting incentives (Garicano and Rayo, 2016). Therefore, a key decision in organisations involves choosing how much co-workers communicate with each other. While this issue is at the core of a large body of theoretical work, empirical evidence on communication in organisations has largely lagged behind.

This paper empirically studies the determinants and consequences of communication between co-workers in teams. We do this by taking advantage of an extremely rich dataset and a unique natural experiment in a large and complex public sector organisation. In our setting, individuals working in teams are always able to communicate electronically. Some teams, exogenously chosen by a computerised system allocating tasks to workers, can also communicate in person. Therefore, our experiment identifies the effects of being able to communicate face-to-face, in addition to electronically.

Our study makes two contributions. Firstly, we provide the first evidence on the positive causal link between the ability to communicate face-to-face and team productivity. More importantly, we provide evidence in support of a fundamental role of communication as a 'help' (Itoh, 1991) or 'information subsidy' (Hall and Deardorff, 2006) activity. Consistently with this mechanism, we show that communication improves the productivity of its receiver while generating a time cost on the part of its sender. We use a simple theoretical model to show that workers' observed behaviour is consistent with them understanding this trade-off and reacting to its costs and benefits in their communication decisions. For instance, we find that face-to-face communication increases when its potential sender has weaker incentives to maximise their own performance and can therefore afford to help their colleague. Similarly, we show that communication is also higher when its receiver needs more help, for instance, as a result of dealing with a more important problem from a social welfare perspective.

This Study We study the branch in charge of answering emergency calls and allocating officers to incidents in the Greater Manchester Police (GMP). To understand the role that face-to-face communication plays in our setting, we must briefly describe the production process. Each incoming call is answered by a *call handler*, who gathers the details and describes the incident in the internal computer system (see Figures 1 and 2). A *radio operator* then reads the description and allocates a police officer on the basis of incident characteristics and officer availability. These two workers have (partially) different objectives. Operators

are responsible for minimising the response time of their incidents (i.e. the time until an officer reaches the incident location). Handlers' main objective is to minimise the time that incoming calls spend waiting in the call queue, and they are therefore expected to be ready to take new calls as soon as they have finished a previous call.

Reading and understanding the incident's description takes the operator some time, which slows the speed of response. This processing time can however decrease if the operator can communicate face-to-face with the handler who created the incident. Unfortunately the handler's assistance is not costless, as communication takes time, during which the handler cannot be ready to receive new calls (i.e. must be in 'not ready' status). Understanding empirically how workers react to the trade-off between the benefit of communication (i.e. lower operator *response time*) and the cost (i.e. higher handler *not ready time*) is a core objective of the paper.

To fix ideas, we develop a queueing theory model where communication is subject to a trade-off: it helps its receiver but it is costly to its sender. Problems arrive and must be sequentially processed by two workers. If a worker is busy dealing with a problem, incoming problems accumulate in their queue. For each problem the first worker learns information that, if communicated, will allow the second worker to process the problem faster. The cost of communication is that it occupies the first worker's attention, thereby preventing them from dealing with new problems. We assume that the first worker chooses the amount of help provided, giving positive weight to the queueing and processing delays at the two levels.

The model generates several predictions regarding the comparison between a setting where communication is possible and another where it is not. Firstly, the second worker processes problems more quickly (and the first worker more slowly) when 'communication as help' is available. Secondly, these effects are larger when: (a) the second worker is busier on average, as a result of being inherently slow or receiving a high inflow of problems; (b) the first worker is less busy; and (c) the first worker assigns a higher weight to the objective of decreasing the second worker's delay. The model further predicts that an improvement in the efficiency of the communication technology (i.e. in the benefit to the second worker per unit of communication effort) decreases the second worker's delay, without necessarily increasing the first worker's.

We identify the effect of being able to communicate in person with the help of a natural experiment. In the GMP, handlers and operators are spread across four rooms in separate buildings. Each room contains the operators responsible for the surrounding neighbourhoods as well as a subset of the handlers, who can take calls from anywhere in Manchester. Our empirical strategy exploits the fact that the computerised queueing system matching incoming calls to newly available handlers creates exogenous variation in the co-location of handler and operator. As a result of this system, operators sometimes receive incidents cre-

ated by handlers located in the same room. For other (exogenously determined) incidents, the information will instead have been entered by handlers based in another location.

When co-located, the workers can engage in face-to-face communication, which comprises of several features. Firstly, it is two-way, allowing for the quick succession of questions and answers between the interlocutors. Secondly, it is oral. Thirdly, it is visual and in person. Given our setting we cannot disentangle the separate effect of these features, and it may be that an alternative technology (e.g. telephone) including only some of these features may achieve a similar result. Our main focus is studying when the workers in our setting will choose to use this additional communication channel at their disposal.

Results Our baseline results support the existence of a trade-off associated with face-to-face communication: operator response time is 2% lower and handler not ready time is 2.5% higher when the two workers are based in the same room. We confirm these findings by exploiting an organisation-wide relocation of workers and finding that *the same pair of workers* that used to work in the same room cease to be associated with differential productivities when not co-located. We also find that this reorganisation, which permanently separated handlers and operators, had the effect of increasing response time by 8% (around one minute, calculated at the median) for incidents classified as violent crimes.

The heterogeneity of the baseline effects (i.e. lower response time and higher not ready time under co-location) is broadly consistent with the comparative statics predictions of the theoretical model. Specifically, the effects are larger when the operator is an intrinsically slow worker or is suddenly very busy. Conversely, the effects are smaller when the handler is busy, proxied by the number of recently received incidents. Lastly, the effects are larger when handler and operator are the same gender, similar age and have a longer history of working together; characteristics that we would expect generate higher altruism on the handler towards the operator.

To test the remaining prediction on the role of the efficiency of the communication technology, we use the proximity within the room when handler and operator are co-located. Our expectation here is that teammates with neighbouring desks require less walking time to engage face-to-face, so each second spent by the handler should translate into more information transmission and therefore a higher benefit to the operator. Consistently with this prediction, we find that response time is faster even when *the same pair of workers* are located closer together inside the room.

We also find that the career incentives of handlers determine the help that they are willing to provide to their colleagues. Communication is higher when a handler has just been upgraded in pay, and has weaker incentives to minimise their not ready time. Along the same lines, communication decreases in the month of a handler's performance review meeting,

when we would expect handlers to be more concerned about their own performance. We regard these results as particularly interesting, as they illustrate how seemingly-unrelated institutional features of the organisation can affect the amount of help between co-workers. An immediate implication is that handlers do not fully internalise the welfare of the public ringing the police in their communication decisions.

While handlers may not be fully optimising social welfare, we show that they are still partially responsive to (proxies for) it. We find that types of incidents for which victim satisfaction depends more on response time, are associated with more communication. Similarly, we find that co-location leads to more face-to-face communication when the handler, upon picking up the phone, learns that there is a crime ongoing.

Overall, the pattern of heterogeneity indicates that handlers respond to the (private and social) costs and benefits of communication when determining how much face-to-face communication to engage in.

Related Work As Dewatripont (2006) and Garicano and Prat (2013) argue, theoretical work in economics has identified two main obstacles to the transmission of information in teams and organisations. On the one hand, conflicting incentives and incomplete contracting can make communication difficult or even impossible (Crawford and Sobel 1982, Prendergast 1993, Dessein 2002, Garicano and Santos 2004, Alonso et al. 2008, Rantakari 2008 and Friebel and Raith 2010). Another body of work instead assumes away incentive problems and posits that communication is directly costly, for instance because it uses the agents' valuable time (Radner 1993, Bolton and Dewatripont 1994, Van Zandt 1999, Garicano 2000, Dessein and Santos 2006, Dessein et al. 2016). While the details of the models in this second literature vary significantly, a common feature is the trade-off between the benefit in terms of better or faster decision-making, and the (time) cost incurred to communicate. The mechanism that we highlight in this paper belongs to this second class of models and, to the best of our knowledge, ours is the first paper to provide empirical support for this broad trade-off. However, we also provide evidence on the role of incentives, in showing that the career concerns of workers determine the amount of communication they are willing to engage in.

Field evidence in economics on workplace communication is relatively scant, and often does not identify effects on productivity (Gant et al. 2002, Palacios-Huerta and Prat 2012, Bloom et al. 2014). A recent exception is Menzel (2019), who implements an experiment to encourage workers to share their knowledge and measures the resulting benefits. However, he does not explore the cost-benefit trade-off, and workers' reaction to it, which are the focus of this paper.[1]

[1] A related literature studies how the patterns of scientific collaboration depend on the ability to commu-

Outside economics, the study of organisational communication reflects contributions from psychology, sociology, and operations research (Jablin et al. 1987, Pace and Faules 1994, Harris 2002). Broadly speaking, two separate paradigms dominate this literature. The earliest is the 'engineering-centric' view (Shannon and Weaver 1949, Kmetz 1998), which treats organisations as information-processing systems, and internal communication as mechanical information transmission between linked processors. Later, the 'transactional' view (Barnlund, 1970) contributed the notion that communication occurs between humans, and therefore is affected by variables such as its context, the medium used, and the relation between sender and receiver. Our paper combines aspects of these two views: we model workers as information-processing nodes but empirically acknowledge that the efficiency of their communication may depend on their incentives, their history together, or the match in their demographics.

The finding that face-to-face communication allows co-workers to help each other links the paper with the wider literature on teamwork (Gaynor et al. 2004, Chan 2016) and, more narrowly, with field studies on employee cooperation (Itoh 1991, Drago and Garvey 1998, Hamilton et al. 2003, Berger et al. 2011). This latter work is mostly concerned with the role of incentives and is typically silent about the actual mechanism through which this cooperation occurs. We contribute to this work both by highlighting interpersonal communication as a leading mechanism and by identifying determinants additional to the incentive structure.

Plan The paper is organised as follows. Section 2 describes the institutional setting. Section 3 presents and solves the theoretical model. Section 4 outlines its empirical predictions as adapted to our institutional setting. In Section 5 we present the data and the main empirical strategy. In Section 6 we present and interpret the baseline results. In Section 7 we examine the heterogeneity of these findings. Section 8 discusses implications for social welfare. Section 9 concludes.

2 Institutional Setting

Organisation and Production The Operational Communications Branch (OCB) is the unit in charge of answering police 999 calls and allocating officers to the corresponding incidents. We focus on the team consisting of its two primary workers: call handlers and radio operators. Figures 1 and 2 visualise the production process.

nicate remotely (Agrawal and Goldfarb 2008, Forman and van Zeebroeck 2012) or in person (Catalini 2017, Catalini et al. 2018). The closest paper here is Catalini (2017), who uses the relocation of departments in a French university to analyse how search costs, monitoring costs and the associated research productivity vary with physical proximity between academics.

< INCLUDE Figure 1 and 2 here>

Incoming calls are allocated to handlers using a first-come-first-served system, matching the call at the front of the queue with the next handler that becomes available. The handler questions the caller, chooses the opening code (i.e. the 'type' of incident) and the grade (a coarsely defined degree of urgency), describes the incident in its log and ticks a box to officially create the incident. This information is recorded by the software GMPICS. The handler then indicates their status as 'not ready' or instead 'ready' to receive new calls. If 'ready', a new call can arrive at any point and must be immediately answered by the handler.

When an incident is created, it immediately appears on the GMPICS screen of the operator overseeing the subdivision where the incident occurred. The operator processes the information in the log and allocates a police officer, who attends the incident scene. The allocation of incidents to operators is deterministic, since at any point in time there is only a single operator in charge of a specific subdivision. Therefore, handlers do not decide to which operator they assign an incident (they can observe the operator's ID number in GMPICS).

Face-to-Face Communication as 'Help' from Handler to Operator Reading the log and gathering the information necessary to allocate an officer takes time. This time can be decreased if the operator is able to interact face-to-face with the handler, for two reasons (see Appendix A for an extended discussion). Firstly, information is sometimes unclear in the log. Operators have several channels through which they can clarify doubts or gather additional details, but conversing with the handler is a fast and efficient way to fill information gaps.[2]

Secondly, the logs sometimes contain *too much* information, rather than too little. For obvious reasons, handlers typically record many more details than are needed for allocation purposes. This implies that operators have to sift through the log and extract the specific elements guiding the allocation of officers. This challenge is compounded by the fact that the information is often not structured optimally (from the operator's perspective). Operators often take less time asking the handler to concisely provide the important details, than extracting these details themselves from the log.

Assisting operators is not costless for handlers, as it implies that they must be 'not ready' to answer incoming calls.[3]

[2]Alternative channels include conducting targeted searches on individuals or addresses in the GMP databases, and contacting the caller directly. In addition, operators can electronically message handlers. The availability of electronic messages enabling real-time Q&A implies that our study may be identifying features of face-to-face communication that are absent in real-time electronic messaging systems, such as oral communication and the presence of visual cues.

[3]A handler can set 'not ready' anticipating that there may be questions or switch from 'ready' to 'not ready' when approached by an operator. However, handlers cannot interrupt an ongoing call to discuss an earlier case with the operator.

The Natural Experiment From November 2009 to January 2012, OCB staff were spread across four buildings or 'rooms', in different parts of Manchester: Claytonbrook, Leigh, Tameside and Trafford. Every room accommodated the operators overseeing the surrounding subdivisions (Figure 3 displays the areas overseen from each location). Handlers were also dispersed across (and uniquely linked to) the four locations.

< INCLUDE Figure 3 here>

As discussed earlier: (a) incoming calls were deterministically matched with the operators in charge of the subdivisions from which they originated, and (b) the first-come-first-served queueing system exogenously matched incoming calls to available handlers. This meant that, for exogenous reasons, operators would sometimes be reading the descriptions of incidents created by same room handlers, while on other occasions the handlers were based in a different part of Manchester. As we argue in Section 5, this exogenous variation provides the foundation for the main empirical strategy of the paper.

In January 2012, a major reorganisation of the OCB reassigned all handlers to a single location (Trafford), while radio operators were divided between Claytonbrook and Tameside. This put an end to the natural experiment that we study here.

Workers and Performance Indicators While there are no educational prerequisites to work at the OCB, a high school diploma is in practice necessary to be a succesful applicant. Once selected, workers undergo intensive training programs. Salaries for handlers are slightly below the median Manchester salary and turnover is quite low (around 5% annually).

