Team Players: How Social Skills Improve Team Performance

Ben Weidmann, Harvard University

David J. Deming, Harvard University and NBER[1]

May 2020

Abstract

Most jobs require teamwork. Are some people good team players? In this paper we design and test a new method for identifying individual contributions to team production. We randomly assign people to multiple teams and predict team performance based on previously assessed individual skills. Some people consistently cause their team to exceed its predicted performance. We call these individuals "team players". Team players score significantly higher on a well-established measure of social intelligence, but do not differ across a variety of other dimensions, including IQ, personality, education and gender. *Social skills* – defined as a single latent factor that combines social intelligence scores with the team player effect – improve team performance about as much as IQ. We find suggestive evidence that team players increase effort among teammates.

[1] Emails: david_deming@harvard.edu; benweidmann@g.harvard.edu. We would like to thank Maria Bertling for indispensable help in running the experiment, along with excellent support from Gabe Mansur and the Harvard Decision Science Lab. Thanks also to Doug Staiger, Derek Neal, Angela Duckworth, Steven Sloman, Duncan Watts, Erika Weisz, Mina Cikara and participants in the Cikara lab work-in-progress seminar, the Duckworth Lab, and the E-con of Education seminar for invaluable feedback. Thanks to Victoria Johnson for superb research assistance, and to Brad Duchaine and Karen Sullivan for generously sharing access to materials for the memory task. All errors are our own.

Section 1: Introduction

Teamwork is increasingly important in the modern economy. In 2017, 78 percent of U.S. employment was in occupations where group work was judged either a “very” or “extremely” important part of the job (O*NET, 2020). Employer surveys consistently find that collaboration, communication and ability to work in a team are among the most desired attributes of new hires (e.g. NACE 2019). Since 1980, occupations requiring high levels of social interaction have grown nearly 12 percentage points as a share of all jobs in the U.S. economy, and have experienced faster wage growth at the same time (Deming 2017).

The economic payoff to social skills arises because teams often operate more efficiently than people working in isolation (e.g. Lindbeck and Snower 2000, Hamilton, Nickerson and Owan 2003, Lazear and Shaw 2007, Boning, Ichniowski and Shaw 2007, Bloom and van Reenen 2011, Edmonson 2012). Yet while teamwork skills are highly valuable in principle, in practice it is difficult to isolate individual contributions to team performance. A large literature in economics estimates productivity spillovers across workers and peers (e.g. Falk and Ichino 2006, Mas and Moretti 2009, Arcidiacono et al. 2012, Herbst and Mas 2015, Cornelissen, Dustmann and Schonberg 2017, Feld and Zolitz 2017, Isphording and Zolitz 2019). Yet this evidence is only useful for the relatively small number of jobs in which individual productivity can be reliably measured. In contrast, while there are many studies of the determinants of team success, team-level performance differences are not easily attributed to individual members of the group.[2] How do we know which people are good team players?

In this paper we design and test a new experimental method for identifying individual contributions to team performance.[3] We first assess individual performance on several different tasks. We then randomly assign individuals to multiple teams, and we measure each team’s performance on tasks that are identical or very similar to those that were administered individually. We use the individual scores to generate a

[2] An important exception is Almaatouq, Yin and Watts (2020), who evaluate individual skill and then use it as a mediator to understand variation in group performance. More broadly, a large literature in organizational psychology studies the determinants of effective teamwork. For an overview, see Driskell, Salas and Driskell (2018). Characteristics such as group average IQ, personality, and knowledge and experience of and attitudes toward teamwork are all positively correlated with team performance (Devine and Phillips 2001, Morgeson, Reider and Campion 2005, Bell 2007, Mumford et al. 2008, Driskell, Salas and Hughes 2010). Of particular interest is the literature on “collective intelligence” (CI), which identifies a common factor predicting group performance across a wide range of tasks (Woolley et al. 2010, Engel et al. 2014). Woolley et al. (2010) find that CI is predicted by the group’s average emotional perceptiveness, conversational turn-taking, and the share of the group that is female. However, some recent work has questioned the distinctiveness of CI from other factors such as group average IQ (Barlow and Dennis 2016, Crede and Howardson 2017, Bates and Gupta 2017). Hansen and Vaagan (2016) argue that we are still not close to establishing why some groups perform better than others.

[3] Our experimental design, statistical analysis plan, and main outcomes of interest were pre-registered with the American Economic Association Randomized Controlled Trial registry as AEARCTR-0002896.

prediction for the performance of each team. We then ask whether some teams consistently outperform their prediction when an individual is randomly assigned to them. We call these individuals *team players*.

Team players improve group performance, conditional on their own skill in the task at hand. If we added a chess grand master to a chess-playing team, that person would clearly increase team performance but would not necessarily be a team player by our definition. Instead, team players are individuals who consistently cause their team to produce more than the sum of its parts.

Our first finding is that team players exist. In our pre-registered model, an individual who scores one standard deviation higher on the estimated team player index increases team performance by 0.13 standard deviations. This effect is economically significant and is about 60 percent as large as the impact of individual task-specific skill. We validate the existence of the team player effect by showing that team players improve team performance on a novel, out-of-sample problem-solving task. Our results are robust to a variety of alternative ways of measuring the team player effect and are consistent across task types.

Our second finding is that team players score significantly higher on the Reading the Mind in the Eyes Test (RMET), a well-established and psychometrically valid measure of social intelligence (Baron-Cohen et al. 2001, Adams et al. 2010, Woolley et al. 2010, Baker et al. 2014, Engel et al. 2014). After controlling for task-specific skills, IQ does *not* predict whether someone is a good team player. The team player effect is also uncorrelated with gender, age, education, ethnicity and scores on the “Big 5” personality factors. Each of these tests was part of our pre-analysis plan, and we report all those results in the first part of the paper before moving on to exploratory analyses.

The correlation between social intelligence and the team player effect holds in models that condition on a variety of other individual characteristics. In fact, the RMET alone has more predictive power than all the other characteristics combined. If we treat the team player index and the RMET as two noisy measures of the same construct, that construct – which we will call *social skill* – predicts team performance about as much as IQ. Consistent with the theoretical model in Deming (2017), social skills improve the productivity of teams and thus are more valuable in workplace settings where more teamwork is required.

Our experiment is designed to establish the existence of individual differences in the ability to contribute to team production. Importantly, we show that the skill of being a “team player” is correlated with social intelligence, but independent of general cognitive ability. However, the results are consistent with multiple theoretical models of team production. Many studies treat social or “non-cognitive” skills as additively separable contributions to a skill vector in a Mincerian earnings regression (e.g. Heckman, Stixrud and Urzua 2006, Yamaguchi 2012).[4] In Deming (2017), social skill reduces coordination frictions in team production, which implies that cognitive skill and social skill are complements in a wage equation. We are unable to fully adjudicate between different mechanisms for the impact of being a good team player, including improved communication and integrative thinking, increased allocative efficiency of participants to tasks, and others.

However, we provide two pieces of suggestive evidence that team players increase effort among teammates. First, groups with good team players are more likely to persist on a task and use their full allotment of time, which is positively correlated with team performance. Second, the team player effect holds even when sub-tasks are performed separately by individual team members, with little direct interaction. This suggests that team players might motivate teammates to exert more individual effort. However, we emphasize that the effort channel may operate alongside other mechanisms, which should be the subject of future study.

Our paper makes three main contributions. First, we develop a new methodology for estimating individual contributions to group performance. We show that *repeated* random assignment is necessary to estimate individual contributions to team performance. Additionally, isolating the "team player" effect from other factors requires conditioning on individual skill in closely related tasks. While the lab setting helped us carefully control these conditions, our experimental approach generalizes to the field and to more complicated real-world tasks (Falk and Heckman 2009, Charness and Kuhn 2011). Our work is similar in spirit to the literature in economics which estimates productivity by separately identifying worker and firm effects on wages (e.g. Abowd, Kramarz and Margolis 1999, Card, Heining and Kline 2013, Cornelissen, Dustmann and Schonberg 2017) and the literature on estimating teacher effectiveness (e.g. Rockoff 2004, Kane and Staiger 2006).

Second, we uncover a direct mechanism for the high economic payoff to social skills in the labor market. Workers with higher social skills causally improve team performance, beyond what their individual task-specific skills would suggest. Our findings are consistent with many other studies showing labor market returns to social skills and "non-cognitive" skills (e.g. Kuhn and Weinberger 2005, Heckman, Stixrud and Urzua 2006, Borghans et al. 2008, Almlund et al. 2011, Lindqvist and Vestman 2011, Heckman and Kautz

[4] In Cunha, Heckman and Schennach (2010), there are complementarities in the *development* of cognitive and non-cognitive skills across different stages of the life-cycle. McCann et al (2015) develop a model where individuals endogenously invest in production or communication skills early in life, with those who specialize in communication becoming managers and teacher and everyone else as workers.

2012, Weinberger 2014, Deming 2017). A closely related body of work in economics and psychology finds that prosociality is associated with positive labor market outcomes (Becker et al. 2012, Falk et al. 2019, Kosse et al. 2020). These studies most often estimate wage differences for individuals with different skill endowments but cannot directly link skills to job performance.

Effective teamwork requires individuals to tacitly read and react to their teammates' emotional states, and to adjust their own behavior accordingly. An expert panel judged the RMET to be one of the best measures of the abilities to recognize emotions in adults (Pinkham et al. 2014). Moreover, group average scores on the RMET have been shown to predict team performance across a range of tasks (Woolley et al. 2010).

Our third contribution is practical - large productivity gains are possible for employers who can accurately identify and recruit team players. While our experiment is conducted in a lab, there are several reasons to believe that the results might generalize to more realistic settings. Herbst and Mas (2015) review the literature on productivity spillovers and find that lab and field experiments yield strikingly similar magnitudes. Moreover, other studies provide circumstantial evidence supporting our findings. Woolley et al. (2010) and Engel et al. (2014) find that average social intelligence predicts group performance, while Deming (2017) finds that individual social skills increase earnings and lead to sorting into teamwork-intensive jobs. Several studies highlight the role of individual scientists in team production of research (Azoulay et al. 2010, Oettl 2012, Waldinger 2012, Jaravel, Petkova and Bell 2018). Arcidiacono, Kinsler and Price (2017) and Devereux (2018) estimate individual spillovers onto team performance in professional sports, while many other studies investigate the contribution of teamwork and team-specific capital to team performance (e.g. Wuchty, Jones and Uzzi 2007, Bartel et al. 2014, Chan 2016, Brune, Chyn and Kerwin 2017, Neffke 2019, Park 2019).

Our lab tasks are relatively simple, requiring only basic coordination among teammates, and there is almost no scope for repeated interactions. If anything, the lab results might understate the full impact of being a team player. Nonetheless, we find that the team player effect is about 60 percent as important as individual skills in explaining group performance. We also find that social skills have roughly the same predictive power as IQ for team success. This suggests that the individual assessments used in nearly all educational and employment settings miss a lot of information about worker productivity. To identify good team players, you must measure performance in team settings.

The remainder of the paper proceeds as follows. Section 2 describes the experiment and the data. Section 3 outlines our measurement framework. Section 4 presents our main, pre-registered results. Section 5 explores mechanisms, and Section 6 concludes.

Section 2: Description of experiment and data

2.1 Overview of experiment

Our experiment had two phases, summarized in Figure 1. In the first phase, participants completed a series of online tests to measure their individual skill at three problem-solving tasks: Memory, Optimization, and Shapes. Section 2.2 describes these tasks in detail. We also assessed participants' social intelligence / emotional perceptiveness (using a shortened version of the Reading the Mind in the Eyes Test, described in Baron-Cohen et al., 2001)[5] and personalities (using a short version of the Big 5 inventory, from Goldberg, 1992).

The Reading the Mind in the Eyes Test (RMET) measures participants' ability to recognize emotions in others and, more broadly, their 'theory of mind' (i.e. their ability to reason about the mental state of others, Baron-Cohen et al., 2001). Relative to other measures of social intelligence, the main value of the RMET is that it has right and wrong answers, has relatively high test-retest reliability, and can be quickly and reliably administered (Pinkham et al. 2014). The test presents participants with photos of faces, cropped so that only the eyes are visible (see example in Figure 2). For each set of eyes, participants are asked to choose which emotion, from four options, best describes the person in the image. We made definitions of all the words available via links to an online dictionary.

