



Topics in Cognitive Science 00 (2023) 1–20


© 2023 The Authors. *Topics in Cognitive Science* published by Wiley Periodicals LLC on behalf of Cognitive Science Society.

ISSN: 1756-8765 online

DOI: 10.1111/tops.12706

This article is part of the topic “How Minds Work: The Collective in the Individual,” Nat Rabb and Steven Sloman (Topic Editors).

The Effects of Group Composition and Dynamics on Collective Performance

Abdullah Almaatouq,^a  Mohammed Alsobay,^a Ming Yin,^b
Duncan J. Watts^{c,d,e}

^a*Sloan School of Management, Massachusetts Institute of Technology*

^b*Department of Computer Science, Purdue University*

^c*Department of Computer and Information Science, University of Pennsylvania*

^d*The Annenberg School of Communication, University of Pennsylvania*

^e*Operations, Information, and Decisions Department, University of Pennsylvania*

Received 11 July 2022; received in revised form 4 October 2023; accepted 24 October 2023

Abstract

As organizations gravitate to group-based structures, the problem of improving performance through judicious selection of group members has preoccupied scientists and managers alike. However, which individual attributes best predict group performance remains poorly understood. Here, we describe a preregistered experiment in which we simultaneously manipulated four widely studied attributes of group compositions: skill level, skill diversity, social perceptiveness, and cognitive style diversity. We find that while the average skill level of group members, skill diversity, and social perceptiveness are significant predictors of group performance, skill level dominates all other factors combined. Additionally, we explore the relationship between patterns of collaborative behavior and performance outcomes and find that any potential gains in solution quality from additional communication between the group members are outweighed by the overhead time cost, leading to lower overall efficiency. However, groups exhibiting more “turn-taking” behavior are considerably faster and thus more efficient. Finally, contrary to our expectation, we find that group compositional factors (i.e., skill level and social

Correspondence should be sent to Abdullah Almaatouq, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: amaatouq@mit.edu

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

perceptiveness) are not associated with the amount of communication between group members nor turn-taking dynamics.

Keywords: Collective performance; Group composition; Collective intelligence; Virtual labs

1. Introduction

Problem-solving in groups is ubiquitous throughout the economy and society. Business firms have long been highly reliant on teams for functions as diverse as engineering, design, and marketing, but other domains, including science, are also increasingly group-based (Wuchty, Jones, & Uzzi, 2007). Naturally, questions about how to most effectively construct and manage groups have also preoccupied researchers across a variety of fields, including psychology, economics, management science, and, more recently, complexity science (Bahrami et al., 2010; Jones, Wuchty, & Uzzi, 2008; Mukherjee, Huang, Neidhardt, Uzzi, & Contractor, 2019); (Woolley, Chabris, Pentland, Hashmi, & Malone, 2010); (Wu, Wang, & Evans, 2019; Wuchty et al., 2007).

In spite of this attention, research on the performance of groups of problem-solvers has often reached inconsistent or conflicting conclusions. One such area of disagreement pertains to the effects of different group compositions on group performance. For example, several lab studies found that average ability was the most consistent predictor of group performance (Bell, 2007; Devine & Philips, 2001; Laughlin & Adamopoulos, 1980; LePine, 2003); (Riedl, Kim, Gupta, Malone, & Woolley, 2021; Stewart, 2006). Other studies, however, have argued the opposite: that average ability is less relevant to group performance than factors such as social perceptiveness (aka emotional intelligence) (Engel, Woolley, Jing, Chabris, & Malone, 2014; Kim et al., 2017; Lillis, 2007); (Woolley et al., 2010), skill diversity (Hong & Page, 2004; Page, 2008), and cognitive style diversity (AlShebli, Rahwan, & Woon, 2018; Aggarwal & Woolley, 2019; Bendor & Page, 2019; Ellemers & Rink, 2016).

Motivated by these seemingly conflicting results, we advocate moving from a traditional “explanatory” framework, in which one seeks to test the sign and statistical significance of a single factor of theoretical interest, to a “predictive” framework, in which one instead seeks to minimize out-of-sample prediction error on some outcome of interest using all available features (Hofman et al., 2021a, 2021b; Watts, 2017; Yarkoni & Westfall, 2017).

Concretely, imagine starting with certain information about a sample of potential group members that includes, for example, their previously demonstrated skill on a related task, their social perceptiveness, and their cognitive style. Armed with these “features,” the prediction approach argues for constructing a model with which one then makes predictions about which combination of features will yield the best performance for the group as a whole. Previous theoretical and empirical findings suggest that any of these features, when considered in isolation, should be somewhat predictive of the outcome of interest; however, they offer little guidance on two questions of central relevance to the prediction problem. First, because different features are emphasized in different studies, it is difficult to quantify the relative importance and predictive power of different features: for example, does average skill account

for more or less variance than social perceptiveness or skill diversity, and if so, by how much? Second, because the traditional method for establishing a causal effect of a given feature is to reject the null hypothesis that the effect size is zero, it is difficult to know how much variance of the outcome of interest (group performance) can be explained at all by any combination of features (Lo, Chernoff, Zheng, & Lo, 2015; Ward, Greenhill, & Bakke, 2010). In other words, whereas previous work has focused on demonstrating the existence of specific, theoretically motivated effects, we are concerned with comparing the relative importance and out-of-sample predictive power of multiple effects, all of which may influence performance either individually or collectively. Finally, because the relative importance and even direction of all these effects may depend on the details of the task instance in question, the predictive “model” must also be evaluated over the relevant variations in the task class (Almaatouq, Alsobay, Yin, & Watts, 2021; Yarkoni, 2020).

Motivated by this predictive framework, we conducted a novel “two-phase” experiment to answer two main questions (preregistered at AsPredicted.org #13123): **(1)** Which of several competing group compositions dominate group performance in a problem-solving task? **(2)** Are these results robust to variations in the task parameters? In phase 1, we measured several relevant attributes for individual workers (i.e., skill, social perceptiveness, and cognitive style); then, in phase 2, we used this information to construct groups with desired combinations of individual attributes (i.e., group-level skill, skill diversity, group-level social perceptiveness, and cognitive style diversity). Our approach differs from previous work in several respects:

1. We manipulated four widely studied attributes of groups simultaneously, allowing us to quantify the relative importance of these attributes both individually and collectively. We emphasize that, by design, we only compared effects that had previously been claimed to be important in explaining group outcomes. In other words, the intention of our study was not to identify novel effects but rather to evaluate the relative and cumulative importance of previously identified effects.
2. In contrast with many previous studies that have used generic measures of skill (e.g., general cognitive ability) and cognitive style (Blazhenkova & Kozhevnikov, 2009), we directly measured individual skill and problem-solving style (i.e., cognitive style) on the task in question before assignment to groups. Although generic measures have some advantages in generalizing to multiple tasks, our purpose was to explain as much variance as possible for the task in question; thus, we wanted to tie our classification of individual workers as closely as possible to the task.
3. In contrast with previous studies that establish effects on performance indirectly (e.g., null-hypothesis testing, factor analysis, collective intelligence factor, etc.), we evaluated relative importance directly in terms of the features’ ability to predict the outcome of interest in an out-of-sample manner. We emphasize that out-of-sample predictive performance is a much stronger test of a feature’s importance than showing that it is correlated with performance or rejecting the null hypothesis that it has no effect at all (Hofman et al., 2021b; Yarkoni & Westfall, 2017).