The GMP has a small number of key performance indicators (Appendix Figure A1 displays them in a recent Annual Report). The most important are: (a) the *allocation time* of incidents (the time between their creation by the handler and the allocation of an officer), (b) the *response time* (the time between creation and the officer reaching the incident's scene), and (c) the *call queuing time* (the average time that incoming calls spend in the queue before being answered). These measures are critical to the GMP (see Appendix Figure A2), for two main reasons. Firstly, nation-wide numerical targets were introduced by the UK Home Office in 2008. For instance, the target for call queuing times was for 90% of calls to be answered within ten seconds.[4] Secondly, these measures are important

[4]For Grade 1 incidents, the targets were for a maximum of 2 and 15 minutes for allocation and response time, respectively. The equivalent targets for Grade 2 (respectively Grade 3) were 20 and 60 minutes (respectively 120 and 240 minutes). While these targets were nominally scrapped in June 2010, police forces continued to regard them as objectives and to believe that they were being informally evaluated on their basis (Curtis, 2015). Furthermore, information on response times continued to be discussed in the reports produced by the HMIC (the central body that in the UK regulates and monitors police forces). For an example, see HMIC (2012). They were also discussed in the reports by the GMP to the Manchester City Council Citizenship and Inclusion Overview and Scrutiny Committee.

determinants of public satisfaction. For instance, UK-wide survey evidence suggests that response time is one of the most important variables predicting citizens' satisfaction with the police forces (Dodd and Simmons, 2002/03).[5]

Operators are held responsible for the allocation and response times of the incidents that they personally deal with, while handlers are held responsible for the call queueing time. Handlers' responsibility is a joint one, as they all take calls from a common queue. As is the case with other public sector organisations (Burgess and Rato, 2003), there is no performance pay providing explicit incentives to handlers. However, handlers do have career incentives to contribute to the reduction of call queueing time. Specifically, handlers can be moved to a higher pay grade (while continuing to perform the same job) and they can be promoted to other jobs of higher status and salary (specifically radio operator and handler supervisor). Career progression depends partly on a handler's supervisor evaluation, which takes place annually following their performance review meeting. An important ingredient in this evaluation is a handler's average 'not ready' time, as it is deemed that handlers being too often 'not ready' are not contributing to the group objective of reducing the call queueing time.

In Table 1 we provide suggestive evidence on the importance of 'not ready' time for handlers' career progression. Specifically, we regress a handler's promotion or upgrade in a year on a set of lagged performance indicators, controlling for handler and year fixed effects. We find that a higher value of average 'not ready' time decreases a handler's likelihood of being promoted or upgraded the following year. Note on the other hand that the coefficient for average response time, which is not handlers' direct responsibility, is much smaller and statistically insignificant. This evidence supports our claim that handlers are incentivised to not spend too much time 'not ready', and that they are held responsible for incoming calls but not for incidents after they have been created.

< INCLUDE Table 1 here>

[5]While important, these measures do not directly capture the 'quality' of the GMP dealing with an incident. For instance, they do not reflect whether the appropriate officer was allocated, or whether the attending officer was in possession of all the relevant information prior to arrival. Measuring every dimension of quality is of course difficult. Nevertheless, in Section 8 we replicate our baseline regressions using additional dependent variables such as whether the incident escalated to becoming a crime, and, if so, whether the crime was cleared. A superior 'quality' measure of performance is whether the crime victim was satisfied with the police response. Unfortunately, we are unable to use this measure as additional dependent variable in the main analysis of the paper, as the number of survey responses is very low and it mostly falls outside our baseline sample period.

3 A Model of Communication in Team Production

In this section we develop a simple model to capture the idea that communication can help its receiver work faster, but has a time cost for its sender.[6]

Organisation and Production An organisation consists of two levels and n_i workers in level $i = 1, 2$. The organisation receives a flow of problems and each problem must be processed as quickly as possible, first in level 1 and then in level 2. An example of this production process would be that of architects and engineers, who provide different inputs, typically sequentially, in construction projects. More generally, knowledge workers with different specialisations must often work together in teams and provide their inputs in a sequential manner.

Problems are ex ante homogenous and arrive following a Poisson process with rate λ. An incoming problem is immediately and randomly allocated to one of the $n_1 \times n_2$ potential pairs of workers, who will comprise the team dealing with that specific problem.

Workers can only process one problem at a time. The baseline processing time at level i follows the same distribution for all problems, which we assume to be exponential with rate θ_i.[7] Conditional on the two distributions, the realised processing times of a problem at the two levels are independent of each other. If a problem reaches a worker's desk and the worker is free, they start processing it immediately. If the worker is instead occupied with an earlier problem, we assume that the incoming problem joins the back of that worker's queue, which has an infinite number of positions.

Communication as Help We model communication as a 'help' (Itoh, 1991) or 'information subsidy' (Hall and Deardorff, 2006) activity, which reduces the processing time of the receiver, but at the cost of increasing it for the sender. Communication helps because, in dealing with an incoming problem, the level 1 worker learns information that is valuable to their level 2 teammate. For instance, an architect may learn features of a project that are not embedded in its blueprint but are useful information to the project engineer. The level 2 teammate may then spend time gathering that information themselves or, alternatively

[6] We discuss the assumptions of the model and the relation with existing theoretical literature in Appendix B. Many organisations could be described as 'networks of queues' (Beggs 2001, Arenas et al. 2010), but tractability is a key challenge in the study of these organisations because of the relative scarcity of formal results in queueing theory (Cooper, 1981). From a technical perspective, the main contribution of the model is in providing a tractable framework to study the optimal allocation of processing resources for these organisations. Our framework could be used to study questions of organisational structure (e.g. the allocation of a fixed set of processors across levels of the organisation) and task delegation (e.g. the routing of problems across workers).

[7] We assume that $\theta_i > \lambda/n_i \; \forall \; i$. This ensures that problems never arrive faster than they can be processed, and that the queue does not grow infinitely.

and more efficiently, learn about it from the level 1 teammate.

Formally, the level 1 processing rate of a problem is $\theta_1 - x$, where $x \geq 0$ is the communication effort.[8] Because processing times follow exponential distributions, the expected processing time is then $(\theta_1 - x)^{-1}$. A higher x increases processing time because it takes time to communicate, which detracts from the level 1 task.[9]

The expected level 2 processing time is $(\theta_2 + \pi x)^{-1}$, where $\pi > 0$ is a parameter capturing the 'efficiency' of the communication technology. An increase in π decreases the expected processing time at level 2, for a given communication effort.[10] Intuitively, the same time spent on communication by the level 1 teammate leads to more learning about the problem by the level 2 teammate.

Incentive to Help We assume that the communication effort is exclusively determined by the level 1 worker. This is a reasonable assumption because this is the teammate providing help to their colleague. Define D_i as the time that it takes to deal with the average problem in level i, where D_i includes both the average processing time and the average time that problems spend waiting in the queue. We assume that level 1 workers choose x to minimise the weighted sum of the delays at the two levels, $D_1 + \omega D_2$.

The parameter ω captures how much level 1 workers internalise the level 2 delay, and can therefore be interpreted as their 'incentive to help'. While we do not microfound it, we would expect ω to reflect four different elements: (a) how much the worker cares about improving their own performance (which may be tied to the average length of the level 1 queue), (b) how much the worker cares about their colleague (and therefor the level 2 queue), (c) how much the worker internalises the performance of the organisation and how this performance depends on the two queues, and (d) how the two queues affect the customers being served by the organisation, and how much the level 1 worker internalises their welfare. For instance, an organisation in which workers are strongly encouraged to maximise their own performance should be associated with a lower ω (Itoh, 1991). Conversely, ω should be higher when the level 2 task is more important to the organisation (provided of course that

[8] We use the label 'effort' here in the sense of Dewatripont and Tirole (2005). As there, the sender makes an effort to communicate, and the amount of information reaching the receiver depends also on other parameters of the model. The cost of this effort, which is not microfounded in Dewatripont and Tirole (2005), arises in our model because higher communication effort prevents the level 1 worker from processing their own problems.

[9] The model is agnostic as to what activities are included in the baseline processing times θ_1^{-1} and θ_2^{-1}. The level 1 activity could include, for instance, some compulsory but basic amount of communication with level 2. x could then refer to the choice to provide *additional* communication. In our empirical setting, handlers always spend some time communicating electronically and that can be regarded as included in θ^{-1}. We then study the consequences of *additional* face-to-face communication effort x.

[10] In the absence of communication, processing time at level 2 is θ_2^{-1}. The time saved by communication is $\pi x/\theta_2(\theta_2 + \pi x)$. We assume that this time saved is already net of the time that the receiver spends listening to the communication from level 1.

the level 1 worker internalises this importance).

Solving the Model Problems arrive with rate λ and are assigned to one out of $n_1 \times n_2$ potential teams, which implies that each worker at level i receives a flow with rate $\frac{\lambda}{n_i}$. We have also seen that problems are processed at rates $\theta_1 - x$ and $\theta_2 + \pi x$, respectively. Standard results in queueing theory imply that we can study this organisation as a collection of independent M/M/1 queues.[11] It can easily be shown that the average delays at the two levels are $\frac{1}{\theta_1 - x - \lambda/n_1}$ and $\frac{1}{\theta_2 + \pi x - \lambda/n_2}$, respectively (see for instance Section 3.4 in Cooper, 1981). We assume that there is a single x that is chosen once and then fixed for the whole organisation.[12] Given the above average delays, we can write the (weighted) delay, which level 1 workers minimise, as

$$D = D_1 + \omega D_2 = \frac{1}{s_1 - x} + \frac{\omega}{s_2 + \pi x} \tag{1}$$

with $s_i = \theta_i - \lambda/n_i$. Solving for the optimal communication effort from this problem gives (all proofs are in Appendix C):

Lemma 1 *The optimal level of communication effort is*

$$x^* = \begin{cases} 0 & \text{if } \sqrt{\omega\pi} s_1 \leq s_2 \\ \dfrac{\sqrt{\omega\pi} s_1 - s_2}{\pi + \sqrt{\omega\pi}} & \text{otherwise} \end{cases} \tag{2}$$

Lemma 1 provides several insights regarding the optimal communication effort. Firstly, note that s_i is the difference in the rates at which problems arrive (i.e. λ/n_i) and are processed (i.e. θ_i) at level i, in the absence of communication. A high s_i indicates that problems are processed very fast relative to their arrival rate, the level i queues are often empty, and the level i workers enjoy a lot of free time. Therefore, we can interpret s_i as the

[11] In queueing theory, an M/M/1 queue is a queue with one server/worker, an arrival process that is Poisson (Markov) and a processing time distribution that is exponential (Markov). A characteristic of Poisson processes is that when they split or merge, the resulting processes are also Poisson. This property, together with the random formation of teams and the fact that level 1 workers are only fed from outside the organisation, leads to the M/M/1 nature of the queue for level 1 workers (Jackson, 1957). Burke (1956) shows that the output process of an M/M/1 queue is also Poisson, and that the number of problems in the queue at time t is independent of the departure process before that date. It follows from these results and from the Poisson nature of merged Poisson processes that the arrival process for each of the level 2 workers in our organisation is also Poisson.

[12] This is a relatively strong assumption as, if allowed, workers would want to make x contingent on the state of the level 1 queue at a point in time. Specifically, they would want to communicate more when the level 1 queue is empty and less when it is full. This contingent-communication model would be substantially more complicated to solve, but the insights from Propositions 1 and 2 would be likely also present there.

'slack' enjoyed by workers at level i.[13] Equation (2) shows that x^* is higher when the level 1 enjoys more slack, relative to level 2.

Secondly, note that the level 1 worker may decide to not communicate at all, for instance if they enjoy very little slack (low s_1) or have little interest in helping their colleagues (low ω).

Lastly, differentiating (2) with respect to π we have

$$\frac{\partial x^*}{\partial \pi} = \frac{\sqrt{\frac{\omega}{4\pi}} s_1 - (1 + \sqrt{\frac{\omega}{4\pi}}) x^*}{\pi + \sqrt{\omega \pi}} \gtrless 0 \tag{3}$$

Equation (3) shows that an improvement in the communication technology need *not* result in more communication effort when x^* is already quite high. To understand the reason, we separately differentiate the two terms in (1) to plot the marginal cost and the marginal benefit of communication in Figure 4. We find that an increase in π makes the marginal benefit curve *rotate clockwise*, as it has two effects on the marginal benefit of communication: a *productivity* effect and a *stock* effect. On the one hand, a higher π makes additional communication effort relatively more efficient, which increases its marginal benefit. On the other hand, an increase in π also makes the existing stock of communication effort translate into a lower delay at level 2, which decreases the marginal benefit of reducing it further. When the existing equilibrium x^* is higher than s_2/π, the stock effect dominates, and hence the clockwise pivot of the curve. Intuitively, when communication is already very high (and the delay at level 2 very low), an increase in its effectiveness is optimally used in part to decrease the delay at level 1, which requires a decrease in the communication effort x.

< INCLUDE Figure 4 here>

Comparative Statics We now derive comparative statics regarding the variables proxied for in the empirical section. Firstly, consider the level 1 worker. The average processing time when the level 1 worker can help the level 2 worker is $(\theta_1 - x^*)^{-1}$. If the level 1 worker was not able to help through communication (i.e. if a lack of access to a communication technology forced $x = 0$), processing time would instead be θ_1^{-1}. We define

$$\Delta P_1 = \frac{1}{\theta_1 - x^*} - \frac{1}{\theta_1} = \frac{x^*}{\theta_1(\theta_1 - x^*)} \tag{4}$$

as the level 1 increase in processing time when communication is possible.

Now consider the level 2 worker. In our empirical setting we do not directly observe the processing time, but can instead measure the overall delay between a problem reaching the

[13] Not coincidentally, the metric $\lambda/\theta_i n_i$ denotes in queuing theory the 'utilisation factor' or 'traffic intensity' for workers in level i (i.e. the expected fraction of time that they are busy). See again Section 3.4 in Cooper (1981).

level 2 worker's desk and the problem being dealt with. The delay in level 2 was $(s_2+\pi x^*)^{-1}$, whereas it would be s_2^{-1} in an organisation where communication was not possible. We define

$$\Delta D_2 = \frac{1}{s_2 + \pi x^*} - \frac{1}{s_2} = \frac{-\pi x^*}{(s_2 + \pi x^*)s_2} \tag{5}$$

as the reduction in the level 2 delay when the two members of the team are able to communicate with each other. Proposition 1 follows immediately.