Lab participants were also assessed on three dimensions of the Big 5 personality inventory that are positively associated with group performance in other studies – Conscientiousness, Extraversion, and Agreeableness (Bell et al. 2007). Conscientiousness is often used as a measure of "non-cognitive" skills in economics and is positively associated with employment and earnings (e.g. Almlund et al. 2011, Heckman and Kautz 2012). We administered the 10-item version of each personality sub-scale, based on Goldberg (1992) and available at IPIP (2018).

The second phase of the experiment focused on testing participants in groups. Participants came to the lab and were randomly assigned to groups of three people. Each group completed a collective version of

[5] To limit the length of our test battery, we included 26 of the 36 original items. The items we removed were an equal balance of male and female faces.

the individual problem-solving tasks: Memory, Optimization and Shapes. Participants visited the lab twice. During each visit, participants worked in two separate groups. Over the course of the experiment, participants were allocated to a total of four groups. The average time between the individual assessment and the first lab visit was 11 days. The average time between the first and second lab visit was 11 days.

To ensure that participants never saw the same problem twice, we incorporated five different versions of each task type and deterministically grouped them into "batteries" A through E, as shown in Figure 1.

Each group of three people worked face-to-face in a single room. The tasks were computer based and each participant was provided with a laptop. We began each group session by asking group members to introduce themselves. Then, groups were required to nominate a 'Reporter'. The Reporter was responsible for entering their group's answers. Participants also gathered around the Reporter's laptop for some tasks.[6] Before each problem-solving task, groups were prompted to discuss their strategy. In batteries B and D groups completed practice versions of each task.

2.2 Individual and Group Tasks

We chose tasks to satisfy three criteria. First, tasks must be feasible to administer to both individuals and groups, with only minor modifications between the individual and group versions. This enabled us to estimate group performance *controlling for individual task-specific skill.* Second, tasks needed to be objective in the sense that we could easily rank performance across individuals and groups. Third, since we are interested in studying teamwork, we looked for tasks where cooperation among group members would plausibly improve performance. The three tasks we use to estimate our "team player" effects - Optimization, Memory and Shapes - meet each of these three criteria. This section describes the individual and group versions of these problem-solving tasks.

Optimization Task

The goal of this task was to find the maximum of a complex function.[7] Some example functions are presented in Figure 3 (left panel). In the individual Optimization task participants were given a function,

---

[6] We deliberately framed the role of the Reporter as one in which people follow a 'collaborative' rather than a 'consultative' approach to help facilitate teamwork (Curseu et al. 2013). In a pre-specified secondary analysis we examined whether there was a relationship between being nominated as a reporter and the Team player index. We found no evidence of an association between the team player index and whether someone was nominated to be the Reporter for the group.

[7] We developed the Optimization task specifically for the purposes of this experiment. We were inspired by Mason et al. (2008), who use a numerical optimization task to study how innovations propagate across networks. The individual task was piloted in a MTurk sample.

which was hidden to them, and had 15 guesses to find the maximum. They entered guesses between 0 and 300. For example, a participant attempting to find the maximum of function b (f_b) would see the interface presented in the right panel of Figure 3. For each guess, the computer returns $f_b(guess)$. Once participants had entered 15 guesses, they were asked to submit their answer for the input value that maximized the output. In Battery A, individuals completed the Optimization task three times. A different underlying function was used each time.

In the group version of the task, each group member was allocated 5 guesses. Collectively, the group had a total of 15 guesses. Each group member entered their own guesses on their own laptop.

A critical feature of this task was the need to involve all three group members. After the group had entered its 15 guesses, the Reporter was asked to enter the group's answer for the output-maximizing input. Each group solved the Optimization task twice. Every time participants attempted the Optimization task, they engaged with a new underlying function. Success on the group Optimization task required collective planning and the sharing of unique information. Both these factors have been shown in previous small-group research to predict group performance across a range of contexts (Driskell et al., 2018; Mesmer-Magnus & DeChurch, 2009; Weingart, 1992).

Memory Task

This task focused on short-term memory, which is closely associated with fluid intelligence / IQ (Colom et al., 2006; Nisbett et al., 2012). We tested participants' ability to memorize three different types of stimuli: words, images and stories.[8]

In Phase 1 of the experiment, individuals' short-term memory for each type of stimuli was measured sequentially. Participants began by completing the words test. This involved memorizing a list of 12 target words over 24 seconds (the stimuli come from the Hopkins Verbal Learning Test, reported in Brandt, 1991). After the memorization period, participants were presented with sets of three words and were asked to identify which, if any of the three, were target words. Next, participants completed the images test, in which they were given 20 seconds to memorize six target faces (the stimuli come from the

[8] We drew on a model of memory that emphasizes three subsystems: verbal, visualspatial and episodic (Baddeley, 2001). Our three stimuli map onto these subsystems: verbal → words; visualspatial → images; episodic → stories. We note that the Baddeley model focuses on working memory, not short-term memory. The two concepts, however, are very closely linked, as discussed in Colom et al. (2006). The reason we focus on short-term memory is that the subtests are easier to translate into a practical task for groups to perform when working face-to-face in a lab setting.

Cambridge Face Memory Test, described in Duchaine & Nakayama, 2006). Participants were then presented with 15 sets of three faces and asked to identify target faces. Last, participants completed the stories test in which they had 40 seconds to read two short paragraphs, of roughly 60 words each. The stimuli were adapted from Wechsler Logical Memory III (Wechsler, 1997). At the end of the memorization period, participants were asked nine multiple choice questions about the two paragraphs.

Once participants had completed the three individual memory tests, we provided feedback about their results. This included information on an individual's overall performance relative to other participants and emphasized the test on which they scored highest. Our goal with the feedback was to provide people with information they might use in the group phase of the experiment to select sub-tasks on which they were most proficient.

In the group version of the task, we combined established measures of individual memory into a collaborative memory challenge. Each group was given 40 seconds to collectively remember 12 words, 6 images, and 2 stories. We added story and images stimuli to those described above, so that each time a group encountered the Memory task they were asked to memorize unseen material.[9]

Each member of the group viewed their own laptop and could view any of the three stimuli. Participants could change the stimuli they were memorizing during the 40 second memorization period. In the example presented in Figure 4, participant A is memorizing images (cars), participant B is memorizing stories, and participant C is memorizing words. During the 40 second memorization period, participants could change the stimulus they were viewing at any time by using the buttons in the top left of their screens. Before the memorization period began, groups were prompted to discuss their strategy.

After the memorization period, all three team members gathered around the Reporter's laptop to answer a set of 24 questions about the stimuli. There were an equal number of questions about each type of stimuli. The structure of the questions mirrored those used in the individual assessments.

Shapes Task

This task relied on two well-established measures of fluid intelligence: the Culture Fair Intelligence Test (CFIT, Scale 3) and the Raven's Advanced Progressive Matrices (Ravens). In the individual testing phase, participants completed 14 Ravens items (even numbered items, ranging in difficulty from across sets I and

[9] We supplemented stories with shortened versions from Sullivan (2005). For images, we added related tests focused on cars, bikes and bodies, described in Dennett et al. (2012).

II; see Raven, 2003).

This task centers on pattern recognition and spatial reasoning. Participants are asked to look for a pattern and determine 'what comes next'. As an example, consider the pattern established in the left-most box of Figure 5, which has missing a piece. Participants were asked to find the missing element (from options a to f).

The group version of this task employed the CFIT, which is very similar to the Ravens task. All group members gathered around the Reporter's laptop and collectively decided on the group's answer for each item. In each battery contained a different form of the CFIT. An example item is provided in the right-hand panel of Figure 5.

Validation Task: Cryptography

We use the three tasks above to estimate individual contributions to group performance, as described in Section 3. We chose the Cryptography task as a fourth, out-of-sample validation measure of group performance. The Cryptography task is a decoding problem in which each letter from A to J represents a unique number from 0 to 9. Groups were asked to decode the value of each letter by entering mathematical expressions that would return an output (e.g. if A=5, B=1, C=4, and D=0, an entry of A+B+C would return the value "BD", for 10). An example is shown in Figure 6.

The procedure for decoding each letter is somewhat complex and is well described elsewhere (Larson, 2010).[10] The goal of the task is to find the value of each letter in the fewest number of steps. We administered this task twice: once as a practice, to make sure that groups understood the process – and a second time to assess their performance. Cryptography is one of the very few established tasks that demonstrates 'strong synergy' in the sense that groups perform better than the sum of their parts (Larson, 2010).[11] This task was only administered in Battery E, the last set of group tasks.

[10] In brief, the process involved three steps. *Step 1: enter an 'equation'*. An equation is a set of letters with '+' or '–' operators; e.g. A+B+C. The computer then returned the answer. If A=3, B=1, C=2, D=6, then the computer would reply A+B+C=D. *Step 2: make a hypothesis*. Here, a group might guess that D was a large number (as it's the sum of 3 numbers). So, they might guess "D=7". The computer would reply "FALSE". *Step 3: guess all the values*. The group is allowed, but not compelled, to submit a value of each letter. If all their guesses are correct, the task ends. If not, the group goes back to step 1.

[11] The reason may be that the task naturally lends itself to people taking on different roles. While some people are figuring out what the next equation should be according to the current strategy, others can consider better strategies. This gives groups the potential to be strategically flexible. Individuals, on the other hand, find it extremely challenging to simultaneously execute a strategy *and* to consider a new one, perhaps due to constraints of attention and working memory (Larson, 2010, p. 154). See also Laughlin et al. (2006) and Laughlin et al. (2002). Note that the underlying feature of the task that enables differentiation is that it is possible to switch strategies at any point in the task, without incurring a cost.

2.3 Recruitment and Sample

We recruited our sample from the Harvard Decision Science Lab participant pool. The pool comprised 41% undergraduate students and 25% graduate students. There were two exclusion criteria: participants needed to be under 60 years old and fluent in English. These restrictions were based on several pilot sessions and were intended to minimize the risk of floor effects with our tasks. Participants who completed the study were paid a total of $100: $10 for completing the individual tests; $30 for lab visit 1; and $60 for lab visit 2. We elected not to explicitly pay groups for better performance, relying instead on intrinsic motivation through priming.

The number of participants in each stage of the experiment is shown in Figure 7, which presents a simple participant flow diagram. 434 participants successfully completed the individual tests.[12] 332 of these participants attended the first lab session. Of these, 274 came to a second lab session. We excluded groups from the final analysis for three reasons: groups that did not have three people,[13] groups in which a participant wasn't eligible for the study based on their responses to the individual tests,[14] and groups who did not complete the full battery of tasks due to technical issues. The final sample was made up of participants who were observed in at least 3 successful groups. This was the case for 255 of the 274 people who participated in two lab sessions.[15]

Table 1 shows descriptive statistics of the study population. Relative to the US, our final analysis sample was younger, more female and less white.

2.4 Randomization

Lab sessions typically consisted of n=9 or n=12 participants.[16] After participants completed the individual tasks, they were immediately eligible to sign-up for group sessions. Lab staff contacted participants and created time slots in which nine or 12 participants were available. When participants did not divide evenly into three-person groups due to no-shows, "extra" participants were paid a small show-up fee and asked

[12] In addition to these 434 participants, 152 people completed some part, but not all, of the online tasks. 9 participants completed all the tasks but were ineligible for the study, based on a series of attention checks. These were included because participants completed the individual tasks outside of the lab, and we wanted to ensure that participants carefully read the instructions.
[13] This happened when an incorrect group ID was used unnoticed by the researchers.
[14] Two participants who were not eligible nonetheless managed to attend a lab session. We removed all affected groups.
[15] 217 participants were observed in four groups, with the remaining 38 being observed in three groups.
[16] From a total of 343 groups, 147 were formed in sessions of 12 participants; 145 were formed in sessions of 9; 51 were formed in sessions of 6.

to come back again. Lab sessions were evenly spaced between August and November of 2019, and participants signed up 6 days ahead on average.

The session signup process was haphazard, but not explicitly randomized. However, there is no clear pattern of average participant scores on the RMET or Ravens over time. Reassuringly, when we compare group means in terms of average Ravens or RMET scores, we fail to reject the null hypothesis that they are equal (p=0.72 and p=0.30 respectively).

We randomly assigned participants to teams, but with two deviations from simple randomization. First, we wanted to maximize the importance of any team player effect by creating groups with similar levels of individual skill. Second, we wanted to minimize scenarios where the same people worked together multiple times, so that our results would not be contaminated by familiarity among teammates.

At the start of each session, we conducted a blocked randomization procedure. Participants were ordered according to their mean performance across the individual problem-solving tasks (Memory, Optimization and Shapes). From this ordering, we formed three blocks: higher-skill; medium-skill; lower-skill. Each block had the same number of participants, $\frac{n}{3}$.