4. By systematically varying the task's complexity over a wide range (from "very low" to "very high" complexity) without changing the nature of the task, we determined how, or if, the relative importance of different attributes is robust to changes in task complexity (e.g., does social perceptiveness, or skill diversity, matter more for the most complex tasks than for simple tasks?).
5. We used a block randomization scheme that intentionally oversampled infrequent combinations of individuals (e.g., "all high skill and high social perceptiveness"), thereby greatly increasing our statistical power.
6. We preregistered our research questions and analysis plan, thereby increasing the replicability of our findings (Simmons, Nelson, & Simonsohn, 2011). See Section S1 in the Supporting Information of Almaatouq et al. (2021) for exceptions to the preregistration plan.

2. Experiment design

We tackle our two questions in a two-phase web-based experiment implemented using the Empirica virtual laboratory platform (Almaatouq et al., 2021). We note that our focus in this paper is on compositional differences between interacting groups, not on the comparison between nominal groups and interacting groups.¹

2.1. Room assignment task

The main task in question was a "room assignment" problem in which participants—first as individuals and then in groups—were required to assign N "students" to M "rooms" where each student had a specified utility for each room. The objective was to maximize total student utility while also respecting Q constraints (e.g., "*Students A and B may not share a room or be in adjacent rooms*"; see Figs. S1 and S2 for screenshots of the experiment; see Section S2.1 for more details about the task). In phase 2, participants were allowed to communicate via text-based chat and move different "students" simultaneously; therefore, they could perform parallel processing, but they were blocked from moving the same student at the same time (i.e., to avoid generating both human confusion and software errors).

We chose the room assignment task for three reasons, similar to those presented in prior research that leveraged this task to study group performance (Almaatouq et al., 2021). First, it is a specific instance of a more general class of complex problems known as "constraint satisfaction and optimization problems" (CSOPs), which are widely studied in artificial intelligence and operations research (Tsang, 2014); thus, our findings will inform collective solutions of CSOPs in general (Almaatouq et al., 2021). Second, as with other complex problems, the payoff function for CSOPs can be described as a "rugged landscape" characterized by many locally optimal but globally suboptimal solutions (Baumann, Schmidt, & Stieglitz,

¹ Our analysis of the conditions under which groups of interacting problem-solvers outperform autonomous individuals is published in Almaatouq et al. (2021)—as per our preregistration.

2019; Shirado & Christakis, 2017; Shore, Bernstein, & Lazer, 2015; Yahosseini & Moussaïd, 2019). Correspondingly, CSOPs are amenable to potentially many solution strategies and cognitive styles, where no single strategy is universally superior (Wolpert & Macready, 1997). Third, the complexity of CSOPs can be systematically varied by adjusting a few key parameters—in our case, by changing the numbers of students (N), rooms (M), and constraints (Q). The analyses in Section S4 demonstrate that the manipulation of task complexity was effective, and led to increased “experienced” complexity (Almaatouq et al., 2021; Liu & Li, 2012). More details about the room assignment task can be found in Almaatouq et al. (2021).

2.2. Phase 1 of the experiment

In phase 1, 1200 participants recruited from Amazon’s Mechanical Turk completed five room-assignment tasks with varying complexity levels, as well as a standard “Reading the Mind in the Eyes” (RME) test (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), which is commonly used as a measure of social perceptiveness (see Sections S2.2 and S2.3 for more details).

In the RME test, participants were shown 36 pairs of eyes, where for each pair of eyes they had to choose one of four words describing the corresponding emotion (see Fig. S3 for screenshots of the task). This test was used by a number of recent studies relating social perceptiveness to group performance (Kim et al., 2017; Lillis, 2007; Riedl et al., 2021; Weidmann & Deming, 2021; Woolley et al., 2010), and it has been shown to be equally predictive of group performance for both face-to-face groups (interacting freely in a room) and online virtual groups interacting via text-based chat (i.e., they cannot see each other’s eyes or facial expressions at all) (Engel et al., 2014; Kim et al., 2017). These findings support that the RME test captures a deeper, domain-independent aspect of social reasoning, not merely the ability to recognize facial expressions of mental states. For instance, one hypothesis is that what makes socially perceptive individuals better group performers is their superior ability to calibrate the weight they assign to one another during discussions (Moussaïd, Noriega Campero, & Almaatouq, 2018). After the completion of phase 1, we evaluated all participants on skill level, social perceptiveness, and cognitive style (see Section S2.4.2 for phase 1 details).

2.2.1. Skill

Our primary definition of skill was the sum of scores on the two moderately complex room-assignment tasks (for more details and robustness checks, see Almaatouq et al., 2021). The score a participant earned in a room assignment task considers both the total utility of students based on their room assignments and penalties arising from constraint violations (see Sections S2.1 and S2.4.3 for details). Individuals who scored above/below the median skill score were classified as high/low skill, respectively. Note that the task order was not randomized in phase 1 to eliminate noise in measuring individual skill level (e.g., to control for learning effects).

2.2.2. *Social perceptiveness*

We defined social perceptiveness simply as the number of RME questions correctly answered. As with skill, individuals above/below the median social perceptiveness score were classified as high/low social perceptiveness, respectively. Our sample was similar in mean and variability on this test to the original general population sample in Baron-Cohen et al. (ours: $M=27.6$; $SD = 4.3$; $N=1200$; original student population: $M=28.0$; $SD = 3.5$; $N=103$; original general population: $M=26.2$; $SD=3.6$; $N=122$).