Proposition 1 *In an organisation where communication is possible (relative to one where it is not possible),*

1. *the processing time at level 1 is higher (i.e. $\Delta P_1 \geq 0$), and*
2. *the overall delay at level 2 is lower (i.e. $\Delta D_2 \leq 0$).*

Proposition 1 represents both the trade-off at the core of the theoretical model and its main empirical prediction. The next proposition studies how changes in the parameters of the model affect ΔP_1 and ΔD_2.

Proposition 2 *Compare an organisation where communication is possible with another organisation where it is not. Then,*

1. *an increase in the level 1 worker incentive to help, ω, leads to a greater decrease in the delay at level 2 (i.e. $\frac{\partial \Delta D_2}{\partial \omega} \leq 0$) and a greater increase in the processing time at level 1 (i.e. $\frac{\partial \Delta P_1}{\partial \omega} \geq 0$).*
2. *an improvement in the communication technology, π, leads to a greater decrease in the delay at level 2 (i.e. $\frac{\partial \Delta D_2}{\partial \pi} \leq 0$), and has an ambiguous effect on the processing time at level 1 (i.e. $\frac{\partial \Delta P_1}{\partial \pi} \gtrless 0$).*
3. *an increase in the slack at level 1, s_1, leads to a greater decrease in the delay at level 2 (i.e. $\frac{\partial \Delta D_2}{\partial s_1} \leq 0$) and a greater increase in the processing time at level 1 (i.e. $\frac{\partial \Delta P_1}{\partial s_1} \geq 0$). An increase in the slack at level 2 has the exact opposite effects (i.e. $\frac{\partial \Delta D_2}{\partial s_2} \geq 0$ and $\frac{\partial \Delta P_1}{\partial s_2} \leq 0$).*

We formally show these results in Appendix C and discuss them here. The intuition of Part 1 is immediate. When the level 1 worker internalises the level 2 delay more, they will provide more help, at the expense of their own processing time.

Now consider Part 2. We have established in (3) that an increase in π may or may not increase communication effort. The ambiguous effect of π on ΔP_1 then follows from the fact that ΔP_1 depends on π only through x^*. To understand the positive effect of π on ΔD_2, consider the level 1 worker's objective as that of maximising production (i.e. minimising

delay) across the two tasks. An increase in π represents an absolute improvement in the technology of production that also makes level 2 production more efficient relative to level 1 production. Depending on parameter values, the level 1 worker may use the improvement to increase level 2 production and decrease level 1 production, or instead to produce more of the two tasks. Because the level 2 task is now more efficient in both absolute and relative terms, it is never worth it to decrease production at level 2.

Part 3 is also straightforward in its intuition. When the level 1 worker enjoys a lot of slack, helping their colleague is not very costly, as this time is typically taken away from waiting for a new problem as opposed to picking the first problem from a long queue. Being able to communicate then translates into a lower delay at level 2, and a higher level 1 processing time. The opposite result follows when the level 2 worker enjoys a lot of slack.

4 Empirical Predictions

The model above is agnostic about the technology of communication available to the organisation. In our empirical setting, we have empirical variation in the co-location of handlers and operators, and therefore in their ability to communicate face-to-face. Therefore, face-to-face communication is the empirical focus of this paper. In this section we discuss the conversion of the insights from Propositions 1 and 2 into empirical predictions.

The Effect of Being Able to Communicate Face-to-Face Proposition 1 predicts that D_2 will be lower and P_1 will be higher when communication is possible. The most natural empirical counterpart of D_2 in the OCB is the allocation time of an incident, as this captures the time elapsed between the incident appearing in the operator's computer and their allocation of an officer. The response time is a complementary measure which extends until the officer reaches the incident's scene.

We proxy ΔP_1 (the additional handler's processing time under co-location) as follows. As discussed in Section 2, talking with the operator absorbs the handler's time and attention, and is typically incompatible with being available to receive new calls. Following the creation of co-located incidents handlers should therefore be more unavailable to take new calls. Our proxy for ΔP_1 is then the time spent by the handler 'not ready' to take new calls following an incident's creation (or, more specifically, the difference across co-located and non-co-located incidents in not ready time).

Proposition 1 indicates that access to a communication technology should translate into the first worker processing more slowly (i.e. higher not ready time), and the second worker producing with a lower delay (i.e. lower allocation time). This is the core trade-off in the model and the main prediction in the paper. We empirically examine this prediction

in Table 3.

The Handler's Incentive to Help Part 1 of Proposition 2 formalises the intuition that communication will increase when the level 1 worker is more eager to help their colleague. We use two sets of proxies to capture the handler's incentive to help the operator.

Firstly, consider career incentives, which we examine in Table 7. In Table 1 we found that handlers' likelihood of being upgraded in pay depends negatively on their not ready time. We posit that recently upgraded handlers may become less focused on minimising their not ready time, and more willing to help their colleagues. The reason is that there are only two pay grades, and while handlers in the highest grade can be promoted to operator or supervisor, this happens only rarely and never shortly after a pay upgrade. Because of this, the period immediately following the pay upgrade can be regarded as one of weak career incentives, and we predict that communication will be higher during this period.

The second proxy for career incentives is based on the fact that handlers are evaluated annually by their supervisors, on the basis of indicators such as average not ready time. This evaluation follows a performance review meeting between handler and supervisor, which takes place in different months of the year for different handlers. We posit that handlers may become more focused on their own performance, and less willing to help operators, in the month of their performance review meeting.

The second set of proxies for the handler's incentive to help is based on the idea that handlers may be more motivated to spend time assisting operators who they share more in common with. We proxy for this common experience with: (a) whether the workers are of the same gender, (b) whether the workers are of similar age, and (c) whether the teammates have worked together often in the past. We posit that in teams with these characteristics, handlers may more strongly internalise the operator's responsibility of reducing response time and we examine these predictions in Table 8.

The Efficiency of the Communication Technology Part 2 of Proposition 2 predicts that the operator's productivity improvement when communication is possible, ΔD_2, should be increasing in the efficiency of the communication technology π. Our proxy for π is the physical distance between the desks of two co-located workers. Even when workers are co-located, face-to-face communication requires walking across the room and standing alongside each other. Two teammates with neighbouring desks will obviously require less walking time. As a result, the additional time spent by the first worker (before being ready to take a new call) will be devoted to more actual information transmission when the two desks are closer, and should lead to a lower delay for the second worker. We provide evidence on this prediction in Table 9.

The Operator's Slack In our stylised theoretical model, a single communication effort is chosen once and then remains constant over the long term. However, in most organisations the amount of slack that workers enjoy will vary, for instance across teams or over time, thus generating variation that we can use to evaluate whether the empirical patterns are consistent with the intuition of Part 3 of Proposition 2.

Operators are uniquely responsible for all incidents occurring in their specific subdivision. Our first measure of an operator's slack is therefore the number of incidents in the operator's subdivision during the hour prior to an incident's arrival. We use the innate (i.e. average) operator speediness in the allocation of incidents as an additional complementary measure. Our expectations here are that operators that are inherently slower and have recently received a large inflow of incidents, should enjoy less slack. They should therefore benefit more (i.e. lower allocation time but higher handler not ready time) from being able to communicate with the incident's handler. This evidence is presented in Table 10.

The Handler's Slack Part 3 of Proposition 2 predicts that ΔD_2 will decrease (and ΔP_1 will increase) in the slack of the level 1 worker, s_1. In choosing a good proxy for s_1 we have to take into account an important difference between the empirical setting and the theoretical framework. Unlike in the model, handlers in the OCB take calls from the same *common* queue rather than from their own individual queues.[14] Therefore, our measure of a handler's slack when receiving an incident cannot be individual, but must instead be computed at the organisational level.

We use the number of calls per on-duty handler received during the previous half hour, as a lower value of this measure should increase handlers' slack and increase the effect of co-location on response time and not ready time.[15]

5 Empirical Strategy

Dataset Our baseline dataset contains every incident reported through the phone to the GMP between November 2009 and December 2011.[16] We restrict our attention to incidents

[14]The model assumption that the level 1 workers process incoming problems from their own individual queues allowed us to keep the model tractable. When all level 1 workers share a common queue, the mathematical expressions become difficult to work with. In addition, Burke's (1956) result on the Poisson nature of an M/M/1 output process does not immediately generalise to M/M/c queues (with many servers/workers). However, the intuition from the current model should extend easily to one in which all level 1 workers share the same queue.

[15]Because most calls are quite short and there is a large number of calls and handlers, the reference time period is shorter here than for the operators (i.e. half hour rather than full hour). We find qualitatively similar results when using a full hour period.

[16]We include calls both to the 999 and 101 lines. The 999 line is theoretically reserved for incidents which the caller perceives to be urgent and serious, while the 101 line is meant for less serious incidents. In the

where the handler allocated the call a grade below or equal to three, therefore transferring responsibility to a operator rather than to a divisional commander. For every incident we observe, the allocation and response time, the location of the incident, the grade and (horizontal) opening code, the identity of the handler and operator, and the desk position from which the handler took the call. An exception is the not ready time variable, which is only observed starting on the second semester of 2010 and is then (for exogenous reasons) further missing for some months. The dataset was made available to us under a strict confidentiality agreement.

Table 2 provides basic summary statistics for the main variables in our study. Note that our sample size is very large, as it includes close to one million incidents. In around one in four observations the handler and operator are in the same room. The outcome variables are highly right-skewed. For response, for instance, the median response time is 20 minutes, while the average time is more than four times larger.[17] The average (median) not ready time is 1.15 minutes (25 seconds) and we find that handlers spend some time in not ready status after 30% of the calls.

< INCLUDE Table 2 here>

Intuition of Empirical Strategy The allocation of calls to handlers works as follows. As calls come in, they join the back of the call queue, and the system matches the call at the front of the queue with the next handler that becomes available. If the call queue is empty and several handlers start to become available, they form their own queue, and the handler at the front of the handler queue is matched with the next incoming call. This matching system implies that, conditionally on the hour on which calls arrive, the characteristics of the resulting incidents are plausibly orthogonal to the handlers matched with these incidents.

The main independent variable in this paper is $SameRoom$, which takes value one when handler and operator are co-located. The queuing system outlined above makes $SameRoom$ also plausibly orthogonal to incident and worker characteristics, subject to an important caveat. Some rooms (for instance Trafford) are bigger than others (e.g. Leigh) and therefore contain a larger number of handlers. This implies that the likelihood of $SameRoom = 1$ will be mechanically higher if the call originates in a Trafford neighbourhood, relative to a Leigh neighbourhood. Calls originating from Trafford and Leigh may also have different characteristics, which could independently affect their average allocation and response times. Therefore, our claim regarding the exogeneity of the variable $SameRoom$ is only conditional

empirical analysis we include both types of calls because many calls that turn out to be urgent arrive through the 101 route, and vice versa. In every regression we control for the call source (999 versus 101).

[17] The maximum value is almost 15 days, likely the result of some error in the classification of the incident. The fact that the left hand side variables in our regressions are in logarithmic form should dampen the effect of outlying observations.

on (handler and operator) room fixed effects, in addition to hour (i.e. year X month X day X hour) fixed effects.

After introducing the above controls, the identification assumption on the exogeneity of the variable *SameRoom* would only be threatened if, for instance, urgent incidents were more likely to be matched with co-located handlers. A truly first-come-first-served queueing system does not take into account the characteristics of the calls or handlers and therefore prevents this type of non-random matching from occuring.

Estimating Equation Our baseline estimating equation is:

$$y_i = \beta SameRoom_{j(i)k(i)} + \theta_{t(i)} + \lambda_{j(i)} + \mu_{k(i)} + \pi_{g(i)} + \gamma_{h(i)} + \mathbf{X_i} + \epsilon_i \tag{6}$$

where y_i is an outcome measure (i.e. allocation time, response time, not ready time following the incident) for incident i. Throughout our paper, outcome measures are measured in log form, both for ease of interpretation of the coefficients and in the presence of right-skewness to minimise the effect of outlying observations. Consistently with our earlier discussion, we control for $\theta_{t(i)}$ (the fixed effect for the hour t at which the incident arrived) and $\lambda_{j(i)}$ and $\mu_{k(i)}$ (the fixed effects for the rooms j and k from which the incident was handled and dispatched). Our main independent variable of interest is the dummy $SameRoom_{j(i)k(i)}$, which takes value 1 when rooms j and k coincide.

In our baseline specification we also control for $\pi_{g(i)}$ and $\gamma_{h(i)}$ (the fixed effects for individual handler g and operator h assigned to the incident) and by other incident characteristics (such as the assigned grade) included in the vector $\mathbf{X_i}$. These latter controls are not essential for identification, but contribute to the reduction of the standard errors. We cluster these standard errors at the subdivision and year/month level.[18]

To examine the balance of characteristics across the co-location of handler and operator,

[18] As discussed in Abadie et al. (2017) and Cameron and Miller (2015), the decision on the adequate clustering level is not straightforward and typically requires additional information about the design of the study or the institutional setting. Our decision to cluster standard errors at the subdivision and year/month level follows from the expectation that serial correlation may be strongest for incidents occurring close both geographically and in time. For instance, unobserved temporary shocks such as traffic congestion could lead to this correlation. In addition, such a correlation may arise from the fact that operators are geographically specialised at the subdivision level. An operator being particularly slow on a specific day may affect all incidents dealt with on that day. Our clustering strategy further allows for correlation across incidents within the same month. The numbers of clusters in the baseline regression are 1,408 (allocation and response) and 832 (not ready), respectively. In Appendix Table A1 we investigate the robustness of the baseline standard errors to allowing for alternative clustering strategies, such as the handler/operator pair, subdivision, date, year/month, year/week, handler, operator, and year/month/operator. We further allow for multiway clustering (Cameron et al., 2011) across several of these dimensions. We find the standard errors to be remarkably robust to the choice of clusters.

we run the regression:

$$SameRoom_{j(i)k(i)} = \sum_{s=1}^{S} \beta_s x_i^s + \theta_{t(i)} + \lambda_{j(i)} + \mu_{k(i)} + \epsilon_i \quad (7)$$

where the variable x_i^s is a characteristic s of incident i, and the control variables are defined as above. We include incident (grade, incident scene location, opening code), worker (gender, age, desk location, current workload) and room time-varying (measures of current average workload) characteristics as the x_i^s variables.[19] We find that the F-statistic of joint significance of the variables x_i^s is 1.19 (p-value = .15), indicating that *SameRoom* is uncorrelated with these characteristics.[20] This finding supports the identification assumption in this paper that *SameRoom* is conditionally orthogonal to worker and incident characteristics.