We then randomly generated groups of three people. Each group had a member from each block. Participants randomly drew balls from bags, under the supervision of the experimenters. For example, in a session with n=9 participants, groups from the higher-skill block each randomly drew a ball from the set {A,B,C}; participants from the medium-skill block randomly drew a selection from the set {D,E,F}, and the participants from the lower-skilled block randomly drew a ball from the set {G,H,I}. During each lab session, participants were randomly assigned to two groups. Each participant's randomly-assigned letter defined both of their groups. Figure 8 provides an example of the randomization scheme.

Section 3: Measurement Model

This section describes our conceptual framework and empirical approach. Our analysis strategy was pre-registered at the AEA RCT registry.[17] Deviations are noted in the footnotes.

Let individuals be indexed by $i = 1, \dots, n$. We allocated individuals to groups of three people, where groups are indexed by g. We ultimately observed $n_g = 343$ groups in our final sample. Let I_g^i be an indicator of whether participant i is in group g. I_g^i is a vector of length n_g, where:

[17] The trial number is AEARCTR-0002896.

$$I_g^i = \begin{cases} 1 & if\ i\ is\ in\ g \\ 0 & otherwise \end{cases}$$

Next, we have a set of variables describing task performance. Let X_{ik} denote the performance of individual i on task type k. Similarly, let G_{gk} denote the performance of group g on task $k \in$ {Optimization; Memory; Shapes}. We rescale group scores G_{gk} for each task to account for potential differences in task difficulty. Let b indicate task battery, $b \in \{B, C, D, E\}$.

Rescaled scores are calculated as: $\tilde{G}_{gkb} = \frac{G_{gkb} - \hat{\mu}_{kb}}{\hat{\sigma}_{kb}}$ where $\hat{\mu}_{kb}$ and $\hat{\sigma}_{kb}$ are the sample mean and standard deviation for task k in battery b.[18] $\tilde{G}_{gk}$ is our main measure of group performance.

Some groups may perform better on tasks purely because groups have different endowments of task-specific skill. Thus our definition of whether someone is a good team player explicitly conditions on individual skill in the task at hand. Consider the following model for how well group g performs on task k:

$$\tilde{G}_{gk} = \alpha_k \sum_i I_g^i X_{ik} + \gamma_1 D_g^{Knew} + \gamma_2 D_g^{Prev} + \epsilon_{gk} \qquad (1)$$
$$\epsilon_{gk} \sim N(0, \sigma_G^2)$$

The term $\alpha_k \sum_i I_g^i X_{ik}$ measures group g's endowment of individual skill on task type k. D_g^{Knew} is an indicator for whether group g contained participants who were acquaintances, friends or colleagues outside the context of the experiment. 95 percent of teams were comprised entirely of participants who had never met. D_g^{Prev} indicates whether group g contained participants who had previously been allocated to the same group in the experiment. 59 percent of teams were comprised entirely of individuals who had never been in the same group in a previous round of the experiment.

Define T_g as a measure of *group level* performance, adjusted for differences in individual task-specific skill. The residuals $\hat{\epsilon}_{gk}$ from equation (1) above provide an estimate of whether each group under- or over-performed on task k relative to the prediction based on task-specific skills. Averaging this residual performance across tasks gives us:

$$\hat{T}_g = \frac{1}{3} \sum_k \hat{\epsilon}_{gk} \qquad (2)$$

[18] After rescaling, we suppress the b subscript for clarity.

With only a single randomization, it is impossible to determine whether variation in $\hat{T}_g$ arises from unmeasured *individual* attributes of particular team members, or from group dynamics between team members. However, with repeated random assignment, we can assess whether $\hat{T}_g$ is correlated for individuals as they join different teams. For each participant, we estimate the *team player index* $\hat{\beta}_i$ as the average $\hat{T}_g$ across all groups that i participated in (up to 4):

$$\hat{\beta}_i = \frac{1}{4}\sum_g I_g^i \hat{T}_g \quad (3)$$

In our framework, $\hat{\beta}_i$ is an estimate of the causal contribution of individual i to team performance. With enough randomizations, we could precisely estimate β_i for each participant. However, with only four team assignments, $\hat{\beta}_i$ is relatively noisy at the individual level. Thus, following our pre-analysis plan, our main focus is σ_β, the standard deviation of the β estimates. We estimate σ_β using a multilevel model.[19]

$$\hat{T}_{gi} = \beta_i + e_{gi} \quad (4)$$
$$\beta_i \sim N(0, \sigma_\beta^2)$$
$$e_{gi} \sim N(0, \sigma^2)$$

Where $\hat{T}_{gi}$ is a vector of skill-adjusted group performance (1x$3n_g$), β_i is a random effect for individual i on group g and e_{gi} is residual error.

We evaluate our results against the null hypothesis that the team player effect – conditional on individual skill, as in equation (1) – is equal to zero. Our preferred approach to hypothesis testing is randomization inference. Specifically, we randomly simulate five thousand allocations of individuals to groups, blocking on task battery so that in every simulated allocation we observe each participant the same number of times as we did in the actual experiment. For each simulated allocation we fit model (4) and estimate $\sigma^2_{\beta(NULL)}$. We then compare our observed team player effect to the simulated distribution under the null, calculating how often the null distribution provides a more extreme value than $\hat{\sigma}_\beta$, i.e. $P(\sigma_{\beta(NULL)} > \hat{\sigma}_\beta)$. This is our p-value (Davison & Hinkley, 1997).

[19] This represents a deviation from our pre-registered analysis plan, in which we planned to estimate σ_β^2 as $\hat{\sigma}_\beta^2 = \widehat{var}(\boldsymbol{\hat{\beta}})$, where $\boldsymbol{\hat{\beta}}$ is a (1xN) vector of team player estimates from equation (3). However, $\widehat{var}(\boldsymbol{\hat{\beta}}) = \hat{\sigma}_\beta^2$. Reporting this figure would overstate the magnitude of the team player effect, which motivates the use of model (4). Appendix Figure A1 shows that the distribution of the raw $\hat{\beta}_i$'s from equation (3) closely matches a normal approximation, which justifies the normality assumption.

While randomization inference is our preferred approach, as a robustness check we also generate p-values for two widely-used methods to estimate uncertainty: Wald, using a normal approximation, and profile likelihood. Since we are interested in the variance parameter σ_β^2, which is non-negative and thus not symmetric around the point estimate, the normal approximation may provide a poor fit. Profile likelihood confidence intervals are based on the chi-squared distribution of the log likelihood ratio test statistic, and thus may be a better fit in this small, non-normal sample (Venson and Moolgavkar 1988).[20]

Section 4: Main Results

In this section we report results from our pre-specified models only, and we explicitly note any deviation from the pre-analysis plan. Section 5 reports post-hoc exploratory analyses and evidence for mechanisms.

4.1 Are some people good team players?

Table 2 reports estimates of the team player effect – the standard deviation of the team player index $\hat{\sigma}_\beta$ – where only the method of hypothesis testing varies by column. Below the estimate of the team player effect we report p-values for each inference method. Coefficient standard errors are presented in parentheses below control variables.

In our pre-registered model, the team player effect is 0.127 standard deviations. The coefficients on each of the task-specific skills in $\sum X_{ik}$ are highly statistically significant (p<0.01), suggesting that a team's endowment of task-specific skill is a strong predictor of group performance. The average magnitude of the task coefficients is about 0.2, which suggests that the contribution of the team player effect (0.127) is worth about 60 percent as much as individual task-specific skill. The exact magnitudes, however, depend on the degree of measurement error in each measure.[21]

Column 1 reports results from our randomization inference approach. The actual team player estimate exceeds the simulated estimates in more than 97 percent of the cases (p=0.029). The p-value is similar when we compute confidence intervals using profile likelihood in Column 2 (p=0.037). The standard normal approximation (Column 3) yields much tighter confidence intervals (p<0.001).

[20] These alternative approaches to estimating uncertainty were not included in our pre-analysis plan.

[21] Disattenuation of the measurement error for the coefficients of $\sum X_{ik}$ requires estimates of the reliability of both group scores and the tests of task-specific individual skill. Because our tasks are novel and were largely developed for the purpose of this experiment, there is considerable uncertainty in these reliability estimates. However, assuming a standard test-retest reliability of 80 percent for both $\tilde{G}_{gk}$ and $\sum X_{ik}$ would revise the average coefficient on $\sum X_{ik}$ upward from 0.199 to 0.249. In that case, a one standard deviation increase in the team player effect is worth just under half as much as an equivalent improvement in task-specific skill. On the other hand, the team player effect is also measured with error.

Table 3 presents results from alternative estimates of the team player effect, varying the set of variables included as controls in equation (1). In each case we compute p-values using randomization inference. Column 1 repeats the main pre-registered results from Table 2. Column 2 replaces task-specific controls (X_{ik}) with controls for the group's average IQ, as measured by the Ravens test. Controlling for general IQ rather than task-specific skills increases the magnitude of the team player effect from 0.127 to 0.173.

Column 3 estimates the team player effect with no controls at all, which increases the magnitude of $\hat{\sigma}_\beta$ to 0.244. We do not think of our estimate of $\hat{\sigma}_\beta$ in column 3 as the impact of being a good team player, but rather the *total* causal impact on group performance of receiving a more talented teammate. In that sense, comparing column 1 to column 3 suggests that about half of the variance in an individual's measured causal contribution to the team can be explained by their individual skill in the task at hand.

Column 4 adds controls for each group's average on three of the "Big 5" personality factors that might be associated with being a good team player – Agreeableness, Conscientiousness, and Extraversion. None of these self-reported personality characteristics predict group performance independently, and thus the magnitude of $\hat{\sigma}_\beta$ remains very similar to the unconditional estimate in Column 3 (0.237 compared to 0.244). Column 5 adds controls for the group's average score on the RMET. The RMET strongly predicts group performance, and the magnitude of the coefficient (0.119) is about 75 percent as large as the coefficient on group average IQ in column 2. Controlling for the RMET reduces the team player effect size to 0.192.

Appendix Table A1 presents additional results where we estimate group performance using a variety of flexible functions of task-specific skills, IQ and the RMET as controls for task-specific skills. To be clear, Column 1 is our preferred, pre-registered estimate of the team player effect as we have defined it. It is not obvious that one wants to include all possible covariates – for example, if we think of the RMET and $\hat{\beta}_i$ as noisy individual measures of the same underlying construct, we would not want to control for the RMET when estimating the magnitude of σ_β. Nonetheless, the bottom line is that none of our results are sensitive to different specifications of group average skills. In a fully flexible "kitchen sink" prediction of group performance, we still estimate a team player effect of 0.120 (p=0.045).

We validate the existence of the individual team player index $\hat{\beta}_i$ by asking whether it predicts group-level performance out-of-sample on the Cryptography task, which is described in Section 2.2. We chose the Cryptography task for out-of-sample validation because the literature on teamwork has shown that it rewards teamwork in the sense that groups typically perform better than the sum of their parts (Larson 2010).[22] 85 groups completed the Cryptography task, always in the last task battery (Battery E).

Unlike the other three group tasks, Cryptography performance was not used to estimate the team player effects shown in Tables 2 and 3 above. Rather, we regress the group's Cryptography score on the group's average team player index $\bar{\hat{\beta}}_i$, as specified in our pre-analysis plan. We also estimate the correlation between Cryptography task performance and other group characteristics such as average IQ and average RMET, as well as other combinations such as the maximum or minimum of each measure. We do this because it is unclear how individual team player effects $\hat{\beta}_i$ should aggregate up to predict group performance.

The results from the Cryptography task validation are in Table 4. Column 1 presents results from our pre-specified model, a bivariate correlation between Cryptography task performance and the team's average team player index $\bar{\hat{\beta}}_i$. The correlation is positive – a 1 standard deviation increase in $\bar{\hat{\beta}}_i$ increases Cryptography task performance by 0.153 standard deviations – but it is also somewhat noisy, and not statistically significant (p=0.162).

Columns 2 and 3 show the same results, except with $\hat{\beta}_{max}$ and $\hat{\beta}_{min}$ as the predictors respectively. The maximum team player index in a group predicts similarly well to the mean (0.165, p=0.131) while the minimum performs worse (0.092, p=0.402). In an exploratory analysis, Column 4 reports the correlation between having a good team player (defined somewhat arbitrarily as someone whose score is >1σ above average) and a group's cryptography score. This is a relatively strong predictor (0.247, p=0.023).