2.2.3. *Cognitive style*

Finally, based on the participants' answers to a post-experiment survey question, we defined an individual's cognitive style as belonging to one of two categories: "optimizer," who indicated a preference for allocating all students to rooms for which they had the highest utility before attempting to resolve conflicts; and "resolver," who indicated a preference to first allocate all students with conflicts before moving students to higher-value rooms. Groups are then labeled as either homogeneous or diverse with respect to cognitive style by checking whether the three group members in the group belong to the same type ("homogeneous") or not ("diverse"). Our definition of cognitive style proceeds from three criteria (Aggarwal & Woolley, 2019): it must be persistent for a given individual (i.e., consistent across tasks); it must be heterogeneous across the sample (ideally, roughly equal numbers would have each style); it must not be highly correlated with skill (see Section S5 for more details and alternative definitions). Our specific measure of cognitive style ("optimizer" vs. "resolver") had an average test-retest reliability of 0.74 (see Supporting Information, page 19), which falls within the recommended range of 0.7–0.9 (Davidshofer & Murphy, 2005). The other measures of cognitive style that we reported in the Supplementary Materials also show considerable test-retest reliability (0.74 for constraint violation tolerance, and 0.71 for preference for efficiency vs. perfection). These survey measures of cognitive style are not only reliable, but also appear to be valid representations of the underlying behavior as measured in phase 1, as the survey answers are strongly correlated with the relevant behaviors during the task (see Section S5 for details).

2.3. *Phase 2 of the experiment*

In phase 2, we recruited the same 1200 participants and allowed 828 of them (as per our preregistration; see Section S1 for sample sizes) to perform a second sequence of five room-assignment tasks (task sequence is randomized) distinct from those completed in phase 1, also of varying complexity (very low, low, moderate, high, very high; all tasks timed out at 10 min in phase 2, regardless of complexity). Based on each participant's phase 1 labels for skill and social perceptiveness (participants were labeled as "high" skill or social perceptiveness if their level exceeded the median, and "low" otherwise), we first assigned each individual into one of the six blocks: HH (all individuals in this block are classified as high skill and high social perceptiveness, $N = 100$); MH (contains a mixture of high/low skill individuals

with high social perceptiveness, $N = 213$); LH (all individuals in this block are classified as low skill and high social perceptiveness, $N = 90$); HL (all individuals in this block are classified as high skill and low social perceptiveness, $N = 97$); ML (contains a mixture of high/low skill individuals with low social perceptiveness, $N = 221$); and LL (all individuals in this block are classified as low skill and low social perceptiveness, $N = 107$). Next, within each block, individuals were randomized to one of two conditions: “group,” in which groups of three randomly selected individuals from the same block were assigned to solve the problem collectively and had the ability to communicate with each other via text-based chat ($N = 591$ participants, forming 197 groups of size 3; data for one group are incomplete, leading to the number of valid interacting groups being 196); and “individual,” in which individual participants solved the problem independently and without communication with others ($N = 237$ participants; data from three individuals are incomplete, leading to the number of valid independent individuals being 234). As noted above, in previous work (Almaatouq et al., 2021), we compared the performance of groups with individuals for different levels of complexity. In contrast, our focus here is exclusively on the effects of compositional differences between interacting groups, hence our analysis only utilizes data from the group condition (see Fig. 1 for overall experimental design; Section S2.4.4 for details on experiment phase 2 design). The main purpose of the block randomization scheme was to oversample statistically less frequent combinations (e.g., all group members had high skills or high social perceptiveness), thereby increasing the statistical power of our experiments; within mixed skill blocks (MH and ML), we rely on the natural variance of skill diversity arising from simple random selection within the block. A secondary benefit of the block randomization approach was that it allowed us to match the distributions of participant skill and social perceptiveness levels in phases 1 and 2 (see Section S2.4.5). Although the median was used to label participants as having high/low skill and social perceptiveness for the sake of block randomization, all analyses use the continuous values of these metrics.

2.4. Performance evaluation

In phase 2, we used three metrics to capture performance in a room assignment task instance: **(1) normalized score**, defined as the actual score obtained in a task instance divided by the maximum possible score for that task; **(2) duration** (or time to completion), defined as the time elapsed from the start of the task until a solution was submitted (or until the task times out at 10 min); and finally, **(3) efficiency**, defined as the normalized score divided by the duration.

All three metrics are natural performance indicators that one might wish to optimize under some circumstances. In the absence of time constraints, for example, normalized score is an obvious measure of solution quality. By contrast, duration is appropriate when the problem-solving time is more important than quality (e.g., quickly come up with a reasonably good plan for resource allocation in a disaster response), and efficiency is appropriate when both quality and speed are important (e.g., in product development).

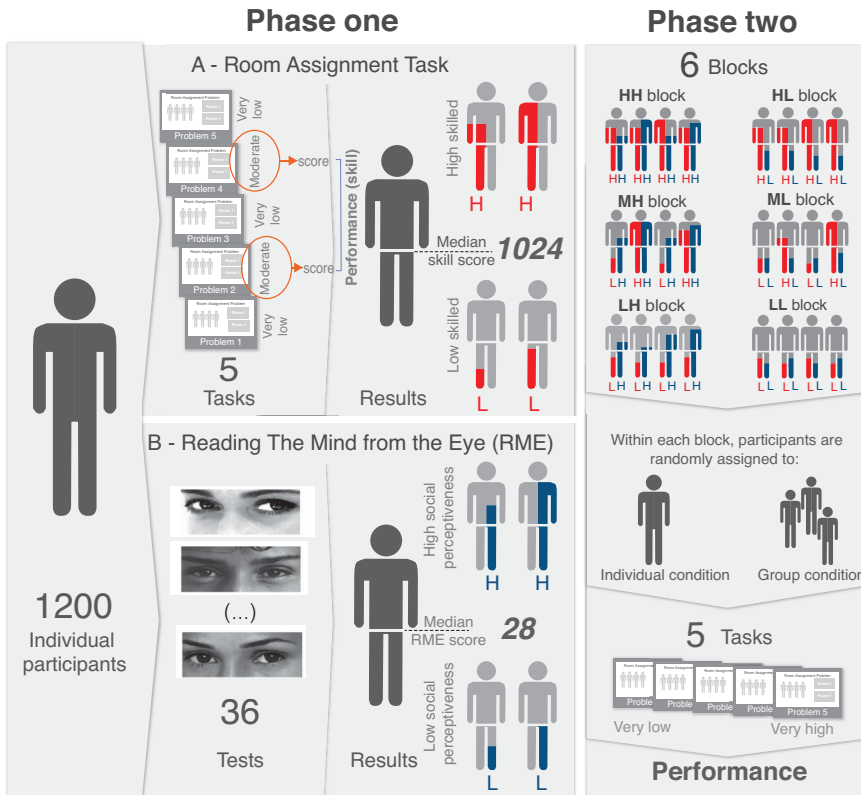


Fig. 1. **Schematic illustration of the study design.** In phase 1, participants completed a sequence of the “room assignment” task and a standard “Reading the Mind in the Eyes” (RME) test. In phase 2, the same participants were assigned to blocks based on skill and social perceptiveness (variation in cognitive style diversity arises only through randomization within blocks), then randomized into “individual” or “group” conditions within blocks before performing the second sequence of five room-assignment tasks.