6 The Effect of Being Able to Communicate

In this section we show that the ability to communicate face-to-face increases the productivity of the communication receiver (faster allocation and response time) while decreasing it for its sender (higher not ready time). We first show that co-location affects the productivity of handler and operator in the directions predicted in Section 4, and that these effects can be interpreted in a causal manner. We then argue that face-to-face communication is the most plausible mechanism through which co-location affects productivity.

The Effect of Co-Location on Productivity Our baseline regressions are variations of equation (6). In the first two columns of Panel A Table 3 we find that allocation and response times are approximately 2% faster on average when handler and operator are co-located. In the last column we find that co-location has the opposite effect on the handler's not ready time: handlers who have just finished creating co-located incidents spend 2.5% more time on average before being available to take a new call. Panel A Table 3 provides support for the first prediction of the theoretical framework above.

< INCLUDE Table 3 here>

Establishing a Causal Interpretation Consistently with the theoretical framework, our interpretation of the findings in Panel A Table 3 is that co-location allows the teammates the

[19] All the variables x_i^s are continuous except for the categorical variables capturing gender, grade and opening code. These latter variables are represented in the regression by a set of dummy variables.

[20] An alternative way of testing for balance is to run a set of regressions such that the x_i^s alternate as dependent variables and *SameRoom* is the main independent variable. We report confidence intervals from these regressions in Appendix Tables A3 and A4 and find for most regressions, the coefficient associated with *SameRoom* is not statistically different from zero.

opportunity to communicate face-to-face. Because of this opportunity, the average productivity (allocation and response time) of the operator is higher while the average productivity (not ready time) of the handler is lower.

Before we proceed further, we confirm that co-location has a causal effect on these productivity measures, as opposed to capturing unobserved components of the handler/location match or the handler/operator match. To understand the first potential confounding effect, note in Figure 3 that co-location occurs when a handler based in a location is assigned an incident from the geographical area surrounding that location. If handlers are more knowledgeable about cases that occur close to their workplaces *and* this knowledge translates into better electronic communication, the findings in the first two columns of Panel A Table 3 may result from proximity to the incident scene, rather than co-location with the operator.

The second potential confounding effect is that co-location might be a proxy for some unobserved dimension of the similarity between teammates. In an extreme example, imagine that workers communicate through room-specific language, which makes electronic communication with individuals outside one's room less efficient. This would be the case if, for instance, there are strong local dialects and workers in a room are drawn from neighbourhoods surrounding that room.[21] In that case, co-location would represent a proxy for the ease of electronic communication between teammates, as opposed to providing an additional channel of communication between them. A similar confounding effect would arise if handlers and operators get to know each other better (and therefore to better electronically communicate) when they are co-located.

Before providing empirical evidence, we must emphasise that these interpretations do not easily explain the not ready time finding. Column 3 of Panel A Table 3 shows that co-located handlers are doing something different *after* entering the electronic information and creating the incident. Our interpretation is that, that something is taking advantage of co-location to communicate face-to-face with the operator. Alternative interpretations relying on unobservables leading to better electronic communication before the incident is created, find it difficult to predict differential behaviour *after* the incident's creation.

In addition to the argument above, we evaluate the plausibility of the aforementioned alternative interpretations by exploiting post-2012 information. As discussed in Section 2, the 2012 OCB reorganisation relocated all the handlers to Trafford, while the operators were split between Claytonbrook and Tameside. In Panel B Table 3 we therefore expand the baseline sample to also include a placebo period comprising of the years 2012 and 2013. We create the variable $SameRoomMatch$ to capture whether a handler/operator pair were

[21] In practice, Appendix Table A3 shows that co-located workers are not more likely to be similar to each other in their observable characteristics. However, they are likely to be similar in their unobservables, such as the area of Manchester where they live.

co-located during the baseline period, and interact this variable separately with a baseline period dummy and with a placebo period (during which no pair is ever co-located) dummy. The interactions with the baseline period replicate the baseline findings from Panel A. The interactions with the placebo period are always statistically insignificant, suggesting that pairs co-located in the 2009-2011 period were *not* associated with different productivity in the 2012-2013 period.[22] In Panel C we add a set of handler/operator pair fixed effects to the expanded sample (the introduction of these effects absorbs the placebo same room variable). We find that *the same pairs of workers* were associated with lower allocation and response time in the earlier period when they were co-located, relative to the later period when they were not.[23]

Understanding the Quantitative Implications of the 2012 Reorganisation The estimates in Panels B and C Table 3 suggest that the 2012 reorganisation had detrimental implications for the speed at which the police was able to respond to the average incident. While the coefficient is 2.4% (Column 3 Panel C), it is important to note that this average effect masks substantial heterogeneity across types of incidents. We devote Section 7 to explore heterogeneous effects in more depth, but we note here that the 2012 reorganisation increased response time substantially for incidents that the handler initially classified as potential violent crimes.[24]

In Panel D we replicate the Panel C specification but now separately identify the effect for violent crimes. We find an overall effect of 8%. Calculated at the mean (respectively median) of response time for this subgroup, the evidence from Panel D implies that separating handlers and operators in 2012 increased response time by 8.5 minutes (respectively, 1 minute) for violent crimes. This suggests that handler and operator were much more willing to use the face-to-face communication channel for incidents that were deemed as more serious and urgent.

Panel D has implications for similar organisations deciding whether to invest in providing workers with an additional communication channel. The ability to communicate faster

[22]Following the reorganisation, operators remained in their previous roles in terms of the subdivisions for which they dispatched officers. Therefore, a post-2012 handler-operator match continues to accurately capture whether the handler is assigned a case from the geographical area around their pre-2012 workplace.

[23]In Appendix Table A8, we perform a complementary analysis at the handler/operator pair level. We first use baseline regressions to estimate handler/operator pair fixed effects, separately for the baseline period and placebo period. We then regress these pair fixed effects on a same room match dummy, a placebo period dummy and their interaction. We find that same room pairs have 4% lower allocation/response time and 5% higher not ready time during the baseline period. As expected, these differences disappear during the placebo period.

[24]We refer to these incidents as potential violent crimes because the final classification is made by the arriving officer. Here we are using the initial classification made by the handler, which is the information available to the OCB workers at the time that an officer is allocated.

may be exercised less often for typical, relatively low-importance, tasks. However, for cases that really matter (such as 10% of incidents in our sample that are classified as violent crimes), having the option to communicate faster can be very useful for the workers of an organisation.

Establishing Face-to-Face Communication as the Mechanism The findings above have established the existence of a causal relation between co-location and productivity. Our preferred explanation is that co-location permits face-to-face interactions which can be used to communicate relevant details about incidents. However, even in the absence of face-to-face communication, workers could silently observe and exert visual pressure on each other, which could potentially affect their behaviour. In our setting, there are two plausible channels through which this *silent* pressure could affect allocation time. Firstly, the handler could react to this hypothetical pressure by exerting more effort in the transmission of electronic information. Secondly, the operator could also react either by exerting more effort to interpret the electronic information or by allocating scarce resources, such as police officers, to co-located incidents and in detriment of other incidents.

Before examining some empirical evidence, we again emphasise that our theoretical model provides a unified framework to understand the differential behaviour under co-location of *both* handler and operator. While the above mechanisms based on co-location leading to mutual silent pressure and better electronic communication could rationalise faster allocation times, they do not naturally predict higher handler not ready time after the incident has been created.

In Table 4 we further test whether observable characteristics of the handler's electronic communication differ depending on co-location. We first use the handler's creation time: the time elapsed between the handler answering the call and the creation of the incident in the GMPICS system. Remember that this creation time takes place before the radio operator is informed of the incident's existence (see Figure 2). If the handler anticipates that a same room colleague will read the incident's description, feels some silent pressure, and transmits more precise electronic communication, we should observe more time devoted to writing the description of the incident, as well as to the elicitation of the information from the caller. In Column 1 of Table 4 we however replicate our baseline specification using creation time as dependent variable and find that it is unaffected by co-location.

< INCLUDE Table 4 here>

As complementary measures of the quality of the electronic communication, we use the number of characters and number of words in the first line of the description of the

incident.[25] In Columns 2 and 3 of Table 4 we find that these variables are not different for co-located incidents. We conclude that higher effort by the handler and the associated better electronic communication (to the extent that we can measure it) do not appear to be the mechanism through which co-location affects productivity.[26]

In Table 5 we investigate the existence of potential spillovers from same room incidents into other contemporaneous incidents. Operators typically have several incidents open (i.e. yet to be allocated) at the same time. Theoretically, same room incidents can generate both positive and negative spillovers. Positive spillovers will occur, for instance, when the time and effort that the operator saves on a same room incident (as a result of being able to gather information more efficiently) is redistributed to other contemporaneous incidents. Negative spillovers are equally plausible. One potential channel would be operators assigning higher priority to incidents that have been created by co-located handlers. If that was the case, the improvement in performance for same room incidents that we document in Table 3 would be, at least partially, at the expense of other contemporaneous incidents, as attention is diverted away from them.

< INCLUDE Table 5 here>

To study whether spillovers are present in our setting, we replicate our baseline specification and use the percentage of incidents assigned to the operator that, in the period surrounding the index incident, are same room incidents as the independent variable. Given the uncertainty about the time horizon on which spillovers might occur, we calculate the independent variable at the 60, 30 and 15 minute horizon. We fail to find that a higher share of same room incidents translates into differential performance for other contemporaneous incidents. However, given that the estimates are positive and the standard errors large, we should regard this test as a relatively weak one. [27]

Overall, we conclude that, with the caveat that some of our tests are underpowered, the mechanism of face-to-face communication represents the best explanation for the set of

[25] Unfortunately, due to a combination of technical challenges and the extreme confidentiality of this information, we were not able to obtain the content of these descriptions. The first line of the incident description consists of a maximum of 210 characters, and serves as a quick summary of the nature of the incident. When operators have more than one incident open at one time, they typically only see the first line of this description, which then plays a role similar to the subject heading of an email in an inbox.

[26] The finding in Table 4 that co-located workers communicating face-to-face do not decrease their electronic communication might seem surprising, as one would expect that different forms of communication are substitutes of each other. Note, however, that the log is not written only for the operator to read, but also for police officers, detectives and a wide variety of other GMP staff. Because of this wide readership, handlers are expected to document all relevant details in the log and they are unlikely to omit some even if they are able to communicate in person with the operator.

[27] Appendix Table A4 repeats the exercise in Table 5 using the additional outcome measures from Table 11. We find that incidents surrounding co-located incidents are not more likely to become crimes, or less likely to be solved quickly conditional on becoming crimes.

results in Table 3.

7 Heterogeneity

In this section we show that the heterogeneity of the estimated effects is broadly consistent with the remaining insights from the theoretical framework, as discussed in Section 4. Before we do these heterogeneity exercises, we first study a more general theoretical insight. Remember that the theoretical framework in Section 3 predicts that co-located incidents with a high(er than expected) not ready time, should often be those with a low(er than expected) allocation and response times. For instance, incidents for which the operator enjoys very little slack should have an effect of co-location that is stronger both for allocation time (negatively) and for not ready time (positively).

In Table 6, we find that these incident-level negative correlations between allocation/response and not ready time are indeed present in our data. This is consistent with the notion that, for some incidents, the two workers take advantage of co-location to communicate face-to-face, and this activity affects their respective productivities in opposite directions. Of course, the analysis in Table 6 does not illuminate what specific incident characteristics are generating this negative correlation. We use the rest of this Section to explore this issue.

< INCLUDE Table 6 here>

The Handler's Incentive to Help We now examine the heterogeneity of the effects with respect to proxies for the handler's motivation to help the operator. We first study the role of handler career incentives. We argue in Section 4 that handlers during the month of the performance review might be more focused on minimising not ready time, as this is the indicator that they are most directly evaluated on. Conversely, handlers recently moved to the highest pay grade might become less focused on their own performance, and more willing to help operators. In Table 7 we find evidence in support of these hypotheses. For instance, the effect of co-location on allocation time is 3.4% higher (i.e. more negative) when the handler has been recently upgraded in pay, while the effect on not ready time is 9.3% higher. The interaction with the performance review month is also statistically significant, both for allocation and response time and for not ready time. These empirical findings confirm that features of handler evaluation and promotion systems play a role in their motivation to help their teammates.

< INCLUDE Table 7 here>

In Table 8 we examine the role of the links between handler and operator, in particular: (a) whether they are of the same gender, (b) the difference in their ages, and (c) the number of past incidents they have worked together in the past. As we discuss in Section 4, our argument is that handlers may be more willing to help operators whom they share more in common with. Therefore, allocation and response time should be lower for these types of teams, and not ready time should be higher.

< INCLUDE Table 8 here>

In Table 8 we replicate the baseline regressions from Panel A Table 3, but interact these proxies with the same room variable (including the uninteracted proxies). To ensure that we are isolating the effect of the handler/operator pair experience, the specification further controls for the individual experiences of handler and operator and their interactions with the same room variable.[28] We find that the three interactions of interest are statistically significant in the allocation and response regressions. The effect of co-location on allocation time is 1.6 percentage points higher when handler and operator share the same gender. Likewise, a 10% increase in the age difference (respectively, number of past interactions) between handler and operator decreases the effect of co-location on performance by 2.5% (respectively, it increases it by 2.1%). We also find that the interactions are statistically insignificant in the not ready time regression. Overall, we interpret Table 8 as providing partial support for the notion that more homogeneous teams are more motivated to help each other through communication.[29]

The Efficiency of Communication: Distance Inside the Room Panel A of Table 3 has established that the co-location of handler and operator affects their respective productivities, relative to them working in rooms in separate areas of Manchester. We now investigate whether these effects change as distance decreases *even when handler and operator are already working in the same room.*

The assignment of desks to workers in the OCB was as follows. Inside a room, a fixed desk would be earmarked for the operator overseeing a specific subdivision. Handlers, on the other hand, were free to work from any remaining and available desk.[30] To measure the within-room distance between desks, we use yearly-updated floorplans of the four OCB

[28]The estimated coefficients for these control variables are omitted from Table 8, but can be found in Appendix Table A5.