Column 5 shows the correlation between a group's average RMET score and Cryptography task performance. A 1 standard deviation increase in mean RMET score increases performance by 0.192 standard deviations, which is statistically significant at the 10 percent level (p=0.08). Column 6 shows the same result but for the group's average IQ. The coefficient on average IQ is 0.217 (p=0.047), making it a modestly stronger predictor of group performance than average RMET and $\bar{\hat{\beta}}$.[23] However, we cannot reject the joint hypothesis that $\bar{\hat{\beta}}$, $\overline{RMET}$, and $\overline{IQ}$ are all equal to each other in their predictiveness. In

---

[22] According to Larson (2010), the primary advantage of group work in the Cryptography task is that some team members can execute the current strategy (e.g. figuring out what the next equation should be given the output of the current equation) while others can consider new strategies. It is extremely challenging for individuals to simultaneously execute a strategy and consider a new one, perhaps due to attention and working memory constraints (Larson 2010).

[23] We do not report results for the maximum and minimum of IQ or RMET, but we generally find that the maximum and the mean perform about as well, while the minima are worse predictors of group performance.

Section 5.1 we explore combining these measures together to maximize the predictive power of individual characteristics for group performance.

4.2 What predicts being a good team player?

We next turn to the predictors of being a good team player. Following our pre-analysis plan, we estimate correlations between an individual's team player index $\hat{\beta}_i$ and the following individual characteristics:

1. Intelligence, as measured by the Ravens test;
2. Emotional perceptiveness, as measured by the RMET;
3. Personality traits, as measured by Agreeableness, Conscientiousness and Extraversion on the Big 5 scale; and
4. Demographic characteristics – years of education, age, gender and ethnicity.

In all of these models, we define each individual's team player index $\hat{\beta}_i$ according to the pre-registered model described in section 3, in which group scores are conditioned on each group's endowment of individual skill, along with indicators of previous familiarity or group membership.

Figures 9 and 10 present a series of scatterplots, where each dot is an individual, the team player index $\hat{\beta}_i$ is on the vertical axis, and the predictors in 1 through 3 above are listed separately on each X axis. Below each figure, we show the bivariate correlations and p-values from our pre-specified model estimate of $\hat{\beta}_i$ (the first column), and a model showing the total effect with no controls (second column). The scatterplot always shows results from the first row, our pre-specified model.

Figure 9 shows results for IQ (left panel). There is no significant association between being a good team player and IQ ($\hat{\rho}$=0.050, p=0.425) in our baseline pre-registered model. However, IQ is strongly correlated with $\hat{\beta}_i$ when group performance is estimated with no controls ($\hat{\rho}$=0.379, p<0.001). An individual's *total* causal impact on group performance is strongly related to IQ, but all of the impact of IQ is mediated by individual skill in tasks.

The right panel of Figure 9 shows the correlation between $\hat{\beta}_i$ and an individual's score on the RMET. We find a clear and statistically significant correlation between being a good team player and emotional perceptiveness ($\hat{\rho}$=0.166, p=0.008). This contrasts notably with IQ and suggests that RMET adds substantial predictive power beyond the impact of individual task-specific skill. The correlation between RMET and $\hat{\beta}_i$ is even stronger when group performance is estimated with no controls ($\hat{\rho}$ =0.306, p<0.001).

Figure 10 show an analogous set of results for Agreeableness, Conscientiousness and Extraversion respectively. The bottom line is that none of these Big 5 Personality factors are correlated with being a good team player, either in our pre-specified models that control for task-specific skill or in unconditional models. Personality scores appear to have no impact on group performance.

Finally, Table 5 shows correlations between the team player index and demographic characteristics. Columns 1 shows results by gender, column 2 by years of education, column 3 by age and column 4 by race and ethnicity.

We find no evidence of gender differences in $\hat{\beta}_i$. This contrasts somewhat with Woolley et al. (2010), who find that teams with more women perform better on group tasks. In Woolley et al. (2010), gender differences in group performance were mediated by gender differences in performance on the RMET. However, we find only a small difference in RMET scores by gender, with women scoring 18.7 on average compared to 18.0 for men, a difference that is on the margin of statistical significance (p=0.11).

Column 2 presents the correlation between the team player effect and years of completed education. We find no association. Column 3 shows the correlation between age and the team player index. There is a negative association between age and group performance, but it is small and not significantly different from zero ($\hat{\rho}$=-0.084, p=0.18). Column 4 shows the correlation between ethnic minority status – self-identifying as African American, Latino/Hispanic and Native American – and the team player index. We find no association.

Appendix Tables A2 through A4 report additional results from our pre-analysis plan. Table A2 shows that there is no association between the team player index and whether a participant was the group reporter. Table A3 shows that the results for the team player index are not driven by one particular task type. Table A4 investigates the role of group diversity for team performance. We find no evidence for associations between team performance and diversity in skills, age, gender or ethnicity.

Overall, we find strong evidence that – holding task-specific individual skills constant – some people consistently improve their group's performance. These individuals are good "team players". The one characteristic that consistently predicts who will be a good team player is that individual's score on the RMET, a well-established test of social intelligence and emotional perceptiveness. Other measures such as IQ and personality scores do not predict whether someone is a good team player. In Section 5, we test a number of potential explanations for these results.

Section 5: Mechanisms

The results are consistent with several different potential mechanisms, and our experiment was not designed to disentangle them. Team players might increase effort among teammates, improve communication about comparative advantage in team production, or encourage better group problem-solving through improved dialogue and integrative thinking. Moreover, these explanations are not mutually exclusive, and team players may improve group performance in multiple ways.

5.1 Social Skills

Across all individual characteristics, only the Reading the Mind in the Eyes Test (RMET) score predicts whether someone is a good team player. The RMET was originally designed by autism researchers to diagnose deficits in the capability to reason about the mental state of others (Baron-Cohen et al, 2001). However, it also has strong psychometric properties in a general, non-impaired population and has become a well-established measure of social intelligence (Baron-Cohen et al. 2001, Adams et al. 2010, Baker et al. 2014, Olderbak et al. 2015).

Since effective teamwork probably requires individuals to read their teammates' emotional states, it is sensible that the RMET would predict whether someone is a good team player. More directly, group average scores on the RMET have been shown to predict team performance across a range of tasks (Woolley et al. 2010, Engel et al. 2014).

In Table 6 we test whether the correlation between the RMET and the team player index $\hat{\beta}_i$ persists after controlling for an increasingly rich set of individual characteristics. Column 1 presents results from the same model as in Figure 9 (right panel) where $\hat{\beta}_i$ is estimated after controlling for skill in each individual task. The bivariate correlation is 0.166 and is statistically significant at the less than 1 percent level (p=0.008). Column 2 adds demographic controls, including gender age and polynomials of years of completed education, which barely changes the results. Column 3 also adds individual IQ and personality scores. If anything, the correlation between RMET and $\hat{\beta}_i$ gets stronger. Notably, the highest adjusted r-squared is the model that *only* contains RMET. Column (4) drops RMET from an otherwise fully saturated model, removing any explanatory power we have.

The results in Table 6 suggest that the team player index $\hat{\beta}_i$ and the RMET might both measure the same underlying construct, which we will call *social skills* (S_i). To explore this construct, we take the average of each individual's team player effect $\hat{\beta}_i$ and their RMET score[24], and test the predictive power of this composite measure $\hat{S}_i$ on group performance in the Cryptography task.

The results are in Table 7. The first column repeats the results from Column 5 of Table 4, which shows the bivariate correlation between a Cryptography task performance and the group's mean score on the RMET (0.192, p=0.08). Column 2 substitutes the RMET with the group's average score on the social skills composite $\bar{\hat{S}}_i$. A one standard deviation increase in $\bar{\hat{S}}_i$ predicts an increase of 0.218 standard deviations on the Cryptography task (p=0.05). Column 3 replaces $\bar{\hat{S}}_i$ with $\overline{IQ_i}$. The coefficient on group average IQ is 0.217 (p=0.05), which suggests that average IQ and average social skills are about equally predictive of group performance on an out-of-sample validation task.

Column 4 includes both together in a "horse race" specification, $\bar{\hat{S}}_i$ and $\overline{IQ_i}$ are still about equally predictive when included in the same specification, and adding each of them increases the R-squared of the prediction by about 60 percent compared to a model that only includes one. This suggests that each makes an independent contribution to predicting group performance. Column 5 adds group means of personality scores and demographic characteristics. None of these dimensions are individually significant. Column 6 substitutes group means with maximums, which yields qualitatively similar results. The bottom line is that social skills have substantial predictive power for group performance, about the same as IQ.

5.2 Effort and Motivation

Good team players may increase group performance by encouraging their teammates to increase effort. We test this in two ways. First, we examine group-level variation in one measure of effort – whether the team used their full allotted time for a task.

We study group time use in the Shapes task. No group in our study got a perfect score on the Shapes task, in part because the time limit was only 3 or 4 minutes and many of the puzzles were cognitively challenging. For this reason, more than 82 percent of groups took their full allotment of time. Of the groups that finished before time, 10 percent "rushed", which we define arbitrarily as groups who submitted answers with more than 15 seconds to spare.[25]

Overall there is a negative association between "rushing" and performance on the Shapes task. Groups that "rushed" answered 54 percent of items correctly, compared to 63 percent among non-rushed groups,

[24] To avoid double-counting RMET, we use $\hat{\beta}_i$ from a model that conditions on group average RMET.
[25] Our results are not sensitive to other cutoffs such as 0, 5, 10 or 30 seconds.

a difference that is highly statistically significant (p=0.005). Groups that had a "good team player" – which we define (again, arbitrarily) as an individual with $\hat{\beta}_i > 1\sigma$ above the average – rushed only 6 percent of the time, compared to 12 percent for all other groups (p=0.04).[26] We also find that group average endowments of $\hat{\beta}_i$, RMET and $\hat{S}_i$ are negatively associated with rushing, as is the group's average score on the Conscientiousness personality factor (even though it is not related to overall performance). This provides suggestive evidence that team players encourage their group to exert more effort on the Shapes task.

Our second test for the importance of effort as a mechanism involves the impact of social skills on group performance in the Memory task, where group performance is almost always the sum of individual contributions. Recall that in the Memory task there were three types of stimuli – words, images and stories. Although group answers were recorded on a single laptop at the end of the task, each member had their own laptop during the memorization period. An intuitive way to approach the group memory task was for each member to bear responsibility for one of the three stimuli. In fact, 92 percent of our groups adopted this strategy, where each member only looked at a single stimulus category for the entire memorization period.

In cases where each group member memorized stimuli that their teammates didn't see, we can measure whether good team players improve their teammates' performance *despite not being directly involved in the sub-task.* We estimate the impact of being randomly assigned to a teammate with high social skills using the following model:

$$G_{gt[i]} = \alpha + \delta\overline{SRMET_{g[-\iota]}} + \phi X_{ti} + \gamma Round_g + \lambda Type_t + \epsilon_{tg[i]} \qquad (5)$$

$G_{gt[i]}$ is group g's average score on sub-task t (words, images, stories) of the Memory task, for groups where only one person looked at each stimulus type – which allows us to attribute the score to individual i (in brackets). $\overline{SRMET}_{g[-i]}$ is the mean social skills of individual i's randomly assigned teammates, leaving out i. X_{ti} is i's individual score on the memory sub-task t, assessed during the phase one of the experiment. $ROUND_g$ and $MemType_t$ are controls for the ordering of the task batteries and the type of memory task respectively, to remove practice effects and control for baseline differences in memory task difficulty. δ gives the spillover effect of being assigned to a group with one standard deviation higher mean

[26] These results are robust to other cutoffs such as 0.5σ, and to using other individual measures such as the RMET or $\widehat{S_\iota}$, the social skills composite.

RMET score. Because the team player effect $\hat{\beta}_i$ is defined by being in groups that outperform expectations, we use only the RMET as a measure of social skill, to ensure that our results are not mechanical.

Column 1 of Table 8 shows baseline estimates of equation (5). After controlling for an individual's own skill on the Memory task, a one standard deviation increase in the average RMET score of their randomly assigned teammates improves their performance by 0.092 standard deviations, an increase that is statistically significant at the less than 1 percent level.

Teammates with higher social skills, as measured by RMET, causally improve an individual's performance on a memory sub-task, even though their teammates view different stimuli and cannot directly help them answer the recall questions. Column 2 adds a control for the individual's own RMET score. Even though RMET is a strong independent predictor of performance, the impact of teammates' RMET only slightly decreases. In both models, we find that teammates with high social skills – as measured by RMET – improve an individual's scores on tasks that are performed independently. This strongly suggests that one mechanism for the team player effect may be increased effort and/or motivation.