3. Results

3.1. The effect of group composition on performance

Fig. 2 shows the absolute and relative effects of all preregistered independent variables on group performance, quantified as normalized score (Fig. 2a), duration of completion (Fig. 2b), and efficiency (Fig. 2c). All three metrics are standardized within each task instance as per our preregistration² (see Section S1).

Averaging across all complexities, Fig. 2a shows that higher levels of average skill, skill diversity, and social perceptiveness—but not cognitive style diversity—were associated with

² In a deviation from our preregistration, cognitive style diversity is not standardized and is kept as a binary variable for ease of interpretation; this choice does not change any of the qualitative conclusions presented here.

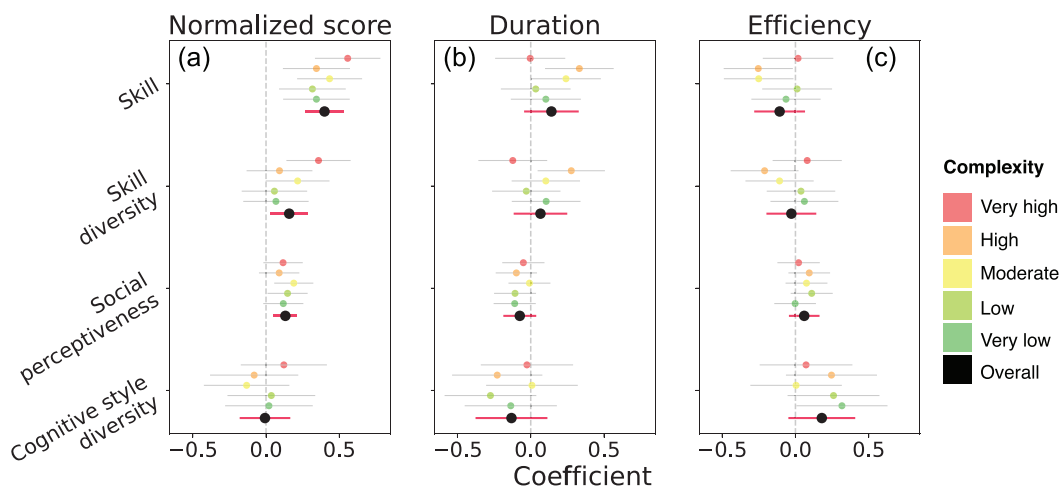


Fig. 2. **Group composition and group performance.** Standardized regression coefficients (OLS within each complexity, mixed effects model for “overall,” which combines data across complexities) for skill, social perceptiveness, skill diversity, and cognitive style diversity as a function of task complexity when predicting (a) normalized score, (b) duration, and (c) efficiency. Error bars indicate 95% confidence intervals. Group performance is standardized within each complexity, while group composition factors are standardized across groups; cognitive style diversity is kept binary. See Section S6 for regression description and tables.

significantly higher group scores ($p < .001$ and 95% CI [0.268, 0.530], $p = .017$ and 95% CI [0.029, 0.287], $p = .001$ and 95% CI [0.054, 0.210], respectively). Meanwhile, Figs. 2b and c show that, after accounting for multiple comparisons, we are unable to detect a statistically significant effect of any of the four group composition factors on task duration or efficiency, respectively (for regression tables, see Section S6). The observation that average individual skill, social perceptiveness, and skill diversity all have positive and significant effects on normalized score is consistent with both the meta-analytical studies that emphasized ability (Bell, 2007; Devine & Philips, 2001; Riedl et al., 2021; Stewart, 2006) as well as the experiments that highlighted the role of social perceptiveness (Engel et al., 2014; Kim et al., 2017; Lillis, 2007; Woolley et al., 2010) and the analytical models that played-up skill diversity (Hong & Page, 2004). The absence of a significant effect for cognitive style diversity is surprising inasmuch as the performance benefits of diversity are believed to derive specifically from a group’s collective ability to think about a problem in qualitatively different ways, a mechanism that is often implied in discussions of diversity (Bendor & Page, 2019; Hong & Page, 2004) even if it is not explicitly represented this way in the underlying formal models.

The absence of any significant effect of group composition on task duration is also nonobvious. One might have expected, for example, that more highly skilled groups would perform both better and faster, but if anything Fig. 2b suggests that higher skill is associated with slower task completion, where interestingly this effect seems to dominate the corresponding gain in score when computing efficiency. Likewise, one might have expected that higher

social perceptiveness would lead to better coordination and hence speed. In this case, the point estimate is consistent with intuition but the true effect, even if nonzero, is likely small.

Focusing again on normalized score (Fig. 2a), we take advantage of our novel design, in which all group composition factors were varied simultaneously, to compare the relative magnitudes of the three statistically significant effects. Strikingly, we find that the average effect of skill on normalized score is more than twice as large as that of skill diversity (Wald chi-square test; $\chi^2 = 32.23$, $p < .001$) and three times as large as that of social perceptiveness ($\chi^2 = 10.33$, $p = .0013$).³ Differences of this magnitude offer important context for recent accounts of non-skill-based factors such as diversity and social perceptiveness. For example, while it may be accurate to say that social perceptiveness exerts a significant and potentially even large effect on team performance, it does not necessarily follow that the optimal strategy for a manager is to focus on identifying “team players” versus simply hiring the most skilled workers. If anything, the three-fold difference in effect of average skill vis-à-vis social perceptiveness suggests the opposite; this in-sample estimate of relative importance is in directional agreement with previous studies focusing on the relative contribution of team composition factors (Weidmann & Deming, 2021).

Unfortunately, while standardized regression coefficients are helpful for comparing effect sizes, they are not well suited to making this sort of comparison; thus, we now reframe the problem in terms of predictive accuracy (Hofman et al., 2021a; Hofman, Sharma, & Watts, 2017; Rocca & Yarkoni, 2021; Salganik et al., 2020; Yarkoni & Westfall, 2017). To illustrate, imagine a hypothetical manager who wishes to compose a group for some task and who has prior information about the skill, cognitive style, and social perceptiveness of prospective group members. In essence, the manager’s task is to predict the group performance of different combinations of individual traits. Specifically, the manager cares about two related questions. First, what is the predictive accuracy of her “model” (i.e., how much of the observed variance can be accounted for by all independent variables in combination)? Second, what fraction of overall predictive performance is accounted for by each independent variable? The answer to the first question quantifies the extent to which group performance depends on the observed individual traits (vs. unobserved traits, factors external to the individuals, random noise, etc.), and hence to what extent it can be “engineered” at all. The answer to the second indicates which of the observed variables to prioritize when selecting group members, which is particularly important when there is a cost associated with measuring the relevant variables.