[29]An alternative interpretation of these results, also consistent with the theoretical model, is that workers who are more similar to each other or have interacted more often in the past are able to communicate more efficiently.

[30]Appendix Figures A5A and A5B display the distribution of handlers spending specific intervals of shifts at their preferred desk(s). The figures show that most handlers vary their location choices significantly, as opposed to always sitting at the same desk(s).

rooms (see Figure 5 for an example).[31] We set distance to zero if handler and operator are not in the same room, and add the interaction of distance and the same room variable to our baseline specification.

< INCLUDE Figure 5 here>

Table 9 presents two sets of results. In Panel A, we replicate the baseline specification (6) and add the within-room distance interaction. We find positive and statistically significant effects on the operator productivity variables. For instance, a 10% decrease in within-room distance is associated with a 2.6% increase in the effect of *SameRoom* on allocation time. Note that, by itself, the evidence in Panel A could confound the treatment effect of within-room distance with the differential selection of workers who choose to sit together. To alleviate this concern, we add a set of handler/operator pair fixed effects (which absorb the same room variable) in Panel B. We find that *the same pair of workers operating from the same room* are associated with higher operator productivity in days when their desks are closer together. The estimated coefficients are in fact almost identical to those in Panel A, suggesting that differential selection does not appear to play a significant role here.[32]

< INCLUDE Table 9 here>

In Section 4 we argued that the within-room distance can be interpreted as a proxy for the efficiency of the communication technology. The argument was that, because handler and operator sitting closer need to walk less to communicate with each other, any additional time spent by the handler should translate into more information transmission and therefore lower allocation time for the operator. An alternative interpretation is that workers sitting close by face lower psychological barriers to communicating with each other, perhaps because they are already within each other's sight.

[31]The floorplans are unfortunately not to scale, which prevents us from measuring distance in metric units and is likely to introduce measurement error in the within-room distance variable. Instead, desks are depicted in the floorplans in a matrix (x, y) format. Our measure is therefore the euclidean distance between desks inside this matrix. $D = \sqrt{[(y_{RO} - y_H)^2 + (x_{RO} - x_H)^2]}$, where y_{RO} is the position of the operator along the row dimension and the other coordinates are defined accordingly.

[32]Because handlers choose the desks where they sit daily (conditional on these desks being unoccupied), within-room distance cannot be considered random. Even after controlling for pair fixed effects, a concern might remain that distance is correlated with within-pair time-varying characteristics. While we cannot fully eliminate this concern, we note three things. Firstly, the effect of within-room distance on allocation and response times is robust to the inclusion of handler/operator/semester fixed effects (see Appendix Table A6). Secondly, the findings remain unchanged when studying handlers starting their shift at times when a large percentage of desks are occupied (Panel A Appendix Table A7). These handlers are more constrained in their location choices, but we find that the effect of within-room distance is similar for them. Thirdly, the findings are also unchanged when studying handlers who are not currently sitting at their preferred desks (Panel B Appendix Table A7).

The Teammates' Slack(s) Part 3 of Proposition 2 predicts that the optimal amount of communication is higher when the operator enjoys very little slack and the handler enjoys a lot of slack. In our third heterogeneity exercise, we study whether proxies for the slack experienced by the teammates correlate with the amount of communication effort undertaken.

Our first proxy for operator slack is the number of incidents created in the subdivision that the operator is overseeing during the hour of the index incident. Because there is only one operator responsible for a subdivision at any one time, this measure captures the likely size of the queue being faced by the operator at that point in time well. We construct a similar proxy to compute the slack being faced by the handler. However, because all handlers share a common queue, we measure this at the organisational level: the number of calls *per on-duty handler* received during the recent past. For ease of interpretation, both variables are entered as above-median dummies.

We use a second, complementary, measure to proxy operator slack: their inherent speed at allocating incidents. Specifically, we use the subsample of non-co-located incidents, and compute the average allocation time of each operator. Our measure is then a dummy variable taking value one for those operators who are faster than the median operator, under non-co-location.[33]

< INCLUDE Table 10 here>

The results are displayed in Table 10. We find that the estimated interactions are broadly consistent with the predictions of the model. In particular, the effects of co-location on allocation and response times are larger when: (a) the operator is inherently slow, (b) the operator has recently received a large number of incidents, and (c) handlers have recently received a low number of incidents per capita. For instance, the effect of co-location on allocation time is 2.9% higher when the operator is above-median in terms of the number of recent incidents received.

In the not ready regressions, two out of the three coefficients have the predicted sign and are of similar magnitude as their counterparts in the allocation time regression. However, these coefficients are estimated less precisely and only one is statistically significant at the 10% level.

[33]Because operators typically work for the same subdivision, this measure reflects the characteristics of the subdivision as well as the characteristics of the specific handler. In terms of generating a similar proxy for handlers, note that it does not make much sense to study the intrinsic speed of handlers at processing incidents (i.e. their creation time plus not ready time, see Figure 2). Because all handlers take calls from the common queue, the average speed of a handler has a negligible impact on the average speed at which the common queue is processed. Instead, in Table 10, we use the average allocation time of the incidents created by a specific handler. We find that the corresponding interaction with Same Room is not statistically significant.

We interpret these estimates as providing overall support for the notion that the relative slack of the two teammates determines the amount of communication effort undertaken.

8 Social Welfare

In the previous sections we have seen that a simple model of 'communication as help' explains how the performance measures of handler and operator are affected by the opportunity to communicate face-to-face. A limitation of our analysis is that the performance measures used so far are only partial, as OCB workers are likely to be motivated by a multiplicity of objectives. Furthermore, their objectives may not necessarily coincide with those of the population at large, making the relation with social welfare less than automatic. For instance, they may place higher-than-optimal weight on response time, in detriment of other more socially valuable outcomes.

The existence of unobserved dimensions of performance implies that a quantitative welfare analysis of the effects of face-to-face communication represents a very challenging task. However, in this section we provide two indirect pieces of evidence supporting the notion that face-to-face communication is associated with an increase in social welfare. Firstly, we show that other outcome measures do not appear to worsen, and even improve slightly, when handler and operator are co-located. Secondly, we show that handler and operator engage in more communication in incidents in which the social benefit of decreasing response time is higher, which suggests that they are considering social welfare in their communication decisions.

Effects on Other Outcome Measures The fact that we cannot observe all dimensions of performance raises the concern that face-to-face communication may be worsening unobserved dimensions, even while it improves response time. For instance, it may be that the higher speed of response documented in Table 3 is at the expense of rushed decision-making and a worse handling of the incident, perhaps even reducing social welfare.

The GMP commissioned surveys of randomly chosen crime victims, in order to measure satisfaction with the police handling of the incident. The survey answers constitute a good incident-level measure of social welfare, or at least victim welfare in the 15% of incidents classified as crimes. Figure 6 shows a strongly negative relation between response time and victim satisfaction, alleviating concerns that faster responses might on average be at the expense of worse decision-making. Unfortunately, we cannot study how victim satisfaction is affected by co-location because the number of survey responses overlapping with our baseline sample period is very small.

< INCLUDE Figure 6 here>

In Table 11 we use the baseline equation (6) to estimate the effect of co-location on other incident outcome measures. In Column 1, the dependent variable is a dummy for whether the incident ended up as a crime. Police behaviour is unlikely to affect this variable for most incidents, but in a small minority (e.g. a verbal altercation that escalates into violence) a fast and efficient arriving officer could prevent the incident from becoming a crime. We find a negative coefficient for co-location (significant at the 10% level) suggesting that handler and operator communicating face-to-face are able to prevent some incidents from escalating.

< INCLUDE Table 11 here>

In Columns 2 and 3 we restrict the sample to include only crimes, and use the likelihoods that they are cleared or cleared within 24 hours as dependent variables. We again find weak evidence that co-location improves the chances that the crime is solved relatively quickly.

In Column 4 we use information on the crimes start and end times, and restrict the sample to include only crimes that are 'in progress' at the time that the handler answers the phone. The dependent variable in Column 4 is the additional duration of the crime timespan. The rationale for this regression is that, because many of these crimes in progress are likely to involve violence, we would expect a shorter additional timespan to be associated with higher victim welfare. We find that co-located crimes in progress have an 11% shorter additional timespan.

Overall, Table 11 is inconsistent with the notion that the decrease in response time under co-location is at the expense of other incident outcomes. Rather, the evidence is the opposite: an improvement in other outcome variables when face-to-face communication is possible. Unfortunately, the fact that some of the measures are only defined for a small percentage of incidents implies that the estimates are noisy, and therefore that we cannot undertake the same type of heterogeneity analysis as we did for response time.

Heterogeneity by Social Welfare Sensitiveness In some types of incidents, social welfare is largely unaffected by how fast the police respond. In other types of incidents, it is instead very important from a welfare perspective that the police arrive quickly. In this subsection, we study whether handler and operator adjust their communication to the sensitiveness of social welfare to response time.

We start by using quantile regressions to study the percentiles of the distribution in which co-location has the highest effects. To do so, we first regress allocation and response time on the baseline set of controls and obtain their residuals. We then estimate the conditional distribution of these residuals on co-location, at different quantiles (Bandiera et al., 2010). The resulting estimates, plotted in Figure 7, decrease almost monotonically as

the quantile increases. For instance, the effect of co-location is around 6-10% for the first quantile, and it becomes statistically indistinguishable from zero around the 70th quantile. Intuitively, we would expect incidents on the left of the allocation/response time distributions to be regarded as more urgent than those on the right of the distributions. Figure 7 therefore suggests that handler and operator take more advantage of co-location to communicate face-to-face for incidents that, even when not co-located, they regard as requiring a faster response.

< INCLUDE Figure 7 here>

The quantile regressions, while indirectly informative, do not indicate whether the incidents benefitting more from co-location are those with the highest potential improvement in social welfare. The first measure that we use to study this question is victim satisfaction, as described in the previous subsection. We regress this variable on response time interacted with the type of incident (i.e. opening code and grade), and use the resulting coefficients to classify incident types on the basis of whether victim satisfaction is strongly affected by response time. We then interact the above-median satisfaction sensitiveness with co-location, and find in Table 12 that response time is lower, and not ready time is higher, for incidents where victim satisfaction is more sensitive to response time. We perform a similar exercise with the likelihood of becoming a crime, and find again in Table 12 that types of incidents for which response time is a stronger determinant are associated with more face-to-face communication.

< INCLUDE Table 12 here>

The last variable that we use is a dummy for whether there is a crime ongoing at the time that the handler answers the phone. To understand the rationale of this interaction, compare two crimes: a burglary discovered by a family upon returning from a holiday, and an assault in progress reported by a bystander. In the second incident, the importance of arriving quickly to prevent further violence is very high. In the first incident, the crime is long over by the time it is reported, and arriving quickly will not typically be as important for social welfare. In Table 12 we find that communication is higher for crimes that are ongoing at the time when handlers are first informed.

We can interpret the evidence in Table 12 through the lens of the Section 3 model. Proposition 2 predicts that communication is increasing in the parameter ω, the extent to which the handler internalises the objective of decreasing response time. Table 12 suggests that handlers (at least partly) internalise social welfare in their communication decisions, and therefore that we can regard the sensitiveness of social welfare to response time as a determinant of ω.

9 Conclusion

We have provided evidence on the interplay between co-location, productivity and a variety of characteristics of the workers and tasks in our organisation. We have shown that the ability to communicate face-to-face improves the performance of the receiver of this communication, but at the time cost of its sender. Therefore, we have argued that communication can play the role of a 'help' or 'information subsidy' activity, through which informed workers assist their less informed colleagues to do their job better. Because this activity entails a trade-off, we have argued theoretically that its optimal level is contingent upon its costs and benefits. We have shown that the workers in our organisation adjust the amount of communication to determinants of these costs and benefits.

Our results bring new evidence to the empirical economics literature both on teamwork and on communication in organisations. In this latter one, theoretical work has largely focused on the role of 'hard features' of the organisation (e.g. the organisational structure, the communication channels, etc.). We instead emphasise 'soft features', i.e. the ability of individual workers to operate within the constraints of a given organisation and adjust the amount of communication flexibly depending on circumstances. In our setting, the 2012 reorganisation deprived workers of the possibility of communicating face-to-face, which illustrates how the hard features of the organisation can restrict or even eliminate workers' ability to use the soft features to maximise performance.

The use of detailed data throughout the production process combined with the exogenous variation created by our natural experiment has allowed us to identify the effects of co-location on productivity and to interpret these effects as resulting from the ability to communicate face-to-face. Precision, however, inevitably comes at the cost of a loss of generality, because the organisation we have studied, as any other, has unique features that can affect the external validity of this study. We conclude the paper by outlining some key features and evaluating how much our main lessons likely depend on them.

Firstly, there are several characteristics of our production process that may have opposite effects on the quantitative importance that face-to-face communication has in complementing electronic communication. On the one hand, the OCB deals with a continuous inflow of (relative to many workplaces) somewhat homogeneous problems. Because over time detailed coding systems and processes have been put in place to increase the efficiency of electronic communication, the effect of being able to additionally communicate face-to-face may be smaller here than in other settings with more complex and heterogeneous problems. Other features of our setting may however be magnifying the importance of face-to-face communication. For instance, there is in our setting a dearth of substitute channels such as communication by phone. Further, the intensity of the time pressure faced by workers

is likely higher than in most organisations, and this makes filtering through the electronic communication to extract the relevant details a less attractive option than perhaps in other settings (see Section 2 and Appendix A). To sum up, these opposite forces make it difficult to extrapolate our estimated (average) effects to other organisations. Indeed, a key lesson of the paper is that, even within a specific organisation, the costs and benefits and therefore the amount of face-to-face communication will depend on the characteristics of the tasks, workflow and workers.

While the quantitative effects may be different in other organisations, three qualitative lessons can be more easily extrapolated. Firstly, communication can play the role of a 'help' or 'information subsidy' activity, and therefore is subject to a trade-off. Secondly, we find that workers respond to social welfare proxies in their communication decisions, but only partially as workers are also influenced by their own career incentives. Our last lesson derives from the finding that the opportunity to communicate face-to-face has a much larger effect for important and urgent incidents (e.g. violent or ongoing crimes). An implication from this is that an additional communication channel may be useful even if it is rarely used, as long as the tasks in which it is used are important enough.