5.3 Allocative Efficiency

Another way that social skills might affect group performance is by facilitating a more efficient allocation of group members to tasks in which they have a comparative advantage. Gains from "trading tasks" was the key mechanism in the model in Deming (2017).

We test this mechanism by again focusing on the memory task. Using individual memory sub-task scores from the first phase of the experiment, we generate an expected score for each group assuming that they adopt the most efficient allocation of people to sub-tasks. We then compare this expected score to a prediction that is based on the actual stimuli that individuals were assigned to memorize by their group. We do this in two ways. First, we create a distance measure that takes the difference between the predicted score of the optimal and actual strategies. Groups that chose the efficient allocation have a distance of zero.[27] Second, we simply order all six possible strategies from best to worst and assign to each group the rank of the strategy they actually choose.

Teams with more efficient allocations have modestly better performance - $(\sigma = 0.11, p = 0.06)$ for the rank measure, and $(\sigma = 0.07, p = 0.21)$ for the distance measure. However, we find no evidence that groups with higher social skill formed better strategies. The correlations between both rank and distance

[27] We reverse the sign so that increases in the distance score correspond to improvements in strategy.

and the average social skills of the group $\overline{S_g}$ are both positive, but very small and not statistically distinguishable from zero.

Finally, we test the hypothesis that cognitive skill and social skill are complements in team production, which is an indirect implication of the allocative efficiency mechanism in Deming (2017). In Deming (2017), social skill improves productivity by lowering communication frictions between teammates, allowing them to more efficiently specialize in their comparative advantage(s). This implies that cognitive skill and social skill are complements, and that the interaction term would be positive in a regression where productivity (or wages) is the outcome.

Appendix Table A5 presents results from a regression of the raw team player index (estimated without controls, as in column 3 of Table 3) on cognitive skill (as measured both by Ravens and task-specific skill), social skill, and the interaction terms. We find no evidence of complementarity in any specifications.

Overall, while allocative efficiency is a potential mechanism by which social skills could affect group performance, we did not find any evidence for it. That said, given that relationship between allocative efficiency and performance was only modestly positive in the first place, this is not a well-powered test. Complementarity might be a more important mechanism as team production becomes increasingly complex, and our tasks were relatively simple, with little scope for allocative efficiency gains.

Section 6: Conclusion

In this paper, we develop a new method for estimating individual contributions to team performance. We repeatedly randomly assign people to teams and find that some people consistently cause their teams to exceed predicted performance. These people are good "team players".

The team player effect is not predicted by demographic characteristics such as age, gender, education, ethnicity or IQ. Yet it is strongly related to individual scores on the Reading the Mind in the Eyes test, a widely used measure of social intelligence. Social intelligence requires the ability to read others' emotional states, which is probably a necessary - but not sufficient - condition for being a good teammate. Social skills predict group performance about as well as IQ. This suggests that being a good team player is an important skill that is distinct from general ability.

Our results uncover a direct mechanism for the growing body of evidence on the importance of "non-cognitive" skills in the labor market (e.g. Kuhn and Weinberger 2005, Heckman, Stixrud and Urzua 2006, Borghans et al. 2008, Almlund et al. 2011, Lindqvist and Vestman 2011, Heckman and Kautz 2012, Weinberger 2014, Deming 2017). We also find suggestive evidence that team players improve group performance by encouraging effort among their teammates.

Our experimental approach highlights one way that organizations can identify good "team players". Future work should focus on scaling up the results of our experiment and testing the viability of measuring the importance of social skills in a variety of other settings.

References

Abowd, J. M., Kramarz, F., & Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica, 67*(2), 251-333.

Adams Jr., R. B., Rule, N. O., Franklin Jr., R. G., Wang, E., Stevenson, M. T., Yoshikawa, S., Nomura, M., Sato, W., Kveraga, K., & Ambady, N. (2010). Cross-cultural reading the mind in the eyes: An fMRI investigation. *Journal of Cognitive Neuroscience, 22*(1), 97-108.

Almaatouq, A., Yin, M., & Watts, D. (2020). "Collective Problem-Solving of Groups Across Tasks of Varying Complexity," Unpublished working paper (found at https://psyarxiv.com/ra9qy).

Almlund, M., Duckworth, A., Heckman, J. J., & Kautz, T. (2011). Personality psychology and economics. In E. A. Hanushek, S. Machin, & L. Woßmann, eds., *Handbook of the economics of education*, Vol. 4 (1st ed.), pp. 1-181. Amsterdam: Elsevier.

Arcidiacono, P., Foster, J., Goodpaster, N., & Kinsler, J. (2012). Estimating spillovers using panel data, with an application to the classroom. *Quantitative Economics, 3*(3), 421-70.

Arcidiacono, P., Kinsler, J., & Price, J. (2017). Productivity spillovers in team production: Evidence from professional basketball. *Journal of Labor Economics, 35*(1), 191-225.

Austin, J. R. (2003). Transactive memory in organizational groups: The effects of content, consensus, specialization, and accuracy on group performance. *Journal of Applied Psychology, 88*(5), 866.

Azoulay, P., Graff Zivin, J. S., & Wang, J. (2010). Superstar extinction. *Quarterly Journal of Economics, 125*(2), 549-89.

Baddeley, A. D. (2001). Is working memory still working? *American Psychologist, 56*(11), 851.

Baker, C. A., Peterson, E., Pulos, S., & Kirkland, R. A. (2014). Eyes and IQ: A meta-analysis of the relationship between intelligence and "Reading the Mind in the Eyes". *Intelligence, 44*, 78-92.

Barlow, J. B., & Dennis, A. R. (2016). Not as smart as we think: A study of collective intelligence in virtual groups. *Journal of Management Information Systems*, *33*(3), 684-712.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “Reading the Mind in the Eyes” test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42*(2), 241-51.

Bartel, A. P., Beaulieu, N. D., Phibbs, C. S., & Stone, P. W. (2014). Human capital and productivity in a team environment: Evidence from the healthcare sector. *American Economic Journal: Applied Economics, 6*(2), 231-59.

Bates, T., & Gupta, S. (2017). Smart groups of smart people: Evidence for IQ as the origin of collective intelligence in the performance of human groups. *Intelligence, 60*, 46-56.

Becker, A., Deckers, T., Dohmen, T., Falk, A. and Kosse, F. (2012). The Relationship Between Economic Preferences and Psychological Personality Measures. *Annual Review of Economics*, *4*(1), 453-78.

Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: A meta-analysis. *Journal of Applied Psychology, 92*(3), 595.

Bloom, N., & van Reenen, J. (2011). Human resource management and productivity. In O. Ashenfelter & D. Card, eds., *Handbook of labor economics*, Vol. 4 (1st ed.), pp. 1697-767. Amsterdam: Elsevier.

Boning, B., Ichniowski, C., & Shaw, K. (2007). Opportunity counts: Teams and the effectiveness of production incentives. *Journal of Labor Economics, 25*(4), 513-50.

Borghans, L., Duckworth, A. L., Heckman, J. J., & Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources, 43*(4), 972-1059.

Brandt, J. (1991). The Hopkins Verbal Learning Test: Development of a new memory test with six equivalent forms. *The Clinical Neuropsychologist, 5*(2), 125-142.

Brune, L., Chyn, E., & Kerwin, J. (2017). Peers and motivation at work: Evidence from a firm experiment in Malawi. In *2017 International Population Conference*. IUSSP.

Card, D., Heining, J., & Kline, P. (2013). Workplace heterogeneity and the rise of West German wage inequality. *Quarterly Journal of Economics, 128*(3), 967-1015.

Chan, D. C. (2016). Teamwork and moral hazard: Evidence from the emergency department. *Journal of Political Economy, 124*(3), 734-70.

Charness, G., & Kuhn, P. (2011). Lab labor: What can labor economists learn from the lab? In O. Ashenfelter & D. Card, eds., *Handbook of labor economics*, Vol. 4 (1st ed.), pp. 229-330. Amsterdam: Elsevier.

Colom, R., Rebollo, I., Abad, F. J., & Shih, P. C. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A reanalysis of key studies. *Memory & Cognition, 34*(1), 158-71.

Cornelissen, T., Dustmann, C., & Schönberg, U. (2017). Peer effects in the workplace. *American Economic Review, 107*(2), 425-56.

Credé, M., & Howardson, G. (2017). The structure of group task performance—A second look at "collective intelligence": Comment on Wolley et al. (2010). *Journal of Applied Psychology, 102*(10), 1483-92.

Cunha, F., Heckman, J. J., & Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, *78*(3), 883-931.

Curşeu, P. L., Jansen, R. J. G., & Chappin, M. M. H. (2013). Decision rules and group rationality: Cognitive gain or standstill? *PloS One, 8*(2), e56454. doi:10.1371/journal.pone.0056454

Davison, A. C., & Hinkley, D. V. (1997). Bootstrap methods and their application (Vol. 1): Cambridge university press.

Deckers, T., Falk, A., Kosse, F., Pinger, P.R. and Schildberg-Hörisch, H. 2020. Socio-economic status and inequalities in children's IQ and economic preferences, *Journal of Political Economy,* forthcoming.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *Quarterly Journal of Economics, 132*(4), 1593-640.

Dennett, H. W., McKone, E., Tavashmi, R., Hall, A., Pidcock, M., Edwards, M., & Duchaine, B. (2012). The Cambridge Car Memory Test: A task matched in format to the Cambridge Face Memory Test, with norms, reliability, sex differences, dissociations from face memory, and expertise effects. *Behavior Research Methods, 44*(2), 587-605.

Devereux, K. (2018). Identifying the value of teamwork: Application to professional tennis. Working Paper Series No. 14, Canadian Labour Economics Forum.

Devine, D. J., & Phillips, J. L. (2001). Do smarter teams do better: A meta-analysis of cognitive ability and team performance. *Small Group Research, 32*(5), 507-32.

Driskell, J. E., Salas, E., & Driskell, T. (2018). Foundations of teamwork and collaboration. *American Psychologist, 73*(4), 334.

Driskell, J. E., Salas, E., & Hughes, S. (2010). Collective orientation and team performance: Development of an individual differences measure. *Human Factors, 52*(2), 316-28.

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia, 44*(4), 576-585.

Edmondson, A. (2012). Teamwork on the fly. *Harvard Business Review, 90*(4), 72-80.

Engel, D., Woolley, A. W., Jing, L. X., Chabris, C. F., & Malone, T. W. (2014). Reading the mind in the eyes or reading between the lines? Theory of mind predicts collective intelligence equally well online and face-to-face. *PloS One, 9*(12).

Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science, 326*(5952), 535-38.

Falk, A., & Ichino, A. (2006). Clean evidence on peer effects. *Journal of Labor Economics, 24*(1), 39-57.

Feld, J., & Zölitz, U. (2017). Understanding peer effects: On the nature, estimation, and channels of peer effects. *Journal of Labor Economics, 35*(2), 387-428.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure*. Psychological Assessment, 4*(1), 26.

Hamilton, B. H., Nickerson, J. A., & Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy, 111*(3), 465-97.

Hansen, M. J., & Vaagen, H. (2016). Collective intelligence in project groups: Reflections from the field. *Procedia Computer Science, 100*, 840-7.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics, 19*(4), 451-64.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics, 24*(3), 411-82.

Herbst, D., & Mas, A. (2015). Peer effects on worker output in the laboratory generalize to the field. *Science, 350*(6260), 545-49.

Hollingshead, A. B. (1998). Group and individual training: The impact of practice on performance. *Small Group Research, 29*(2), 254-280.

IPIP. (2018). Retrieved from https://ipip.ori.org/New_IPIP-50-item-scale.htm

Isphording, I. E., & Zölitz, U. (2019). The value of a peer. *Working paper.*

Jaravel, X., Petkova, N., & Bell, A. (2018). Team-specific capital and innovation. *American Economic Review, 108*(4-5), 1034-73.

Kane, T. J., & Staiger, D. O. (2005). Using imperfect information to identify effective teachers. *Unpublished article.*

Kim, Y., Aggarwal, I., & Woolley, A. W. (2016). How collective intelligence impacts team performance: Exploring transactive memory system as a mechanism. *Unpublished article.*

Kosse, F., Deckers, T., Pinger, P., Schildberg-Hörisch, H. and Falk, A., 2020. The formation of prosociality: causal evidence on the role of social environment. *Journal of Political Economy*, *128*(2), 434-467.