Addressing the first question regarding the degree to which group performance depends on the observed individual traits: With a simple list of covariates discussed in the literature, a linear regression achieved a Q^2 (a measure of goodness of prediction) of 0.24, meaning that the model “explained” about 24% of the out-of-sample observed variance in group performance—other models performed similarly (Joreskog & Wold, 1982; Quan, 1988). Using a longer list of task-specific covariates measured in phase 1 (including demographics

³ We note that because our task has bounds on the score a group can achieve, group-level average skill and skill diversity are highly correlated: as the group’s average skill approaches the maximum, so must the members’ individual skill levels. Consequently, we have reduced power in detecting the effect of these group composition factors.

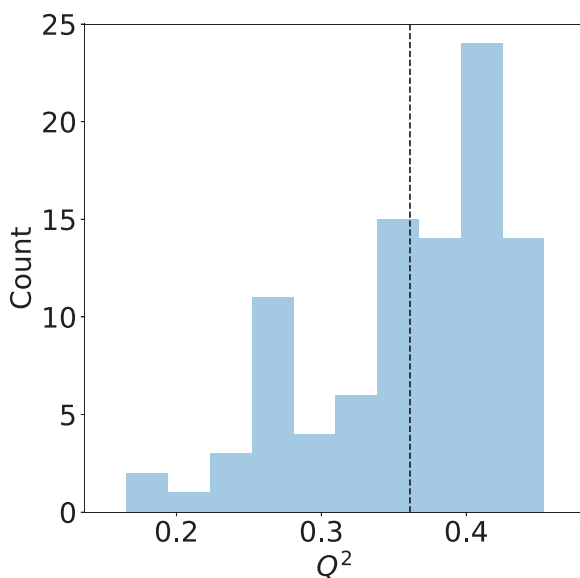


Fig. 3. **Out-of-sample (OOS) prediction of phase 2 performance using a range of models.** Using 19 features capturing the demographics, cognitive style, and performance of group members as measured in phase 1, we conduct a search over 100 models of varying structure and parameters using an “automatic machine learning” framework to predict the group’s phase 2 score. The figure shows the distribution of the Q^2 (a measure of out-of-sample predictive power) achieved by the models. The highest-performing model achieves a Q^2 of 0.45, while the average Q^2 across all models is 0.36 (indicated by the dotted vertical line). See Section S7 for details of the variables and model search procedure.

such as gender and age, behavioral measures of cognitive style, alternative definitions of surveyed cognitive style, and phase 1 task performance), an automated search over 100 models of varying structure (e.g., generalized linear models, gradient boosted trees, and ensembles thereof) and parameters, as described in LeDell and Poirier (2020), yields a model configuration that reaches a Q^2 of 0.45 when predicting the group’s score in phase 2, with the distribution of model performance centered around Q^2 of 0.36 (see Fig. 3; see Section S7 for details of the procedure). These figures are similar to those of recent out-of-sample prediction attempts in the social sciences, such as individual life-course outcomes ($0.03 \leq R^2 \leq .23$) (Rigobon et al., 2019; Salganik et al., 2020) and the size of Twitter cascades ($R^2 \simeq .4$) (Martin, Hofman, Sharma, Anderson, & Watts, 2016). As with these previous studies, the finding that at least two-thirds of observed variance cannot be explained even by state of the art models with access to a large number of precisely measured covariates suggests that group performance is a noisy phenomenon that may be subject to a fundamental “limit to prediction” (Martin et al., 2016), and hence explanation (Hofman et al., 2017), even for the relatively simple, controlled case presented here.

To address the second question of the overall predictive performance being accounted for by each independent variable, we measure the value of a given feature to out-of-sample prediction as its permutation importance (Breiman, 2001). To adapt this procedure to the Q^2

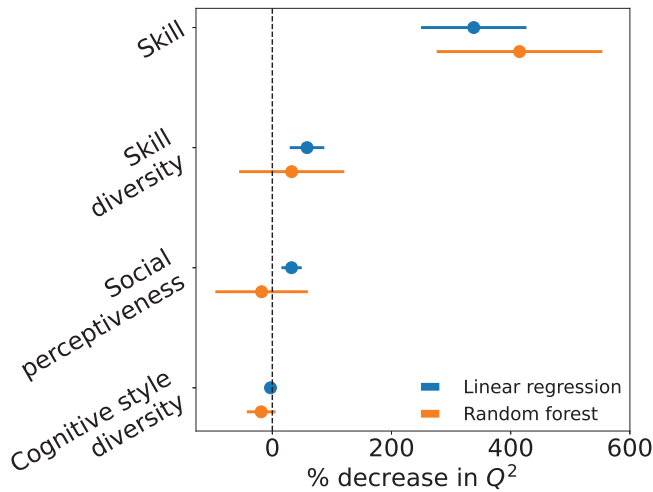


Fig. 4. **Out-of-sample permutation feature importance.** For each model, a feature’s out-of-sample permutation importance is measured as the decrease in Q^2 caused by randomly shuffling the feature during out-of-sample evaluation; this decrease is reported relative to the model’s out-of-sample performance on the unshuffled data. Each feature is shuffled 30 times, and error bars indicate the 95% confidence intervals. See Section S7.2 for a detailed explanation of the procedure and its application to other models and a broader list of features.

metric, we first fit one model per data point, while withholding that point during training (i.e., 196 models total), and measure the baseline Q^2 using the predictions of these models on their respective held-out data points. Then, to measure a feature’s importance to out-of-sample prediction, the feature is randomly shuffled across observations, and the Q^2 in predicting this partially permuted test set is compared to the base performance of the model on the original data. The features that lead to the largest decrease in out-of-sample model performance when permuted are then considered the most important to out-of-sample prediction. In both the literature-driven and extended sets of features, the average skill level of the group’s members consistently emerges as the most important feature—among the features shown in Fig. 4, average skill is approximately five times as “important” to out-of-sample prediction as skill diversity is, and nearly 12 times as “important” as social perceptiveness. Returning to our motivating example, in other words, faced with a choice between recruiting “team players” (i.e., high social perceptiveness individuals) and high skilled individuals, our hypothetical manager would be advised to focus almost exclusively on skill. This is not to say, of course, that for any pool of individuals of a fixed skill level, one would not want to select the best team players among them. Rather, it is just that when deciding how much effort to devote to increasing the average skill level of one’s pool versus the average social perceptiveness, the payoff to the former far outweighs the latter.