FIGURE 1: Operational Communications Branch

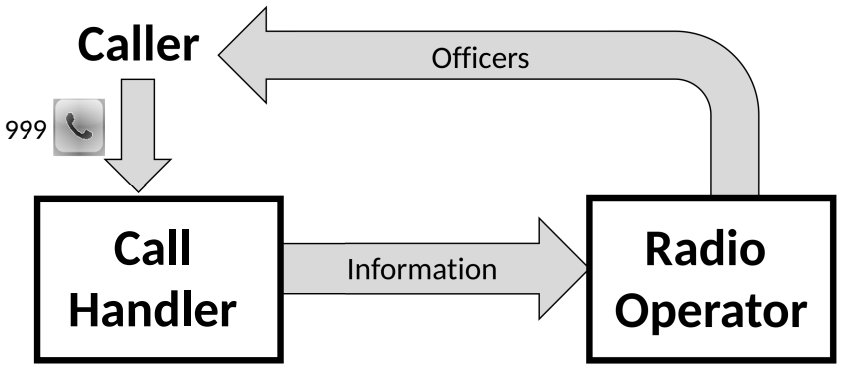


FIGURE 2: Timeline

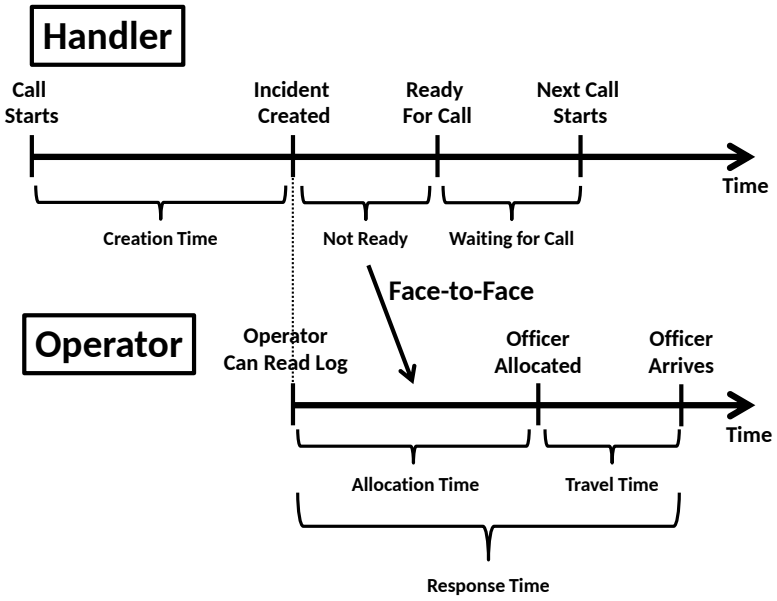


FIGURE 3: Location and Radio Operations Coverage of OCB Rooms

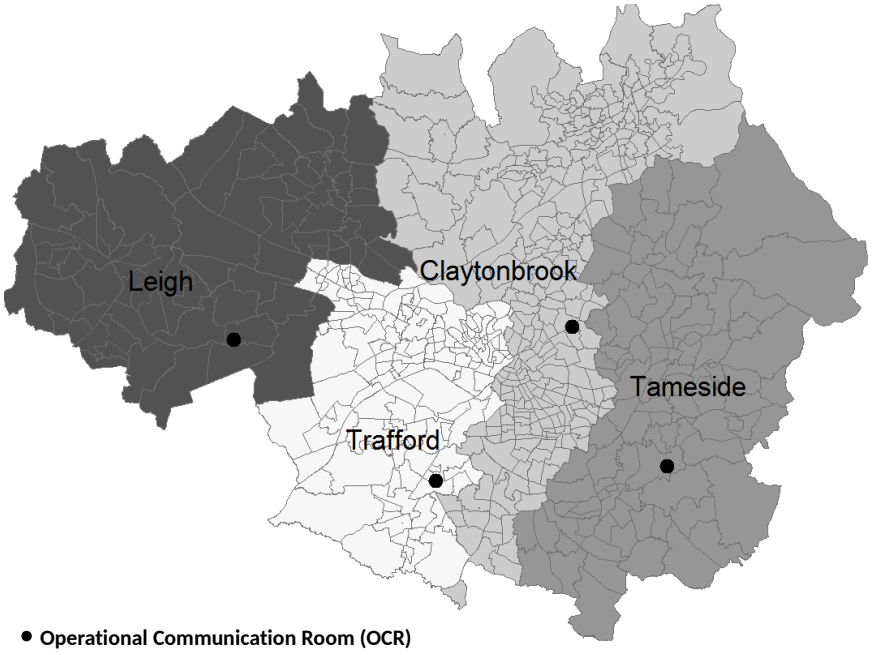
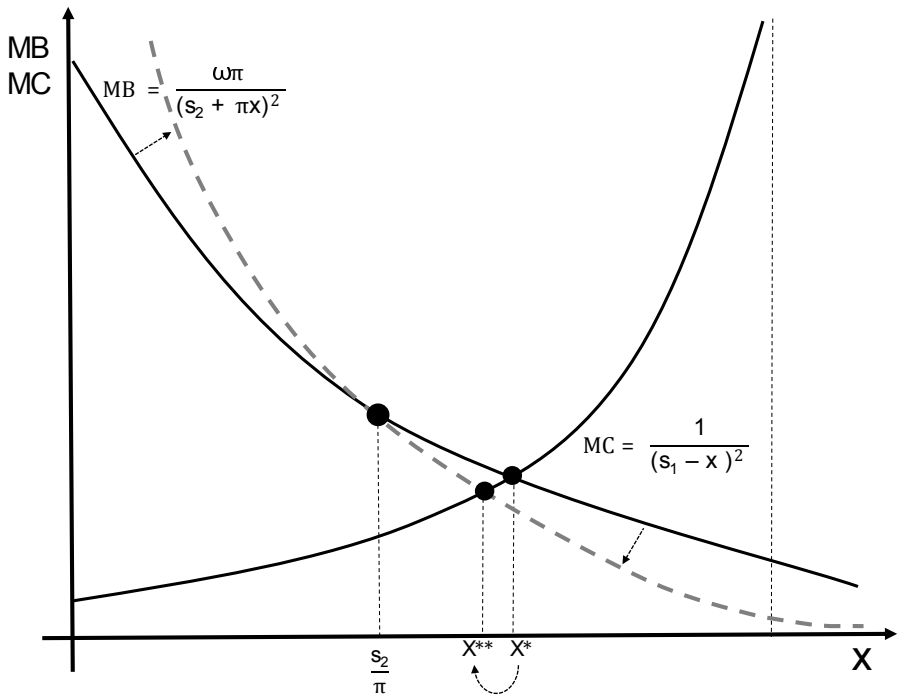


FIGURE 4: Effect of Improvement in the Communication Technology π on Communication Effort x



Note: to compute the marginal cost and marginal benefit curves we differentiate $\frac{-1}{s_1-x}$ and $\frac{\omega}{s_2+\pi x}$, respectively. The marginal cost curve does not depend on π. The derivative of the marginal benefit curve with respect to π is $\frac{\omega(s_2-\pi x)}{(s_2+\pi x)^3}$, which is positive or negative depending on whether x is lower or higher than s_2/π.

FIGURE 5: Example of OCB Room Floorplan

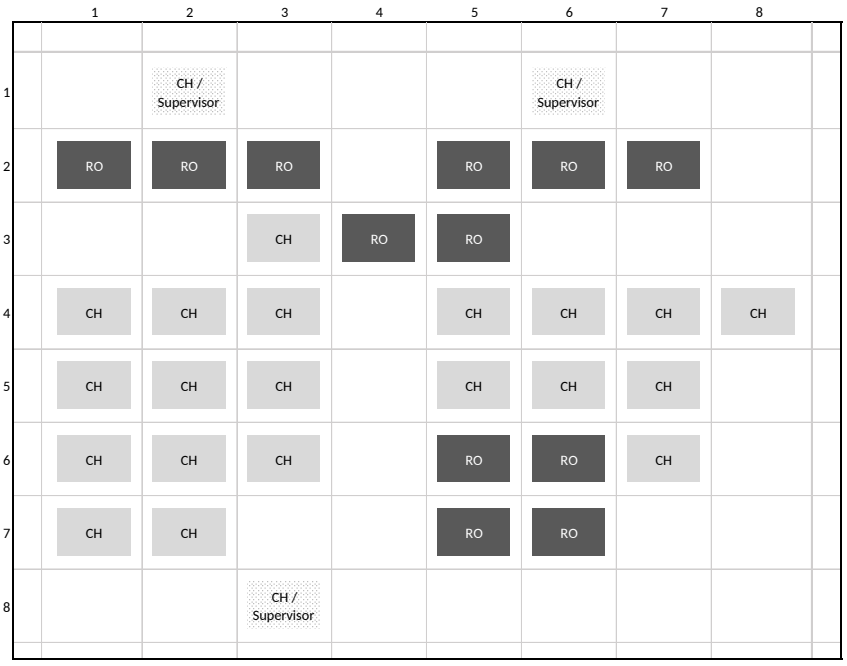
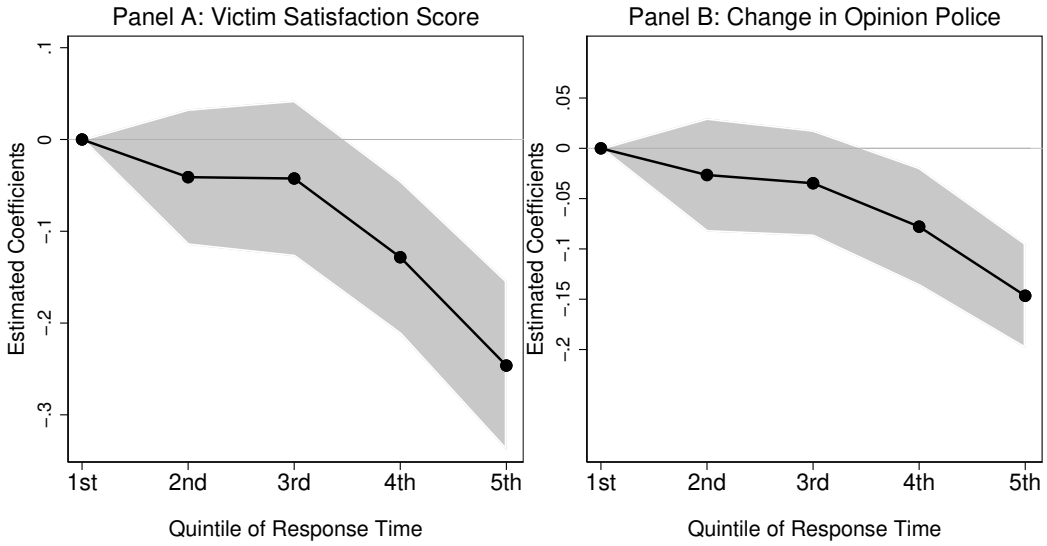
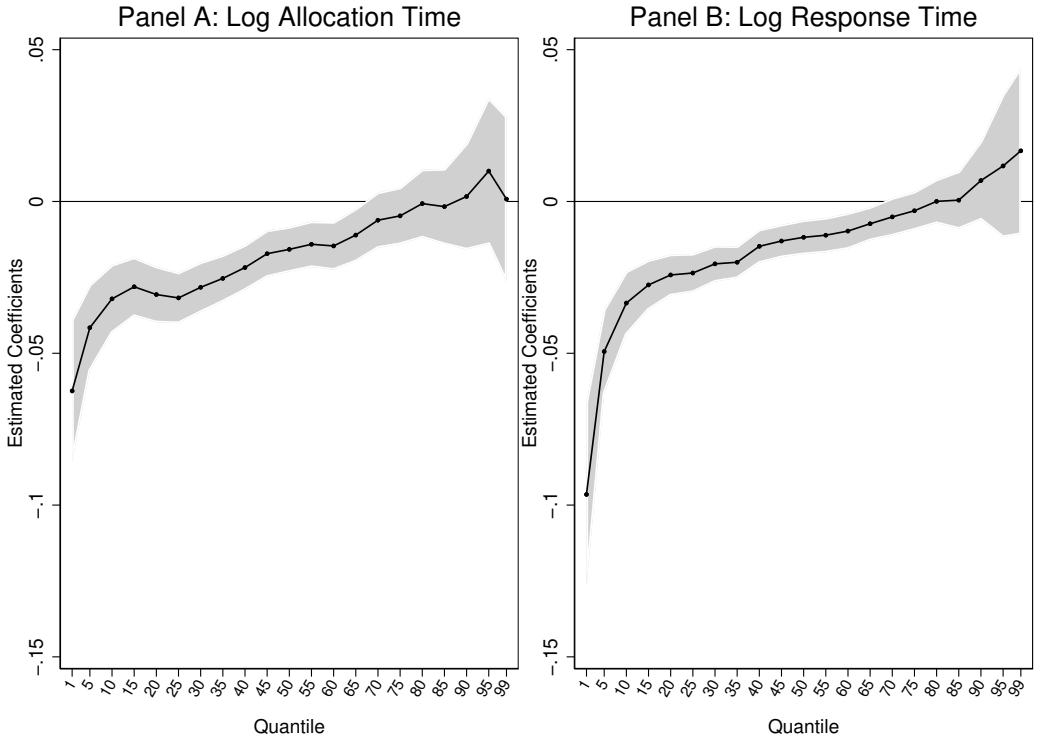


FIGURE 6: Correlation between Response Time and Victim Satisfaction



In the GMP, a subset of callers is regularly questioned about their satisfaction with the treatment they received by the police. The two most important questions are the Victim Satisfaction Score and the Positive Change in the Opinion of the Police. Each panel displays a different regression using the answers to these questions as dependent variables. The displayed coefficients are for the Quintiles of Response Time (the first quintile is added to aid visual analysis). 95% confidence intervals are displayed in the shaded grey area. All regressions control"for Grade, Call Source, Year X Month X Day, Hour of Day, Division and Opening Code. Standard error are clustered at the Year X Division level.

FIGURE 7: Quantile Regressions
Same Room Effect at Different Quantiles of Distribution



Each point displays the coefficient of a different regression. The displayed coefficients are for the effect of Same Room. 95% confidence intervals are displayed in the shaded grey area. The dependent variables in each panel are, correspondingly, the (log of) allocation and the (log of) response time, partialed out from Grade, Call Source, Year X Month X Day X Hour of Day, Radio Operator Room, Call Handler Room, Radio Operator and Handler indicators. Standard errors are estimated using 100 bootstrap replications of simultaneous-quantile regressions (Koenker, 2005).