Kuhn, P., & Weinberger, C. (2005). Leadership skills and wages. *Journal of Labor Economics, 23*(3), 395-436.

Larson, J. R. (2010). *In search of synergy in small group performance*: Psychology Press.

Laughlin, P. R., Bonner, B. L., & Miner, A. G. (2002). Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behavior and Human Decision Processes, 88*(2), 605-620.

Laughlin, P. R., Hatch, E. C., Silver, J. S., & Boh, L. (2006). Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and Social Psychology, 90*(4), 644.

Lazear, E. P., & Shaw, K. L. (2007). Personnel economics: The economist's view of human resources. *Journal of Economic Perspectives*, 21(4), 91-114.

Lewis, K. (2004). Knowledge and performance in knowledge-worker teams: A longitudinal study of transactive memory systems. *Management Science, 50*(11), 1519-1533.

Lindbeck, A., & Snower, D. J. (2000). Multitask learning and the reorganization of work: From Tayloristic to holistic organization. *Journal of Labor Economics, 18*(3), 353-76.

Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics, 3*(1), 101-28.

Mas, A., & Moretti, E. (2009). Peers at work. *American Economic Review, 99*(1), 112-45.

Mason, W. A., Jones, A., & Goldstone, R. L. (2008). Propagation of innovations in networked groups. *Journal of Experimental Psychology: General, 137*(3), 422.

McCann, R. J., Shi, X., Siow, A., & Wolthoff, R. (2015). Becker meets ricardo: Multisector matching with communication and cognitive skills. *The Journal of Law, Economics, and Organization*, *31*(4), 690-720.

Mesmer-Magnus, J. R., & DeChurch, L. A. (2009). Information sharing and team performance: A meta-analysis. *Journal of Applied Psychology, 94*(2), 535.

Morgeson, F. P., Reider, M. H., & Campion, M. A. (2005). Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge. *Personnel Psychology, 58*(3), 583-611.

Mumford, T. V., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2008). The team role test: Development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology, 93*(2), 250-67.

National Association of Colleges and Employers (NACE) (2019). Job outlook 2020.

National Center for O*NET Development (2020). Work context: work with work group or team. *O*NET OnLine*. Retrieved from https://www.onetonline.org/find/descriptor/result/4.C.1.b.1.e?s=1&a=1

Neffke, F. M. (2019). The value of complementary co-workers. *Science Advances, 5*(12), eaax3370.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist, 67*(2), 130.

Oettl, A. (2012). Reconceptualizing stars: Scientist helpfulness and peer performance. *Management Science, 58*(6), 1122-40.

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: toward a brief form for research and applied settings. *Frontiers in Psychology*, *6*, 1503.

Park, S. (2019). Socializing at work: Evidence from a field experiment with manufacturing workers. *American Economic Journal: Applied Economics, 11*(3), 424-55.

Pinkham, A. E., Penn, D. L., Green, M. F., Buck, B., Healey, K., Harvey, P. D. (2014). The social cognition psychometric evaluation study: Results of the expert survey and RAND panel. *Schizophrenia Bulletin, 40*(4), 813-23.

Raven, J. (2003). Raven's Advanced Progressive Matrices, Test Books 1 and 2: Pearson Assessment.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247-52.

Sullivan, K. (2005). Alternate forms of prose passages for the assessment of auditory-verbal memory. *Archives of Clinical Neuropsychology, 20*(6), 745-753.

Venzon, D.J. and Moolgavkar, S.H. (1988). A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 37*(1), 87-94.

Waldinger, F. (2012). Peer effects in science: Evidence from the dismissal of scientists in Nazi Germany. *Review of Economic Studies, 79*(2), 838-61.

Wechsler, D. (1997). WMS-III: Wechsler memory scale. In *Wechsler memory scale* (3rd ed.). San Antonio, TX: Psychological Corporation.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In *Theories of group behavior* (pp. 185-208): Springer.

Weinberger, C. J. (2014). The increasing complementarity between cognitive and social skills. *Review of Economics and Statistics, 96*(5): 849-61.

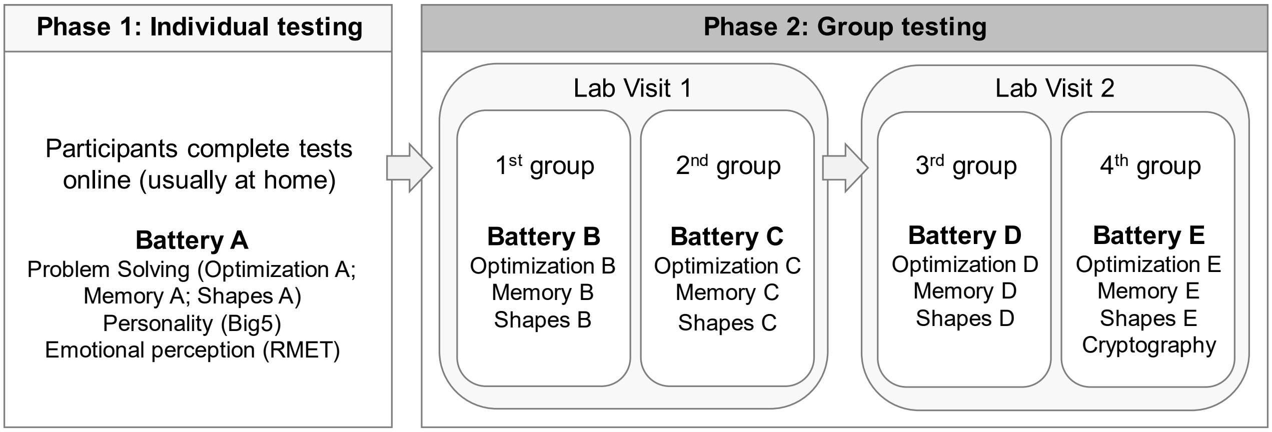
Weingart, L. R. (1992). Impact of group goals, task component complexity, effort, and planning on group performance. *Journal of Applied Psychology, 77*(5), 682.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science, 330*(6004), 686-688.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036-9.

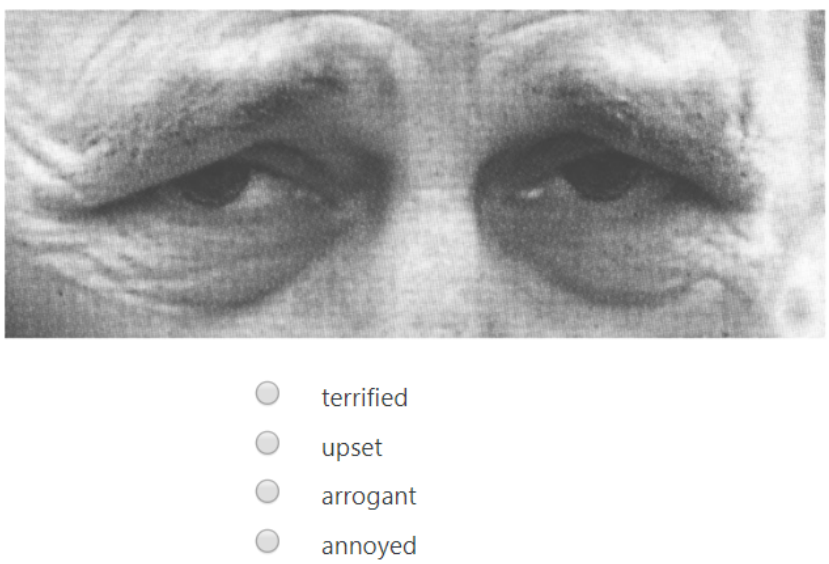
Yamaguchi, S. (2012). Tasks and heterogeneous human capital. *Journal of Labor Economics*, *30*(1), 1-53.

Figure 1: Overview of Experiment



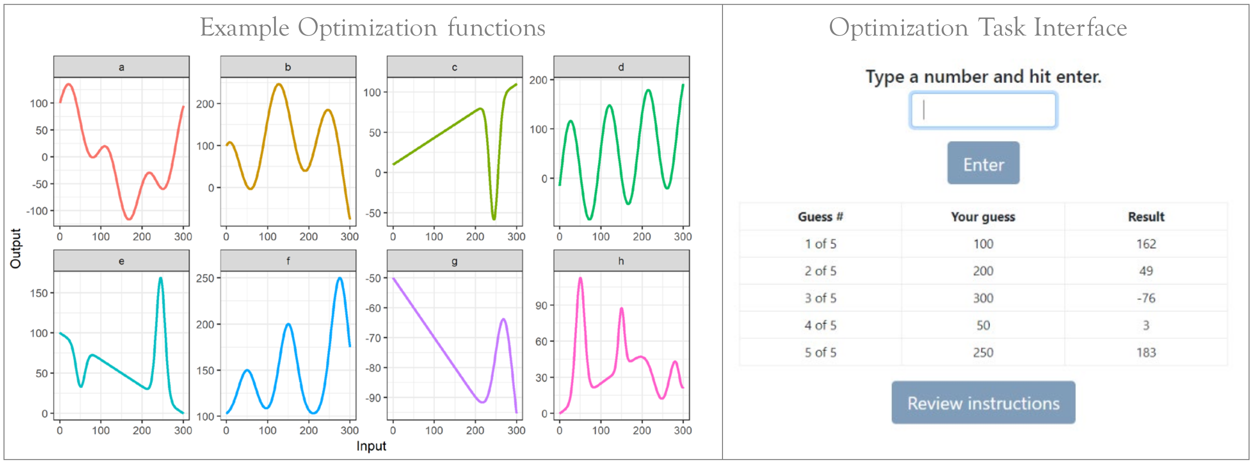
Notes: this figure describes the experimental design from an individual participant's perspective. Tasks are described in Section 2.2. Task batteries represent unique sequences of tasks. Participants never see the same exact task more than once. Lab visits involve 6, 9 or 12 participants, who were randomly allocated to groups of 3 people - see Sections 2.3 and 2.4 for details.

Figure 2: Example Item from the Reading the Mind in the Eyes Test (RMET)



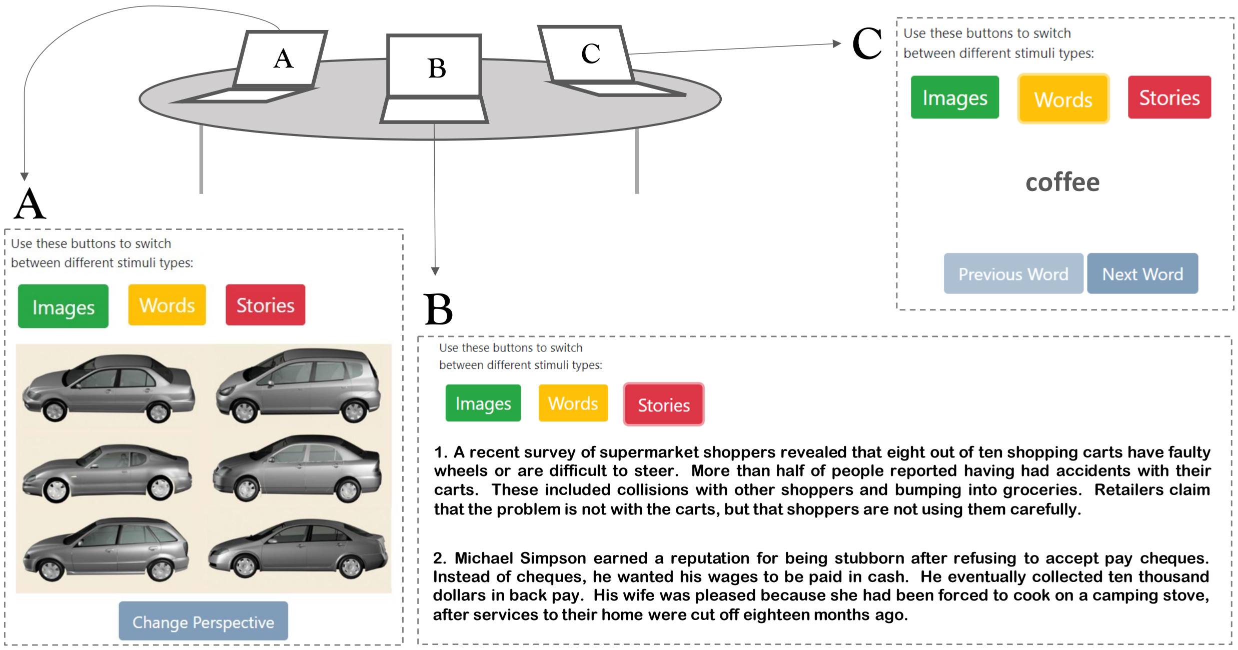
Notes: this is an example item from the Reading the Mind in the Eyes Test (RMET), a well-established and psychometrically validated test of emotion recognition and social intelligence (Baron-Cohen et al, 2001). We administered a 26-item version of the test. The mean score on the test was 18.5 out of 26.