3.2. Compositional differences and collaborative behavior

To explore how groups interact to complete the task, we define two measures of collaboration at the group level: the overall number of chat messages sent within the group during

the game, and the group's turn-taking index. A group's turn-taking index for a given round is measured by dividing the number of turns taken (a turn is an uninterrupted sequence of room assignments made by a single player, each defining an intermediate solution) by the total number of solutions generated on a particular task instance. This measure is intended to differentiate between groups that collaborate in blocks (e.g., Player 1 moves N times, then Player 2 moves N times, then Player 3 moves N times) and groups that collaborate more dynamically (e.g., Players 1, 2, and 3 alternate moves, for a total of $3N$ moves)—in the first example, the number of turns taken is 3, and in the second example, the number of turns taken is $3N$, but the total number of solutions generated is the same in both cases. Contrary to our expectation, we find that both skill and social perceptiveness are not meaningfully associated with the standardized number of chat messages sent across all rounds (Bonferroni-corrected 95% CIs $[-0.187, 0.247]$ and $[-0.161, 0.273]$, respectively) nor the average standardized turn-taking index (Bonferroni-corrected 95% CIs $[-0.126, 0.204]$ and $[-0.233, 0.096]$, respectively, see Section S10). Intuitively, we would expect the conversation to be affected by the group's average social perceptiveness; here, it may be that the time constraints imposed by the task inhibit this mode of socializing and drive the conversation toward more task-related messages, which may not be mediated by social aptitude. Consequently, while the quantity of messages is not predicted by a group's social perceptiveness, the content of the messages could be. Similarly, while social aptitude may lead group members to give each other opportunities to participate (thus increasing their turn-taking index), the pressures of a time-sensitive task may overshadow such a dynamic. This post hoc exploratory analysis highlights the potential significance of these behaviors, and the observation that social perceptiveness is not associated with these behaviors in our data warrants a more focused study of these mechanisms.

3.3. Collaborative group behavior and performance outcomes

In this exploratory (i.e., not preregistered) analysis, we probe the relationship between patterns of collaborative group behavior and performance outcomes. As illustrated in Fig. 5, we find that any potential gains in score from additional communication across task instances are outweighed by the overhead time cost, leading to lower efficiency averaged across all task complexities (Bonferroni-corrected $p < .001$, 95% CI $[-0.484, -0.247]$, see Table S9). While communication enables groups to coordinate more complex strategies (e.g., division of labor, see Section S9), these strategies may be less efficient than expected given the complexity and time constraints involved in the task. Because communication ability was not experimentally manipulated, this measurement is potentially confounded but serves as suggestive evidence to motivate more detailed future work on the role of communication in group performance.

We also explore the association between turn-taking behavior and performance, and find that while turn-taking is not significantly associated with score (Bonferroni-corrected $p > .1$, 95% CI $[-0.086, 0.143]$), groups exhibiting more turn-taking behavior are considerably faster and thus more efficient on average (Bonferroni-corrected $p < .001$, 95% CI $[0.108, 0.301]$, see Fig. 5 and Section S8). This efficiency gain may arise from the increased capacity for “parallel processing”; while one group member acts, another can process the situation and take action instead of waiting for the first to finish (which would result in a possibly longer

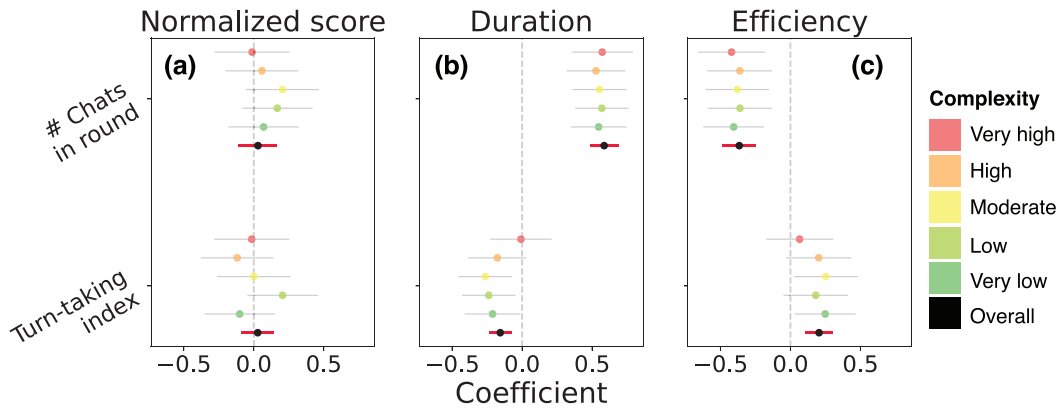


Fig. 5. **Group behavior and collective performance.** Standardized regression coefficients (OLS within each complexity, mixed effects model for “overall” which combines data across complexities) for the number of chat messages and the turn-taking index as a function of task complexity when predicting group performance. For any task complexity, groups that chat more do not score higher on average (a), but take longer to complete the task (b), leading to lower efficiency (c). On the other hand, groups that show more interspersed collaboration (higher turn-taking index) are faster and, consequently, more efficient on average. Error bars indicate the 95% confidence intervals with correction for multiple comparisons. See Section S8 for regression tables.

cycle of individual acting and thinking). Given the observational nature of this analysis, more careful manipulation of collaboration mechanics might be warranted to accurately assess turn-taking’s role in group performance; for example, future work could control the number of successive actions a given player may take.

4. Discussion

The successful pursuit of ambitious objectives such as national security, space exploration, or corporate management necessitates effective teamwork (Mathieu, Hollenbeck, van Knippenberg, & Ilgen, 2017; Mathieu, Maynard, Rapp, & Gilson, 2008; Wuchty et al., 2007). However, understanding team performance is challenging, as group behavior is influenced by many intricate interplaying processes and components, making it difficult to predict even with a good understanding of its individual elements. Therefore, we took a multi-level perspective, considering both the individual and team units of analysis, to investigate how individual attributes contribute to collective performance.

Our results provide mixed support for previous studies and highlight important building blocks, such as the two-stage design, block randomization, relative effect size comparison, and out-of-sample predictive performance, from which a research program could be constructed. For example, only by directly comparing the importance of group composition factors to out-of-sample prediction do we find that the “importance” of average individual skill far outweighs that of other factors, such as skill diversity, cognitive style diversity, and social perceptiveness, that have been emphasized in recent years. Naturally, this is not to say

that the importance of the average individual skill level will dominate in all contexts. For instance, access to the true skill level may—at least in many cases—be either unavailable or prohibitively difficult.

Second, our findings that **(a)** the effect of skill on collective performance is greater than that of skill diversity and **(b)** cognitive style diversity measures are neither positively nor negatively associated with performance appear to contradict widely cited claims (which are largely based on a theoretical model, rather than empirical results) regarding the performance benefits of diversity (Hong & Page, 2004; Page, 2008). Naturally, the lack of reliable diversity effect might not generalize to all types of tasks. Moreover, groups can be diverse with respect to attributes other than skill and cognitive style (e.g., demographics, specialized skills, worldview, etc.), and diversity can affect outcomes other than performance on a task (e.g., satisfaction, legitimacy, social equity, etc.). Thus, our results should not be construed as finding no effect of diversity in general. Nevertheless, they add to other recent results (de Oliveira & Nisbett, 2018; Eagly, 2016; Kurvers et al., 2019; Novaes Tump, Wolf, Krause, & Kurvers, 2018) that positive performance effects of diversity are surprisingly difficult to detect in carefully controlled empirical studies, and highlight the need for a research program that systematically varies task types (along with other contextual factors) while considering a wide range of group composition factors (and operationalizations thereof) to advance the basic science of collective problem-solving.