TABLE 1: DESCRIPTIVE EVIDENCE ON HANDLER CAREER INCENTIVES

Dependent Variable	**Promoted or Upgraded**
Lagged Average Not Ready Time	-.275***
	(.093)
Lagged Average Response Time	-.095
	(.074)
Lagged Average Audit Score	.511*
	(.304)
Lagged Average Creation Time	-.018
	(.097)
Handler Indicators	Yes
Year Indicators	Yes
No Audit Indicator	Yes
Observations	282

This table studies the predictors of promotion or pay upgrade among GMP handlers. The dataset is a panel dataset of handlers and years. The dependent variable takes value one if during the current year the handler: (a) was promoted to the position of supervisor or radio operator, or (b) remained with the exact same job but was moved to a higher pay grade. The main independent variable of interest is the lagged average not ready time. The variable is standardised so the coefficient can be interpreted as the effect of a one standard deviation increase. The lagged average response time and the lagged average creation time are also standardised. The audit score is the score given by the handler's supervisor when listening to a randomly selected call by the handler. The regression controls for handler and year indicators. It further includes an indicator for whether there is no information on the handler audit scores. Standard errors are clustered at the handler level.

TABLE 2: SUMMARY STATISTICS

	Mean	Median	SD	Min	Max
Allocation Time (min.)	64.12	4.58	276.57	0	21331.78
On Target Allocation	.75	1	.43	0	1
Response Time (min.)	87.48	19.93	311.17	.05	21391.92
On Target Response	.88	1	.33	0	1
Not Ready Time (min.)	1.15	.42	2.06	0	166.27
Not Ready Time>0	.3	0	.46	0	1
Creation Time (min.)	3.89	2.85	4.95	0	219.53
Response Time Violent Crimes	83.09	11.27	316.81	.05	24104.59
Response Time Ongoing Crimes	62.74	13.93	248.7	.08	17022.81
Response Time Other Crimes	126.24	31.47	355.5	.07	16748.5
Grade 1	.2	0	.4	0	1
Grade 2	.44	0	.5	0	1
Same Room	.23	0	.42	0	1
Distance inside Room	4.34	4.24	1.78	.5	11.88
Handler Female	.7	1	.46	0	1
Operator Female	.49	0	.5	0	1
Handler's Age	39.13	39	11.37	19	67
Operator's Age	45.18	46	8.48	19	67

This Table reports summary statistics for the baseline sample (N=957,137). An observation is an incident. Allocation time is the time between the creation of the incident by the handler and the allocation of a police officer by the operator. Response time is the time between creation of the incident and the police officer arriving at the scene. On target allocation (respectively, response) is a dummy taking value one if the allocation time falls wihin the UK Home Office targets, which are 2, 20 and 120 minutes (respectively 15, 60 and 240 minutes) for Grades 1, 2 and 3. Not ready time is the time between the handler creating the incident and the handler indicating that they are available to take a new call. Not Ready¿0 is a dummy taking value one when not ready time is positive. This measure is only available for certain months in the sample (N=466,452). Creation time is the time between the handler answering the call and the creation of the incident in GMPICS. Ongoing crimes are defined as those ongoing at the time that the handler answers the phone. Other crimes refer to crimes classified as non-violent and not ongoing. Grade 1 and Grade 2 are dummies for the grade of the incident. Same Room is a dummy taking value one when handler and operator are located in the same room. Handler female and operator female are dummy variables.

TABLE 3: BASELINE REGRESSIONS

Dependent Variable (in logs)	(1) Allocation Time	(2) Response Time	(3) Not Ready Time
Panel A: Baseline Period			
Same Room	-.02***	-.017***	.025***
	(.004)	(.003)	(.009)
Baseline Controls	Yes	Yes	Yes
Observations	957,137	952,495	466,452
Panel B: Adding the Placebo Period			
Same Room Match X Baseline Period	-.02***	-.017***	.026***
	(.004)	(.003)	(.009)
Same Room Match X Placebo Period	.000	.003	.01
	(.006)	(.005)	(.01)
Baseline Controls	Yes	Yes	Yes
Baseline Controls X Placebo Period	Yes	Yes	Yes
Observations	1,371,705	1,364,017	801,955
Panel C: Controlling for Handler/Operator Pair Fixed Effects			
Same Room Match X Baseline Period	-.023***	-.024***	.016
	(.01)	(.007)	(.018)
Baseline Controls	Yes	Yes	Yes
Baseline Controls X Placebo Period	Yes	Yes	Yes
Handler/Operator Pair Fixed Effects	Yes	Yes	Yes
Observations	680,956	677,226	466,697
Panel D: Separating the Effect on Violent Crimes			
Same Room Match X Baseline Period	-.027**	-.023***	.018
	(.012)	(.009)	(.021)
SR Match X Baseline X Violent Crime	-.046*	-.056***	-.031
	(.026)	(.02)	(.042)
Baseline Controls	Yes	Yes	Yes
Baseline Controls X Placebo Period	Yes	Yes	Yes
Handler/Operator Pair Fixed Effects	Yes	Yes	Yes
Observations	680,956	677,226	466,697

This table displays estimates of OLS regressions of allocation time, response time and not ready time on whether the handler and operator are located in the same room. Allocation time is the time between the creation of the incident by the handler and the allocation of a police officer by the operator. Response time is the time between the creation of the incident and the police officer arrival at the scene. Not ready time is the time between the handler creating the incident and the handler indicating that they are available to receive a new call. Not ready time is only available for certain months in the sample. In Panel A we use the baseline period, which includes all incidents received by the GMP between November 2009 and December 2011. In Panel B, we use both the baseline period (2009-2011) and the placebo period (2012-2013). In the placebo period handlers and operators are never located in the same room. The variable same room match takes value one if the handler and the operator were located in the same room in the baseline period. Panels C and D include all incidents in the years 2011 and 2012, in order to study the same individuals in a short time window while being and not being co-located. Violent crime is a dummy taking value one if the incident was classified as a potential violent crime by the handler. The baseline controls are indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Operator Room, Handler Room, Operator and Handler. Standard errors are clustered at the Year X Month X Subdivision level.

TABLE 4: INVESTIGATING EFFECTS ON HANDLER ELECTRONIC COMMUNICATION

Dependent Variable (in logs)	(1) Creation Time	(2) Number of Characters	(3) Number of Words
Same Room	.00446	-.0004	-.00028
	(.00326)	(.00138)	(.0015)
Baseline Controls	Yes	Yes	Yes
Observations	466,125	956,194	956,194

This table displays estimates of OLS regressions of characteristics of handler electronic communication on whether the handler and the operator are co-located. The sample includes all incidents received by the GMP between November 2009 and December 2011. In Column (1) the dependent variable is the log of the creation time (i.e. the time between the handler answering the call and the creation of the incident). In Column (2) the dependent variable is the number of characters in the first line of the description of the incident (maximum number of characters = 210). In Column (3) the dependent variable is the number of words in the first line of the description of the incident. The baseline controls are indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Operator Room, Handler Room, Operator and Handler. Standard errors are clustered at the Year X Month X Subdivision level.

TABLE 5: INVESTIGATING EFFECTS ON OPERATOR ALLOCATION OF RESOURCES

	Spillovers by Same Room Incidents during Period:					
	60 minutes		30 minutes		15 minutes	
Dependent Variable (in logs)	(1) Allocation Time	(2) Response Time	(3) Allocation Time	(4) Response Time	(5) Allocation Time	(6) Response Time
% Same Room Incidents	.005	.004	.006	.007	.009	.007
Received By Operator	(.005)	(.004)	(.006)	(.004)	(.007)	(.005)
Baseline Controls	Yes	Yes	Yes	Yes	Yes	Yes
No Other Calls Indicator	Yes	Yes	Yes	Yes	Yes	Yes
Observations	957,137	952,495	957,137	952,495	957,137	952,495

This table investigates potential spillovers from same room incidents into other contemporaneous incidents. The dependent variables in the OLS regressions are allocation time and response time. The independent variable is the percentage of incidents during the index incident time period for which the handler and operator were co-located, excluding the index incident. In Columns (1) and (2) the period comprises of 60 minutes (respectively, 30 minutes for columns (3) and (4) and 15 minutes for columns (5) and (6)). The baseline controls are indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Operator Room, Handler Room, Operator and Handler. The regressions also include indicators for whether there were no calls received by the operator during the time period. Standard errors are clustered at the Year X Month X Subdivision level.

TABLE 6: CORRELATIONS BETWEEN ALLOCATION/RESPONSE AND NOT READY TIME FOR SAME ROOM AND NON-SAME ROOM INCIDENTS

Dependent Variable	(1) Allocation Time	(2) Response Time
Not Ready	.0075***	-.0008
	(.0013)	(.001)
Same Room X Not Ready	-.0056**	-.0037*
	(.0026)	(.002)
Observations	957,137	957,137

This table displays estimates of OLS regressions of the allocation/response time of an incident on handler not ready time following that incident. We use the residuals of these variables after controlling for the baseline set of controls: Grade, Call Source, Year X Month X Day X Hour of Day, Operator Room, Handler Room, Operator and Handler. The standard errors are bootstrapped with 500 repetitions.

TABLE 7: HETEROGENEITY OF THE SAME ROOM EFFECT BY MEASURES OF HANDLER CAREER INCENTIVES

Dependent Variable (in logs)	(1) Allocation Time	(2) Response Time	(3) Not Ready Time
Same Room X Performance Review Month	.027*	.024*	-.093**
	(.016)	(.013)	(.047)
Same Room X Recent Pay Upgrade	-.034**	-.021*	.096*
	(.016)	(.013)	(.05)
Baseline Controls	Yes	Yes	Yes
Handler X Year F.E.	Yes	Yes	Yes
Handler X Same Room F.E.	Yes	Yes	Yes
Year X Month X Same Room F.E.	Yes	Yes	Yes
Performance Review Month	Yes	Yes	Yes
Recent Pay Upgrade	Yes	Yes	Yes
Observations	891,360	887,017	426,331

This table displays estimates of OLS regressions of allocation time, response time and not ready time on the same room dummy, interacted with: (a) whether the handler is in the month of their yearly performance review, and (b) whether the handler has recently been upgraded to a higher pay grade. The baseline controls include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Operator Room, Handler Room, Operator and Handler. Standard errors are clustered at the Year X Month X Subdivision level.

TABLE 8: HETEROGENEITY OF SAME ROOM EFFECT BY MEASURES OF HANDLER/OPERATOR MATCH

Dependent Variable (in logs)	(1) Allocation Time	(2) Response Time	(3) Not Ready Time
Same Room	-.021	-.031*	.113**
	(.023)	(.018)	(.05)
Same Room X Same Gender	-.016**	-.019***	-.007
	(.008)	(.006)	(.016)
Same Room XLog Difference in Age	.025***	.024***	.003
	(.005)	(.004)	(.01)
Same Room X Log # Past Interactions	-.021***	-.019***	-.008
	(.005)	(.004)	(.01)
Baseline Controls	Yes	Yes	Yes
Gender/Age/Past Interactions	Yes	Yes	Yes
Handler/Operator Experience	Yes	Yes	Yes
Same Room X Handler/Operator Experience	Yes	Yes	Yes
Observations	923,156	918,628	437,169

This table displays estimates of OLS regressions of allocation time, response time and not ready time on the Same Room dummy, interacted with whether the operator and the handler are of the same gender, with the log of their difference in age, and with the number of previous incidents in which they have worked together. Baseline controls include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Operator Room X Year, Handler Room X Year, Operator and Handler. Standard errors are clustered at the Year X Month X Subdivision level.

TABLE 9: HETEROGENEITY OF SAME ROOM EFFECT BY DISTANCE INSIDE ROOM

Dependent Variable (in logs)	(1) Allocation Time	(2) Response Time	(3) Not Ready Time
Panel A: Baseline Controls			
Same Room	-.049***	-.035***	.002
	(.012)	(.01)	(.022)
Same Room X Log Distance	.026***	.018***	.018
	(.009)	(.007)	(.015)
Baseline Controls	Yes	Yes	Yes
Observations	944,448	939,878	466,409
Panel B: Adding Handler/Operator Pair Fixed Effects			
Same Room X Log Distance	.027***	.017**	.021
	(.01)	(.008)	(.017)
Baseline Controls	Yes	Yes	Yes
Handler/Operator Pair Indicators	Yes	Yes	Yes
Observations	932,441	927,871	455,184

This table displays estimates of OLS regressions of allocation time, response time and not ready time on whether the handler and the operator are co-located, interacted with the distance between their desks when they are in the same room. The distance between desks is calculated as the euclidean distance in the floorplans provided by the GMP. Baseline controls include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Operator Room X Year and Handler Room X Year and Operator and Handler Identifiers. Panel B includes Operator/Handler Pair Identifiers. Standard errors are clustered at the Year X Month X Subdivision level.

TABLE 10: HETEROGENEITY OF SAME ROOM EFFECT BY MEASURES OF WORKER SLACK

Dependent Variable (in logs)	(1) Allocation Time	(2) Response Time	(3) Not Ready Time
Same Room	-.019***	-.02***	.048***
	(.008)	(.006)	(.016)
Same Room X High Operator Inflow	-.029***	-.019***	-.004
	(.008)	(.006)	(.016)
Same Room X High Handler Inflow	.013*	.005	-.014
	(.008)	(.006)	(.015)
Same Room X Fast Operator	.012	.021***	-.026*
	(.008)	(.006)	(.015)
Same Room X Fast Handler	-.005	-.004	-.006
	(.007)	(.006)	(.016)
Baseline Controls	Yes	Yes	Yes
High Operator Inflow	Yes	Yes	Yes
High Handler Inflow	Yes	Yes	Yes
Observations	954,454	949,820	465,303

This table displays estimates of OLS regressions of allocation time, response time and not ready time on the Same Room dummy, interacted with measures of the operator's slack, the handler's slack High Operator Inflow takes value one when the incident's operator has received a high (above median) number of incidents during the index hour. High Handler Inflow takes value one for half-hour periods during which the number of calls per on-duty handler has been relatively high (i.e. above median). Fast Operator takes value one for operators who allocate non-co-located incidents faster than the median operator. Fast Handler takes value one for handlers whose non-co-located incidents are allocated faster than for the median handler. All regressions also include indicators for Call Source, Grade, Year X Month X Day X Hour of Day, Operator Room, Handler Room, Operator and Handler, and the uninteracted High Operator Inflow, High Handler Inflow, Fast Operator and Fast Handler. Standard errors are clustered at the Year X Month X Subdivision level.