Figure 3: Description of Optimization Task



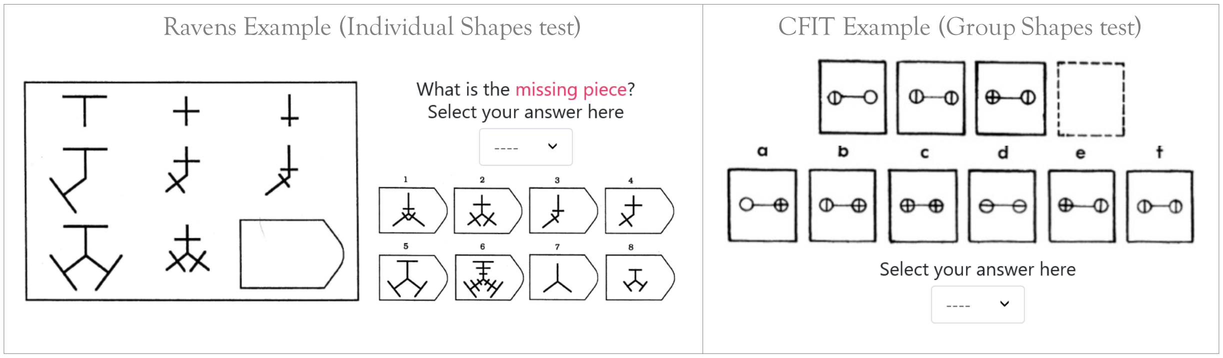
Notes: this figure presents a visual description of the Optimization task. Participants were asked to enter guesses between 0 and 300 (the horizontal axis). They received an output from a complex function (the vertical axis). The left panel contains example functions, which were hidden to participants. The right panel is the participant interface. The goal of the task is to find the maximum of the function. In the first phase of the experiment, individuals received 15 guesses before submitting an answer. In the second phase, each member of a 3-person team received 5 guesses; once all these guesses had been entered, the group agreed upon a final answer. See Section 2.2 for details.

Figure 4: Description of Group Memory Task



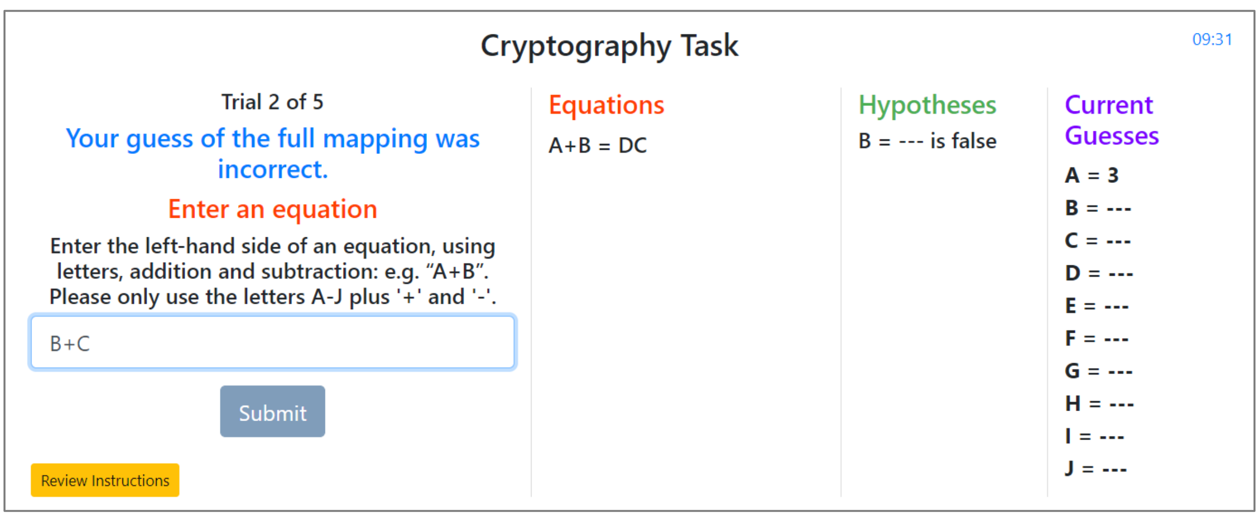
Notes: this figure presents a visual description of the group Memory task. Participants were given 40 seconds in which they could cycle through: 6 different images (Panel A), 2 different stories (Panel B), and 12 different words (Panel C) on their own laptop. Then groups gathered around a single laptop and answered 24 questions together about the three sets of stimuli, with an equal number of questions about each. See Section 2.2 for details.

Figure 5 – Description of the Shapes Task



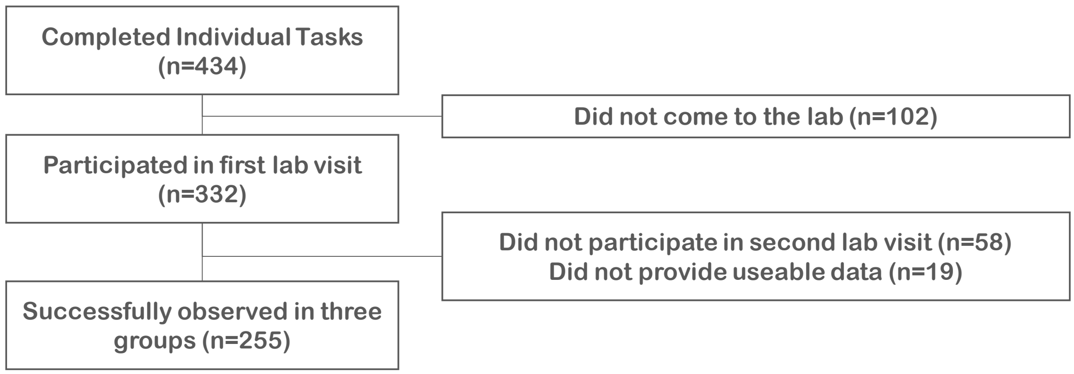
Notes: this figure presents example items from the Shapes task, which was adapted from two well-established measures of IQ or fluid intelligence – the Culture Fair Intelligence Test (CFIT) and the Raven's Advanced Progressive Matrices (Ravens). In the individual phase of the experiment, participants were given 14 items and seven minutes. The mean score was 7.3, and no individual received a perfect score. In the group phase, all members gathered around a laptop and collectively decided on an answer for each item. Mean scores differed by battery but no groups received a perfect score. See Section 2.2 for details.

Figure 6 – Description of the Cryptography Task



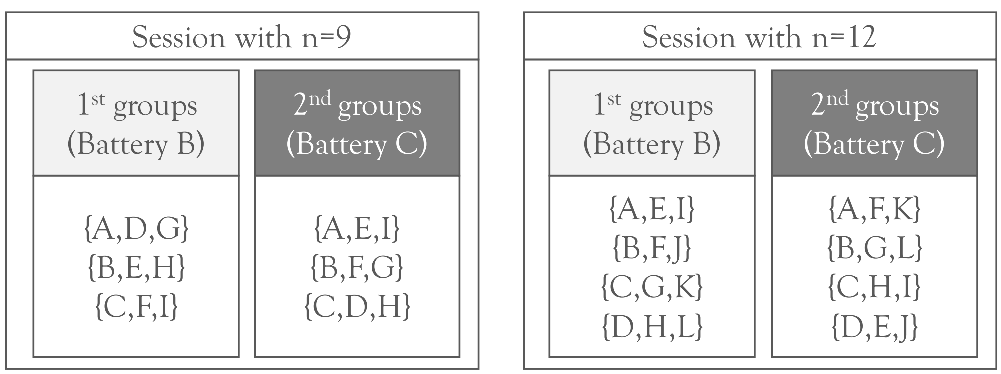
Notes: this figure presents the Cryptography task interface. Cryptography was a fourth, out-of-sample validation task that was not used to estimate the team player index. Each letter from A to J represents a unique number from 0 to 9. Groups were asked to decode the value of each letter by entering mathematical equations that would return an output. The goal was to decode the letters using as few equations as possible. Groups were given one practice try on the Cryptography task, to make sure they understood the task. Each group was allowed to try up to 15 equations on the assessed version of the task. Those who decoded all the letters (81%) used a mean of 7.9 equations. 85 groups attempted Cryptography. See Section 2.2 for details.

Figure 7: Participant Flow Diagram



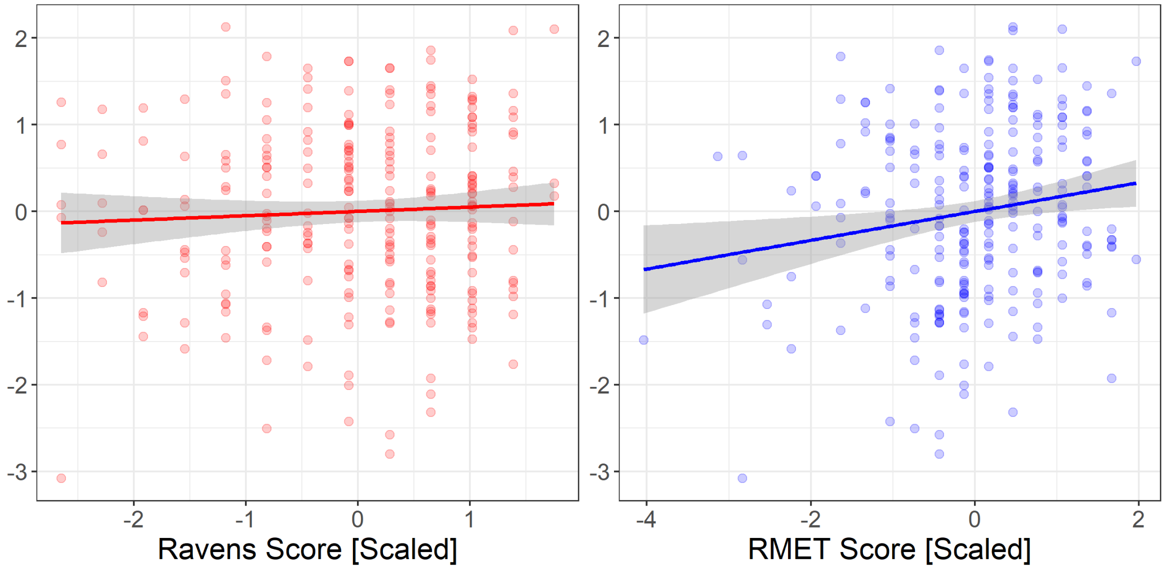
Notes: this figure presents the participant flow for the experiment. For details, see section 2.3.

Figure 8: Example of Blocked Randomization Scheme for First Lab Visit



Notes: this figure presents a visual example of how individuals were randomized to groups over the course of a single lab visit. We use Lab Visit One as an example. The left panel illustrates the randomization process for a session of 9 people; the right panel is the equivalent process for a session of 12 people. Participants were randomized to two successive groups in a single draw, and the randomization was blocked so that, where possible, participants did not have any of the same team members in their second group assignment of a lab session. See Section 2.4 for details.