We recognize that these findings do not resolve all of the conflicting results that motivated this work—at least not in the general sense—and there are many other potential sources that contribute to the inconsistencies in this literature. For example, some theoretical constructs can be vague (e.g., what does “cognitive style diversity” mean?) or ambiguous (e.g., how do you operationalize cognitive style?), potentially causing different studies ostensibly about the same phenomenon (e.g., the impact of cognitive diversity on collective problem-solving) to measure quite different things. Another source of inconsistency could be the presence or absence of other mediating variables (i.e., multiple causes), or the misidentification of causal effects due to false-positive results (e.g., underpowered experimental designs, misspecified or faulty computational models) or bias in publications (e.g., incentives to find counterintuitive results).

We also note that the results of laboratory experiments, including ours, rarely translate directly into the real world. Obtaining results of immediate practical relevance would require running a far more extensive and complicated series of experiments than the one we have presented, one in which we would vary the available time, group size, task type, group interaction parameters, and many other potentially moderating variables. Nonetheless, our experiment is more realistic than previous work in one important sense: that when some hypothetical manager is faced with a situation where she must select individuals about whom she has some prior information to combine into a group, it is effectively out-of-sample predictive performance that she is seeking to maximize. In other words, if the problem we have studied did arise in a real-world context, then the quantity we are measuring, out-of-sample predictive power, would be exactly the quantity that a hypothetical manager would care about in weighing the different pieces of information available to her. In this sense, our work exemplifies a

“solution-oriented” approach (Watts, 2017): by forcing theoretical conjectures to confront the sort of practical questions that a manager trying to assemble a team might ask, our objective has been to advance basic understanding of collective problem-solving.

5. Materials and methods

The study was reviewed by the Microsoft Research Ethics Advisory Board and approved by the Microsoft Research Institutional Review Board (MSRIRB; Approval 0000019). All participants provided explicit consent to participate in this study, and the MSR IRB approved the consent procedure. Our experimental design, sample size, and analyses comparing the performance of groups with different composition were preregistered before the collection of the data (AsPredicted 13123). All other analyses are exploratory.

5.1. Statistical methods

To measure the effect of group composition factors on group outcomes, the composition measures (skill level, skill diversity, social perceptiveness, and cognitive style diversity) were first standardized at the group level, and the outcome measures (score, duration, and efficiency) were standardized within each of the five task complexity levels. The variables were standardized by subtracting the respective mean (e.g., across groups, or within complexity) and dividing by the respective standard deviation.

Effects within each complexity were then estimated by multivariate ordinary least squares regression, with fixed effects for each group composition factor. To estimate the average effect across task complexities (i.e., by pooling data from all tasks), we use mixed effects regression with fixed effects for group composition, and random effects at the group level, accounting for the nested structure of the data. Regressions relating to group behavior are conducted in the same manner, albeit with fixed effects for the behaviors instead of the group composition factors. These methods are detailed in Section S6. For exploratory analyses, we address the problem of multiple comparisons by a Bonferroni correction of the confidence intervals and p -values to maintain a family-wise error rate of 5%, where the “family” of hypotheses is that of all exploratory analyses in the study, the number of which is 91 (not all are included in the report).

Out-of-sample feature importance is measured by applying the permutation importance method outlined in Breiman (2001) to the Q^2 evaluation metric outlined in Joreskog and Wold (1982), and Quan (1988) (see Section S7.2 for details of the modified procedure).

5.2. Data and code availability

Replication data and code are available at the Harvard Dataverse, <https://doi.org/10.7910/DVN/T2ZNHE>. The experiment was developed using the Empirica platform (Almaatouq et al., 2021). The source code for the “room assignment” task can be found at

<https://github.com/amaatouq/room-assignment-csop>, and the source code for the “Reading the Mind in the Eyes” (RME) test can be found at <https://github.com/amaatouq/rme-test>.

Acknowledgments

The authors gratefully acknowledge the Alfred P. Sloan Foundation (G-2020-13924) for financial support.

References

- Aggarwal, I., & Woolley, A. W. (2019). Team creativity, cognition, and cognitive style diversity. *Management Science*, 65(4), 1586–1599.
- Almaatouq, A., Alsobay, M., Yin, M., & Watts, D. J. (2021). Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences of the United States of America*, 118(36), e2101062118.
- Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: A virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, 53(5), 2158–2171.
- AlShebli, B. K., Rahwan, T., & Woon, W. L. (2018). The preeminence of ethnic diversity in scientific collaboration. *Nature Communication*, 9(1), 5163.
- Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, 329(5995), 1081–1085.
- 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 and Allied Disciplines*, 42(2), 241–251.
- Baumann, O., Schmidt, J., & Stieglitz, N. (2019). Effective search in rugged performance landscapes: A review and outlook. *Journal of Management*, 45(1), 285–318.
- Bell, S. T. (2007). Deep-level composition variables as predictors of team performance: A meta-analysis. *Journal of Applied Psychology*, 92(3), 595–615.
- Bendor, J., & Page, S. E. (2019). Optimal team composition for tool-based problem solving. *Journal of Economics & Management Strategy*, 28(4), 734–764.
- Blazhenkova, O., & Kozhevnikov, M. (2009). The new object-spatial-verbal cognitive style model: Theory and measurement. *Applied Cognitive Psychology*, 23(5), 638–663.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Davidshofer, K. R., & Murphy, C. O. (2005). *Psychological testing: Principles and applications*. Hoboken, NJ: Pearson/Prentice Hall.
- de Oliveira, S., & Nisbett, R. E. (2018). Demographically diverse crowds are typically not much wiser than homogeneous crowds. *Proceedings of the National Academy of Sciences of the United States of America*, 115(9), 2066–2071.
- Devine, D. J., & Philips, J. L. (2001). Do smarter teams do better: A meta-analysis of cognitive ability and team performance. *Small Group Research*, 32(5), 507–532.
- Eagly, A. H. (2016). When passionate advocates meet research on diversity, does the honest broker stand a chance? *Journal of Social Issues*, 72(1), 199–222.
- Ellemers, N., & Rink, F. (2016). Diversity in work groups. *Current Opinion in Psychology*, 11, 49–53.
- 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), e115212.
- Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, 355(6324), 486–488.

- Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021a). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188.
- Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021b). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188.
- Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(46), 16385–16389.
- Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science*, *322*(5905), 1259–1262.
- Joreskog, K. G., & Wold, H. O. A. (1982). *Systems under indirect observation: Causality structure prediction*. Contributions to Economic Analysis. London, England: Elsevier Science.
- Kim, Y. J., Engel, D., Woolley, A. W., Lin, J. Y.-T., McArthur, N., & Malone, T. W. (2017). What makes a strong team? Using collective intelligence to predict team performance in League of Legends. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17* (pp. 2316–2329). New York: ACM Press.
- Kurvers, R. H. J. M., Herzog, S. M., Hertwig, R., Krause, J., Moussaid, M., Argenziano, G., Zalaudek, I., Carney, P. A., & Wolf, M. (2019). How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science of Advanced*, *5*(11), eaaw9011.
- Laughlin, P. R., & Adamopoulos, J. (1980). Social combination processes and individual learning for six-person cooperative groups on an intellectual task. *Journal of Personality and Social Psychology*, *38*(6), 941–947.
- LeDell, E., & Poirier, S. (2020). H2O AutoML: Scalable automatic machine learning. *7th ICML Workshop on Automated Machine Learning (AutoML)*. Retrieved from https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf
- LePine, J. A. (2003). Team adaptation and postchange performance: Effects of team composition in terms of members' cognitive ability and personality. *Journal of Applied Psychology*, *88*(1), 27–39.
- Lillis, M. P. (2007). Emotional intelligence, diversity, and group performance: The effect of team composition on executive education program outcomes. *Journal of Executive Education*, *6*(1), 41–54.
- Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics*, *42*(6), 553–568.
- Lo, A., Chernoff, H., Zheng, T., & Lo, S.-H. (2015). Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(45), 13892–13897.
- Martin, T., Hofman, J. M., Sharma, A., Anderson, A., & Watts, D. J. (2016). Exploring limits to prediction in complex social systems. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 683–694).
- Mathieu, J., Maynard, M. T., Rapp, T., & Gilson, L. (2008). Team effectiveness 1997–2007: A review of recent advancements and a glimpse into the future. *Journal of Management*, *34*(3), 410–476.
- Mathieu, J. E., Hollenbeck, J. R., van Knippenberg, D., & Ilgen, D. R. (2017). A century of work teams in the journal of applied psychology. *Journal of Applied Psychology*, *102*(3), 452.
- Moussaid, M., Noriega Campero, A., & Almaatouq, A. (2018). Dynamical networks of influence in small group discussions. *PLoS One*, *13*(1), e0190541.
- Mukherjee, S., Huang, Y., Neidhardt, J., Uzzi, B., & Contractor, N. (2019). Prior shared success predicts victory in team competitions. *Nature Human Behaviour*, *3*(1), 74–81.
- Novaes Tump, A., Wolf, M., Krause, J., & Kurvers, R. H. J. M. (2018). Individuals fail to reap the collective benefits of diversity because of over-reliance on personal information. *Journal of the Royal Society, Interface*, *15*(142), 20180155.
- Page, S. E. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies - New Edition*. Princeton, NJ: Princeton University Press.
- Quan, N. T. (1988). The prediction sum of squares as a general measure for regression diagnostics. *Journal of Business and Economic Statistics*, *6*(4), 501–504.

- Riedl, C., Kim, Y. J., Gupta, P., Malone, T. W., & Woolley, A. W. (2021). Quantifying collective intelligence in human groups. *Proceedings of the National Academy of Sciences of the United States of America*, 118(21), e2005737118.
- Rigobon, D. E., Jahani, E., Suhara, Y., AlGhoneim, K., Alghunaim, A., Pentland, A. S., & Almaatouq, A. (2019). Winning models for grade point average, grit, and layoff in the fragile families challenge. *Socius*, 5.
- Rocca, R., & Yarkoni, T. (2021). Putting psychology to the test: Rethinking model evaluation through benchmarking and prediction. *Advances in Methods and Practices in Psychological Science*, 4(3). <https://doi.org/10.1177/25152459211026864>
- Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., Morgan, A. C., Pentland, A., Polimis, K., Raes, L., Rigobon, D. E., Roberts, C. V., Stanescu, D. M., Suhara, Y., Usmani, A., Wang, E. H., Adem, M., Alhajri, A., AlShebli, B., Amin, R., Amos, R. B., Argyle, L. P., Baer-Bositis, L., Büchi, M., Chung, B.-R., Eggert, W., Faletto, G., Fan, Z., Freese, J., Gadgil, T., Gagné, J., Gao, Y., Halpern-Manners, A., Hashim, S. P., Hausen, S., He, G., Higuera, K., Hogan, B., Horwitz, I. M., Hummel, L. M., Jain, N., Jin, K., Jurgens, D., Kaminski, P., Karapetyan, A., Kim, E. H., Leizman, B., Liu, N., öMser, M., Mack, A. E., Mahajan, M., Mandell, N., Marahrens, H., Mercado-Garcia, D., Mocz, V., Mueller-Gastell, K., Musse, A., Niu, Q., Nowak, W., Omidvar, H., Or, A., Ouyang, K., Pinto, K. M., Porter, E., Porter, K. E., Qian, C., Rauf, T., Sargsyan, A., Schaffner, T., Schnabel, L., Schonfeld, B., Sender, B., Tang, J. D., Tsurkov, E., van Loon, A., Varol, O., Wang, X., Wang, Z., Wang, J., Wang, F., Weissman, S., Whitaker, K., Wolters, M. K., Lee Woon, W., Wu, J., Wu, C., Yang, K., Yin, J., Zhao, B., Zhu, C., Brooks-Gunn, J., Engelhardt, B. E., Hardt, M., Knox, D., Levy, K., Narayanan, A., Stewart, B. M., Watts, D. J., & McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, 117(15), 8398–8403.
- Shirado, H., & Christakis, N. A. (2017). Locally noisy autonomous agents improve global human coordination in network experiments. *Nature*, 545(7654), 370–374.
- Shore, J., Bernstein, E., & Lazer, D. (2015). Facts and figuring: An experimental investigation of network structure and performance in information and solution spaces. *Organization Science*, 26(5), 1432–1446.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.
- Stewart, G. L. (2006). A meta-analytic review of relationships between team design features and team performance. *Journal of Management*, 32(1), 29–55.
- Tsang, E. (2014). *Foundations of constraint satisfaction: The classic text*. BoD – Books on Demand.
- Ward, M. D., Greenhill, B. D., & Bakke, K. M. (2010). The perils of policy by *p*-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4), 363–375.
- Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour*, 1, 0015.
- Weidmann, B., & Deming, D. J. (2021). Team players: How social skills improve team performance. *Econometrica*, 89(6), 2637–2657.
- Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82.
- 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.
- Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744), 378–382.
- Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827), 1036–1039.
- Yahosseini, K. S., & Moussaïd, M. (2019). Search as a simple take-the-best heuristic. *Royal Society Open Science*, 6(10), 190529.
- Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 45, e1.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab of this article:
Supporting Information