TABLE 11: EFFECTS ON OTHER OUTCOME MEASURES

Dependent Variable (in logs)	(1) Escalated to Crime	(2) Cleared	(3) Cleared within 24hrs	(4) Crime Duration
Same Room	-.001*	-.003	.002*	-.116*
	(.001)	(.003)	(.001)	(.066)
Baseline Controls	Yes	Yes	Yes	Yes
Observations	957,164	186,379	186,379	14,595

This table displays estimates of OLS regressions of different outcome variables on whether the handler and operator are co-located, interacted with measures of the importance of response time for social welfare. In Column (1) the dependent variable is a dummy taking value one if the incident was classified as a crime (this classification is typically done by the arriving police officer). In Columns (2) and (3) the dependent variables are dummies taking value one if the crime was cleared, or cleared within 24 hours of the crime being reported. The samples in these columns include only incidents that the police classified as crimes. In Column (4) the dependent variable is the log of the additional crime duration, which is the time elasped between the handler answering the phone and the end time of the crime. The sample in this column includes only crimes that are ongoing as the handler answers the phone. Baseline controls include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Operator Room and Handler Room and Operator and Handler Identifiers. Standard errors are clustered at the Year X Month X Subdivision level.

TABLE 12: HETEROGENEITY OF SAME ROOM EFFECT BY MEASURES OF THE IMPORTANCE OF RESPONSE TIME FOR SOCIAL WELFARE

Dependent Variable (in logs)	(1) Allocation Time	(2) Response Time	(3) Not Ready Time
Same Room	-.001	-.005	-.003
	(.007)	(.005)	(.015)
Same Room X Crime Sensitiveness	-.015*	-.008	.026*
	(.008)	(.006)	(.015)
Same Room X Satisfaction Sensitiveness	-.025***	-.023***	.032*
	(.009)	(.007)	(.016)
Same Room X Ongoing Crime	-.034*	-.029**	.091***
	(.02)	(.014)	(.037)
Baseline Controls	Yes	Yes	Yes
Crime Sensitiveness	Yes	Yes	Yes
Satisfaction Sensitiveness	Yes	Yes	Yes
Ongoing Crime	Yes	Yes	Yes
Observations	781,679	778,324	403,759

This table displays estimates of OLS regressions of allocation time, response time and not ready time on whether the handler and operator are co-located, interacted with measures of whether response time is important to increase social welfare. The crime sensitiveness variable is a dummy which we calculate as follows. We first use the baseline sample to regress the likelihood of an incident becoming a crime on response time interacted with the type of incident. Crime sensitiveness then takes value one if the coefficient on response time for that type of incident is above the median. We calculate the victim satisfaction dummy in the same way, with the exception that the sample used includes only incidents for which the victim satisfaction measure is available. Ongoing crime is a dummy taking value one if the crime is still ongoing at the time that the handler answers the phone. Baseline controls include indicators for Grade, Call Source, Year X Month X Day X Hour of Day, Operator Room and Handler Room and Operator and Handler Identifiers. Standard errors are clustered at the Year X Month X Subdivision level.

REFERENCES

Alonso, R., Dessein, W. and Matouschek, N. (2008), "When Does Coordination Require Centralization?", *American Economic Review*, 98(1): 145-79.

Agrawal, A., and Goldfarb, A. (2008), "Restructuring Research: Communication Costs and the Democratization of University Innovation.", *American Economic Review*, 98(4): 1578-90.

Arenas, A., Cabrales, A., Danon, L., Diaz-Guilera, A., Guimera, R., and Vega-Redondo, F. (2010), "Optimal Information Transmission in Organizations: Search and Congestion", *Review of Economic Design*, 14(1-2): 75-93.

Arrow, K. J. (1974), "The Limits of Organization", Norton.

Barnlund, D. C. (1970), "A Transactional Model of Communication", in Sereno, K. K. and Mortenson, C. D. (Eds.), *Foundations of Communication Theory*, Harper and Row.

Bandiera, O., Barankay, I., and Rasul, I. (2010), "Social Incentives in the Workplace", *Review of Economic Studies*, 77(2): 417-458.

Beggs, A. W. (2001), "Queues and Hierarchies", *Review of Economic Studies*, 68(2): 297-322.**Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J.** (2017), "When Should You Adjust Standard Errors for Clustering?", (No. w24003). National Bureau of Economic Research.

Berger, J., Herbertz, C. and Sliwka, D. (2011), "Incentives and Cooperation in Firms: Field Evidence", *Working Paper*.

Besley, T., and Ghatak, M. (2005), "Competition and Incentives with Motivated Agents", *American Economic Review*, 95(3): 616-636.

Bloom, N., Garicano, L., Sadun, R., and Van Reenen, J. (2014), "The Distinct Effects of Information Technology and Communication Technology on Firm Organization", *Management Science*, 60(12): 2859-2885.

Bolton, P., and Dewatripont, M. (1994), "The Firm as a Communication Network", *Quarterly Journal of Economics*, 109(4): 809-839.

Burgess, S., and Ratto, M. (2003), "The Role of Incentives in the Public Sector: Issues and Evidence", *Oxford Review of Economic Policy*, 19(2): 285-300.

Burke, P. J. (1956), "The Output of a Queuing System", *Operations Research*, 4(6): 699-704.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011), "Robust Inference with Multiway Clustering", *Journal of Business and Economic Statistics*, 29(2): 238-249.

Cameron, A. C., and Miller, D. L. (2015), "A Practitioners Guide to Cluster-Robust Inference", *Journal of Human Resources*, 50(2): 317-372.

Carte, T., and Chidambaram, L. (2004), "A Capabilities-Based Theory of Technology Deployment in Diverse Teams: Leapfrogging the Pitfalls of Diversity and Leveraging its Potential with Collaborative Technology", *Journal of the Association for Information Systems*, 5(11-12): 448-471.

Catalini, C. (2017), "Microgeography and the Direction of Inventive Activity", *Management Science*, 64(9): 4348-4364.

Catalini, C., Fons-Rosen, C., and Gaule, P. (2018), "How Do Travel Costs Shape Collaboration?", *National Bureau of Economic Research.* (No. w24780)

Chan, D. C. (2016). "Teamwork and Moral Hazard: Evidence from the Emergency Department." *Journal of Political Economy*, 124(3): 734-770.

Cooper, R.B. (1981), "Introduction to Queueing Theory", *North Holland*.

Crawford, V. P., and Sobel, J. (1982), "Strategic Information Transmission", *Econometrica*, 50(6): 1431-1451.

Curtis, I. (2015), "The Use of Targets in Policing ", available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/466058/Review_Targets_2015.pdf

Dessein, W. (2002), "Authority and Communication in Organizations", *Review of Economic Studies*, 69(4): 811-838.

Dessein, W., Galeotti, A., and Santos, T. (2016), "Rational Inattention and Organizational Focus", *American Economic Review*, 106(6): 1522-36.

Dessein, W., and Santos, T. (2006), "Adaptive Organizations", *Journal of Political Economy*, 114(5): 956-995.

Dewatripont, M. (2006), "Costly Communication and Incentives", *Journal of the European Economic Association*, 4(2-3): 253-268.

Dewatripont, M., Jewitt, I., and Tirole, J. (1999), "The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies", *Review of Economic Studies*, 66(1): 199-217.

Dewatripont, M., and Tirole, J. (2005), "Modes of Communication", *Journal of Political Economy*, 113(6): 1217-1238.

Dodd, T., and Simmons, J. (Eds.) (2002/03), "British Crime Survey", available at http://webarchive.nationalarchives.gov.uk/20110220105210/rds.homeoffice.gov.uk/rds/pdfs2/hosb703.pdf

Drago, R. and Garvey, G. T. (1998), "Incentives for Helping on the Job: Theory and Evidence", *Journal of Labor Economics*, 16(1): 1-25.

Forman, C., and Zeebroeck, N. V. (2012), "From Wires to Partners: How the Internet Has Fostered R&D collaborations within firms" *Management Science*, 58(8): 1549-1568.

Friebel, G., and Raith, M. (2010), "Resource Allocation and Organizational Form", *American Economic Journal: Microeconomics*, 2(2): 1-33.

Gant, J., Ichniowski, C., and Shaw, K. (2002), "Social Capital and Organizational Change in High-Involvement and Traditional Work Organizations", *Journal of Economics and Management Strategy*, 11(2): 289-328.

Garicano, L. (2000), "Hierarchies and the Organization of Knowledge in Production", *Journal of Political Economy*, 108(5): 874-904.

Garicano, L., and Prat, A. (2013), "Organizational Economics with Cognitive Costs", *Advances in Economics and Econometrics*, 1: 342.

Garicano, L., and Rayo, L. (2016), "Why Organizations Fail: Models and Cases". *Journal of Economic Literature*, 54(1): 137-92.

Garicano, L., and Santos, T. (2004), "Referrals", *American Economic Review*, 94(3): 499-525.

Gaynor, M., Rebitzer, J. B., and Taylor, L. J. (2004), "Physician Incentives in Health Maintenance Organizations." *Journal of Political Economy*, 112(4): 915-931.

GMP (2015a), "Command and Control Databases 2008-2015", *Multiple electronic files*, Retrieved October 01, 2015. Provided to the authors under data processing agreement

2014/07/16.

GMP (2015b), "Human Resources Records 2005-2015", *Multiple electronic files*, Retrieved October 11, 2015. Provided to the authors under data processing agreement 2014/07/16.

GMP (2016), "Workforce Management System and OCB Records 2009-2014", *Multiple electronic files*, Retrieved February 24, 2016. Provided to the authors under data processing agreement 2014/07/16.

GMP (2017), "Victim Satisfaction Surveys 2011-2016", *Multiple electronic files*, Retrieved June 06, 2017. Provided to the authors under data processing agreement 2014/07/16.

GMP (2019), "Crime and Incident Data 2009-2014", *Multiple electronic files*, Retrieved July 01, 2019. Provided to the authors under data processing agreement 2014/07/16.

Hall, R. L., and Deardorff, A. V. (2006), "Lobbying as Legislative Subsidy", *American Political Science Review*, 100(1): 69-84.

Hamilton, B. H., Nickerson, J. A. and Owan, H. (2003), "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation", *Journal of Political Economy*, 111(3): 465-497.

Hamilton, B. H., Nickerson, J. A., and Owan, H. (2012), "Diversity and Productivity in Production Teams. In Advances in The Economic Analysis of Participatory and Labor-managed Firms", 99-138. Emerald Group Publishing Limited.

Harris, T. E. (2002), "Applied Organizational Communication: Principles and Pragmatics for Future Practice", Second Edition *Erlbaum*.

Hayek, F. A. (1945), "The Use of Knowledge in Society", *American Economic Review*, 35(4): 519-530.

HMIC (2012) https://www.justiceinspectorates.gov.uk/hmic/media/greater-manchester-response-to-the-funding-challenge.pdf

Holmstrom, B., and Milgrom, P. (1991), "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design", *Journal of Law Econonmics and Organization*, 7: 24-52.

Imbens, G. W., and Wooldridge, J. M. (2009), "Recent Developments in the Econometrics of Program Evaluation", *Journal of Economic Literature*, 47(1): 5-86.

Itoh, H. (1991), "Incentives to Help in Multi-Agent Situations", *Econometrica*, 59(3): 611-636.

Jablin, F. M., Putnam, L. L., Roberts, K. H., and Porter, L. W. (1987), "Handbook of Organizational Communication: An Interdisciplinary Perspective", *Sage Publications*.

Jackson, J. R. (1957), "Networks of Waiting Lines", *Operations Research*, 5(4): 518-521.

Kmetz, J. L. (1998), "The Information Processing Theory of Organization: Managing Technology Accession in Complex Systems", Ashgate

Lang, K. (1986), "A Language Theory of Discrimination", *The Quarterly Journal of Economics*, 101(2): 363-382.

Lazear, E. P. (1999), "Globalisation and The Market for TeamMates", *The Economic Journal*, 109(454): 15-40.

McCain, B. E., O'Reilly, C., and Pfeffer, J. (1983), "The Effects of Departmental Demography on Turnover: The Case of a University", *Academy of Management Journal*, 26(4): 626-641.

Menzel, A. (2019), "Knowledge Exchange and Productivity Spill-overs in Bangladeshi Garment Factories", *Working Paper*.

O'Reilly III, C. A., Caldwell, D. F., and Barnett, W. P. (1989), "Work Group Demography, Social Integration, and Turnover ", *Administrative Science Quarterly*, 34(1): 21-37.

Pace, R. W., and Faules, D. F. (1994), "Organizational Communication"Third Edition *Prentice Hall*.

Palacios-Huerta, I. and Prat, A. (2012), "Measuring the Impact Factor of Agents within an Organization Using Communication Patterns", *Working paper*.

Shannon, C. E. and Weaver, W. (1949), "A Mathematical Model of Communication", University of Illinois Press

Simon, H. A. (1979), "Rational Decision Making in Business Organizations", *American Economic Review*, 69(4): 493-513.

Staples, D. S., and Zhao, L. (2006), "The Effects of Cultural Diversity in Virtual Teams Versus Face-to-Face Teams", *Group Decision and Negotiation*, 15(4): 389-406.

Radner, R. (1993), "The Organization of Decentralized Information Processing", *Econometrica*, 61(5): 1109-1146.

Van Zandt, T. (1999), "Decentralized Information Processing in the Theory of Organizations", in *Contemporary Economic Issues*. Palgrave Macmillan, London.

Zenger, T. R., and Lawrence, B. S. (1989), "Organizational Demography: The Differential Effects of Age and Tenure Distributions on Technical Communication", *Academy of Management Journal*, 32(2): 353-376.

Data Availability Statement

The data underlying this article were obtained from the internal records of the GMP under strict confidentiality Data Use Agreements which prevent them being shared publicly.[34] Researchers interested in accessing these data should refer to the information provided in the replication package. Code for replicating the results in this paper are available on Zenodo at DOI 10.5281/zenodo.4017410.

[34]GMP(2015a,2015b,2016,2017,2019)