Figure 9: The Team Player Index is correlated with Social Intelligence, but not IQ

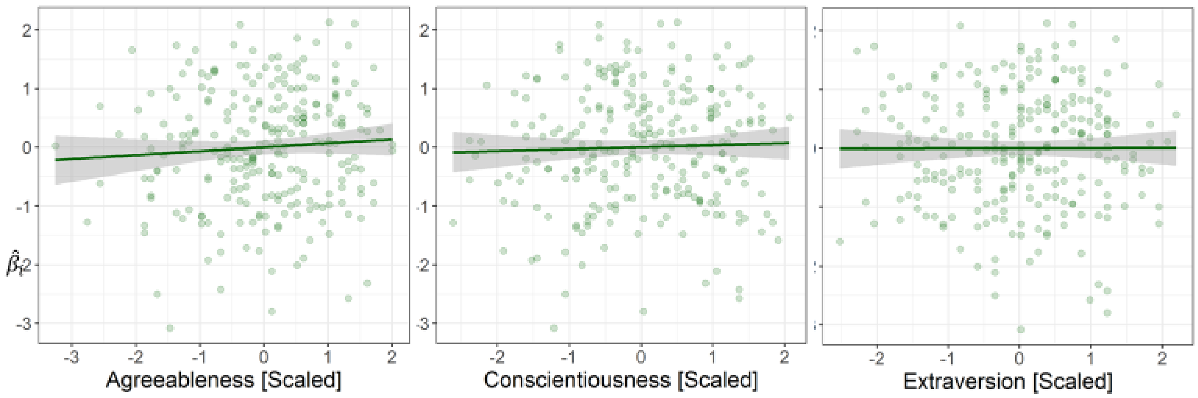


(1)	(2)
0.050	0.379**
(p=0.425)	(p<0.001)

(1)	(2)
0.166**	0.306**
(p=0.008)	(p<0.001)

Notes: each panel presents a scatterplot of an individual's estimated team player index $\hat{\beta}_i$ against their individual Ravens score (left panel) and their individual RMET score (right panel). In both cases $\hat{\beta}_i$ shown in the figures is estimated based on the model in equations (1) through (4), as described in Section 3 and detailed in our pre-registered analysis plan. Ravens is a well-established measure of IQ or fluid intelligence. RMET is the Reading the Mind in the Eyes Test, a well-established test of emotion perception and social intelligence. Beneath each panel, we show coefficients from two different estimates of $\hat{\beta}_i$: (1) our pre-specified model, with controls for task-specific skills and indicators for group familiarity; (2) no controls. See the text for details. The scatterplot always shows estimates from model (1). The same sample was used for all analysis: 1029 group-task observations, 343 groups, 255 participants. *p<0.05; **p<0.01.

Figure 10 – The Team Player Index is Uncorrelated with Personality Scores



(1)	(2)
0.067	0.030
(p=0.285)	(p=0.637)

(1)	(2)
0.033	0.010
(p=0.599)	(p=0.879)

(1)	(2)
0.002	0.012
(p=0.972)	(p=0.849)

Notes: each panel presents a scatterplot of an individual's estimated team player index $\hat{\beta}_i$ against their individual scores on the Agreeableness (left panel), Conscientiousness (middle panel) and Extraversion (right panel) scales of the Big 5 Personality inventory. In all three cases the $\hat{\beta}_i$ show in the figures is estimated based on the model in equations (1) through (4), as described in Section 3 and detailed in our pre-registered analysis plan. Beneath each panel, we show coefficients from two different estimates of $\hat{\beta}_i$: (1) our pre-specified model, with controls for task-specific skills and indicators for group familiarity; (2) no controls. See the text for details. The scatterplot always shows estimates from model (1). The same sample was used for all analysis: 1029 group-task observations, 343 groups, 255 participants. *p<0.05; **p<0.01.

Table 1: Descriptive Statistics

	Study Sample (n=255)	US Population
Age (median)	26	38
Female%	57%	51%
Latino/Hispanic%	13%	18%
Black%	11%	13%
Asian%	27%	6%
White%	35%	60%

Notes: the table presents descriptive statistics for the study population (Column 1), relative to the US population (Column 2). Data for the US population are taken from the 2017 American Community Survey and 2018 estimates from the US Census Bureau.

Table 2 – Are Some People Good Team Players?

`	**Dependent variable: Group Performance $\widetilde{G}_{gk}$**		
	(1)	(2)	(3)
Teamplayer Effect $\hat{\sigma}_{\beta}$	0.127* (p=0.029)	0.127* (p=0.037)	0.127** (p<0.001)
Inference method	Randomization Inference	Profile Likelihood	Normal approximation
Problem-solving skills			
Task-specific skills			
Memory°	0.166** (0.032)	0.166** (0.032)	0.166** (0.032)
Optimization°	0.125** (0.031)	0.125** (0.031)	0.125** (0.031)
Ravens (Shapes) °	0.302** (0.030)	0.302** (0.030)	0.302** (0.030)
Group familiarity controls?	✓	✓	✓
Number of groups	343	343	343
Number of participants	255	255	255

Notes: °Indicates group-level sum of individual skills (e.g. $\sum_i I_g^i X_{gk}$). The team player effect is the standard deviation of our estimate of the causal contribution of each individual to group performance – see Section 3 for details. Group familiarity controls are indicators for whether any members of the group knew each other prior to randomization, and whether any members of the team had been on a team together during a previous battery. Covariate coefficients have standard errors in parentheses. The same sample was used for all analysis: 1029 group-task observations, 343 groups, 255 participants. The results in Column 1 are the model from our pre-registered analysis plan, which can be found in the AEA RCT registry as AEARCTR-0002896. *p<0.05; **p<0.01.

Table 3 –Team Player Effect with Different Controls for Group Performance

`	Dependent variable: Group Performance $\widetilde{G}_{gk}$				
	(1)	(2)	(3)	(4)	(5)
Teamplayer Effect $\hat{\sigma}_\beta$	0.127* (p=0.029)	0.173** (p=0.002)	0.244** (p<0.001)	0.237** (p<0.001)	0.192** (p<0.001)
Problem-solving skills					
Task-specific skills					
Memory°	0.166** (0.032)				
Optimization°	0.125** (0.031)				
Ravens (Shapes) °	0.302** (0.030)				
Ravens°		0.161** (0.018)			
RMET°					0.119** (0.017)
Personality					
Agreeableness°				0.017 (0.020)	
Extraversion°				0.002 (0.020)	
Conscientious°				0.002 (0.019)	
Group familiarity controls?	✓	✓		✓	✓

Notes: °Indicates group-level sum; for task-specific skills, this would be $\sum_i I_g^i X_{gk}$. The team player effect is the standard deviation of our estimate of the causal contribution of each individual to group performance – see Section 3 for details. Estimates of the team player effect ($\hat{\sigma}_\beta$) have p-values in parentheses from the randomization inference procedure. Ravens is a well-established measure of IQ or fluid intelligence. RMET is the Reading the Mind in the Eyes Test, a well-established test of emotion perception and social intelligence. Personality comes from three of the five factors in the “Big 5” personality inventory. Group familiarity controls are indicators for whether any members of the group knew each other prior to randomization, and whether any members of the team had been on a team together during a previous lab visit. Covariate coefficients have standard errors in parentheses. The same sample was used for all analysis: 1029 group-task observations, 343 groups, 255 participants. The results in Column 1 are the preferred model from our pre-registered analysis plan. *p<0.05; **p<0.01.

Table 4 – Validating the Team Player Effect with Out-of-Sample Task Performance

	Dependent variable: Score on Cryptography Task C_{gk}					
	(1)	(2)	(3)	(4)	(5)	(6)
Team player index (β)						
Group Mean	0.153 (0.108)					
Group Max		0.165 (0.108)				
Group Min			0.092 (0.109)			
Group contains someone with $\beta > 1\sigma$				0.247* (0.106)		
RMET group mean ($mean_g\, RMET_i$)					0.192 (0.108)	
Ravens group mean ($mean_g\, Ravens_i$)						0.217* (0.107)

Notes: results from the Cryptography task were available for N=85 groups. See Section 2.2 for details about the task. The team player index comes from our pre-registered model described in section 3. Ravens is a well-established measure of IQ or fluid intelligence. RMET is the Reading the Mind in the Eyes Test, a well-established test of emotion perception and social intelligence. Coefficients have standard errors in parentheses. All variables were standardized to have mean = 0 and sd = 1. The results in Columns 1, 5 and 6 are from our pre-registered analysis plan. *p<0.05; **p<0.01.

Table 5 – The Team Player Index is Uncorrelated with Demographic Characteristics

	Dependent variable: Team Player Index $\widehat{\beta}_i$			
	(1)	(2)	(3)	(4)
Female	-0.007 (0.063)			
Years of education		0.025 (0.063)		
Age			-0.084 (0.063)	
Under represented minority°				-0.071 (0.147)
Observations	252	254	252	254

Notes: each column presents bivariate correlations between estimates of $\hat{\beta}_i$ from our pre-registered model described in Section 3 and the indicated demographic characteristics. °Under-represented minorities are participants who identified as African-American, Latino/Hispanic or Native American. Standard errors are presented in parentheses. All variables were standardized to have mean = 0 and sd = 1.

Table 6: Correlation between Social Intelligence and the Team Player Effect

	Dependent variable: $\hat{\boldsymbol{\beta}}_i$ (pre-specified model)			
	(1)	(2)	(3)	(4)
RMET	0.166**	0.169**	0.192**	
	(0.062)	(0.063)	(0.068)	
Age		-0.085	-0.129	-0.105
		(0.064)	(0.073)	(0.074)
Female		-0.030	-0.042	-0.020
		(0.064)	(0.065)	(0.065)
Years of education		0.037	0.042	0.015
		(0.090)	(0.091)	(0.092)
Years of education^2		-0.031	-0.025	-0.035
		(0.089)	(0.090)	(0.091)
Ravens			-0.073	-0.004
			(0.075)	(0.073)
Personality				
Agreeableness			0.061	0.073
			(0.070)	(0.071)
Extraversion			-0.002	-0.009
			(0.068)	(0.069)
Conscientious			0.041	0.028
			(0.067)	(0.068)
Observations	255	250	250	250
R^2	0.028	0.038	0.048	0.016
Adjusted R^2	0.024	0.018	0.012	-0.017

Notes: each column presents a regression in which the dependent variable is the team player index ($\hat{\beta}_i$) from our pre-registered model described in Section 3. RMET is the Reading the Mind in the Eyes Test, a well-established test of emotional perception and social intelligence. Ravens is a well-established measure of IQ or fluid intelligence. Personality comes from three of the five factors in the "Big 5" personality inventory. Covariate coefficients have standard errors in parentheses. All variables were standardized to have mean = 0 and sd = 1. The same sample was used for all analysis: 1029 group-task observations, 343 groups, 255 participants. *p<0.05; **p<0.01.

Table 7: Predictive Power of Social Skills and Other Factors on the Cryptography Task

	Dependent variable: $\widehat{C}_g$ (Cryptography Score)					
	(1)	(2)	(3)	(4)	(5)	(6)
$\overline{RMET}$	0.192					
	(0.108)					
S (Social skills)						
Group mean		0.218*		0.177	0.198	
		(0.107)		(0.109)	(0.111)	
Group max						0.241*
						(0.111)
Ravens (IQ)						
Group mean			0.217*	0.175	0.205	
			(0.107)	(0.109)	(0.110)	
Group max						0.202
						(0.107)
Personality						
Agreeableness mean					0.008	-0.010
					(0.117)	(0.117)
Conscientious mean					0.174	0.148
					(0.112)	(0.112)
Extraversion mean					-0.159	-0.128
					(0.112)	(0.111)
Observations	85	85	85	85	85	85
R^2	0.037	0.048	0.047	0.077	0.125	0.144
Adjusted R^2	0.025	0.036	0.035	0.054	0.070	0.090

Notes: results from the Cryptography task were available for N=85 groups. See Section 2.2 for details about the task. $\overline{RMET}$ is the group's average score on the Reading the Mind in the Eyes Test, a well-established test of emotion perception and social intelligence. S (Social Skills) is the average of each participant's mean standardized score on the RMET and their estimated team player index. To avoid double-counting, the team player index used in this average controls from group endowments of RMET. Ravens is a well-established measure of IQ or fluid intelligence. Personality comes from three of the five factors in the "Big 5" personality inventory. Covariate coefficients have standard errors in parentheses. All variables were standardized to have mean = 0 and sd = 1. *p<0.05; **p<0.01.

Table 8: Team Players Improve Teammate Performance on Sub-Tasks that are Performed Separately

	Dependent variable: $G_{gt[i]}$ (Memory score)	
	(1)	(2)
Individual Memory ($X_{i,memory,t}$)	0.184** (0.030)	0.157** (0.030)
Mean RMET in $i's$ group ($\overline{RMET}_{g[-i]}$)	0.092** (0.030)	0.088** (0.030)
Individual RMET ($RMET_i$)		0.085** (0.030)
Controls for memory battery (Round)	✓	✓
Controls for memory type (Type)	✓	✓
Observations	921	921

Notes: this table presents results from a regression of the group's average score on a memory sub-task t (words, images, stories) on the average RMET score of participant i's teammates. The model is fit to the 92% of cases in which only one individual looked at each stimulus type during the memorization period. In these teams, individuals memorized separate material, yet having good teammates still improves performance. We also include controls for individual memory scores on task type t as well as fixed effects for task battery and memory type. Covariate coefficients have standard errors in parentheses. Variables were standardized to have mean = 0 and sd = 1. *p<0.05; **p<0.01.