Towards Stable Principles of Collective Intelligence under an Environment-Dependent Framework

by

Abdullah Mohammed Almaatouq
B.Sc., University of Southampton (2012)
S.M., Massachusetts Institute of Technology (2016)

Submitted to the Center for Computational Engineering & the
Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computational Science and Engineering
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2019

Author ..... Signature redacted .....
Center for Computational Engineering
May 3, 2019

Certified by ..... Signature redacted ...
Alex "Sandy" Pentland
Toshiba Professor of Media Arts and Sciences
Thesis Supervisor

Certified by ..... Signature redacted .....
John R. Williams
Professor of Information Engineering
Thesis Supervisor

Accepted by ..... Signature redacted .....
Heidi Nepf
Donald and Martha Harleman Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Accepted by ..... Signature redacted .....
Nicolas Hadjiconstantinou
Co-Director, Center for Computational Engineering

Towards Stable Principles of Collective Intelligence under an Environment-Dependent Framework

by

Abdullah Mohammed Almaatouq

Submitted to the Center for Computational Engineering
on May 3, 2019, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computational Science and Engineering

Abstract

A large body of work has shown that a group of individuals can often achieve higher levels of intelligence than the group members working alone. Despite these expectations of group advantage, many examples of collective failure have been documented—from market crashes to the spread of false and harmful rumors. To reconcile these results, a major effort in the study of collective decision making has been focused on understanding the role of group composition and communication patterns in promoting the "wisdom of the crowd" or, conversely, leading to the "madness of the mob." In the past decades, much of this effort has been devoted to inferring the importance of a particular attribute, in isolation, by its capacity to explain the accuracy of collective judgments. In this thesis, we argue that such a perspective can lead to inconsistent conclusions: an 'incoherency problem.' We assert that the importance of an individual-level or structural attribute may change as a function of the environment in which the group is situated. Hence, we propose a research agenda to investigate the relative importance of the group composition and the structure of interaction networks under an environment-dependent framework. We show that under such a framework, we can reconcile previously conflicting claims from the collective intelligence literature and motivate a future research program to identify stable principles of collective performance. Although implementing such a program is logistically challenging, "virtual lab" experiments of the sort discussed in this thesis, in combination with emerging "open science" practices such as pre-registration, data availability, open code, and "many-labs" collaborations, offer a promising route forward.

Thesis Supervisor: Alex "Sandy" Pentland
Title: Toshiba Professor of Media Arts and Sciences

Thesis Supervisor: John R. Williams
Title: Professor of Information Engineering

Acknowledgments

I consider myself exceptionally fortunate to have been mentored by world-class scientists, surrounded by the sharpest colleagues, and supported by loving family and friends throughout my academic journey. It is difficult to express how much these people mean to me, both professionally and personally.

This work was completed under the supervision of my research advisor, Alex "Sandy" Pentland. I couldn't have asked for a better thesis adviser and mentor than Sandy. He never forced an agenda on my research, but instead he provided me with guidance and a nurturing environment in which I was able to develop my research interests and pursue my own path. His personal assistance and friendly nature has always made me feel at ease with him and I could always look to him for support. Sandy will continue to be a role model to me as an academic entrepreneur who engages with a wide range of different types of people and technical areas.

I want to thank Duncan Watts, a mentor and a committee member who challenged me to keep up with his fine taste in research questions and to maintain a high degree of scientific rigor. Duncan inspires me with his very sharp process of thought and ability to approach just about any problem. Most importantly, he provided me with a standard of excellence in research, a standard which I hope this dissertation and all of my subsequent research lives up to. It is an honor to work with one of your heroes, and I happen to be just such a lucky person.

I would also like to thank Serguei Saavedra, another mentor and committee member. Serguei's 1.873 Fundamentals of Network and Community Ecology class taught me how to question my most fundamental assumptions and helped in forming much of the balance and direction of this dissertation. Serguei's unique approach to research and teaching set an example that I will try to follow.

I have had many other mentors over the years as well, all of whom are special to me and inspired me not only by their research, but also by their professional and personal lives. In particular, I am thankful to Iyad Rahwan, Matthew Salganik, Christopher Bail, David G. Rand, Anas Alfaris, Sinan Aral, Marta Gonzalez, John R. Williams,

Khaled AlGhoneim, and James A. Evans. I am also grateful to Robert Irwin for teaching me the importance of communicating my ideas with clarity and precision.

This work would not have been possible without the support of my peers from the various programs and labs at MIT and other institutions, who profoundly influenced this research through discussions, brainstorming, and insightful criticism. Chief among these have been Alejandro Noriega Campero, Ahmad Alabdulkareem, Peter Krafft, Yan Leng, Tara Sowrirajan, Martin Saveski, Joshua Becker, Yoshihiko Suhara, Eaman Jahani, Abdulrahman Alotaibi, Mehdi Moussaid, Nicolas Paton, Niccolo Pescetelli, Moshe Hoffman, and Simone Cenci.

I am also grateful to the MIT students I have had the pleasure of mentoring through the Undergraduate Research Opportunities Program (UROP) at MIT.

To my fellowship sponsor, King Abdulaziz City for Science and Technology, and the Ministry of Higher Education: thank you for your generous support. I am lucky to come from a country that supports education in every way.

And now, to my friends. I am blessed to have too many of you to list without incurring the dual risks of the boredom of readers and the omission of very important friends: *they who should be here, already know; the curious browser need not.*

Next, to my parents Mohammed Almaatouq and Najah Alsayegh, who are my constant source of wisdom and inspiration. They have always encouraged me to do what makes me happy and demonstrated, through me, what it truly means to guide someone in life. They set a high bar for me as far back as I can remember and were always there for me when I needed them. This dissertation, and my current and future research, will always be in honor of them.

My most important acknowledgemnt is to my wife, Ghadeer, for her endless support and superhuman patience in all of my pursuits. Ghadeer had a huge and lasting positive impact on my quality of life, from her constant love and encouragement to my sleep schedule and work-life balance.

Above all, thanks to God—*Alhamdulillah*—who has given me everlasting joy and purpose in my life, surrounded me with family and friends, and blessed me with the abilities and opportunity to complete this work.

Contents

1 Introduction **23**

1.1 Premise: A Unifying Theory 24

1.2 Reality: An Incoherency Problem 27

1.3 Resolution: An Environment-Dependent Framework 28

1.3.1 Illustrative Example 30

1.3.2 Conceptual Reflections and Dissertation Organization 34

1.3.3 On the Shoulders of Giants 36

2 Varying Environmental Complexity **41**

2.1 Environments of Widely Varying Complexity 42

2.2 Experiment: Optimal Team Construction for a Complex Task 44

2.2.1 Experimental Setup 44

2.2.2 Design of Phase One Experiment 48

2.2.3 Design of Phase Two Experiment 53

2.2.4 Details of Analysis 56

2.2.5 Results 59

2.3 Discussion and Chapter Reflections 67

3 Non-Stationary Information Environments **71**

3.1 Adaptive Systems and Environmental Conditions 72

3.2 Conventional Wisdom on the Wisdom of Crowds 73

3.3 The Role of the Environment, Again 73

3.4 Experiment: Guess the Correlation Game 74

3.4.1 Experimental Design 74
3.4.2 Experimental Results 80
3.5 Numerical Model and Simulations 86
3.5.1 Model Specifications 86
3.5.2 Simulation Results 89
3.6 Chapter Summary and Reflections 91

4 The Virtual Lab: High-throughput Social Science 95
4.1 The Interactive Environment 96
4.2 Towards Expanding the "Lab Experiment" Design Space 98
4.2.1 Size: In Complex Systems, Large is Different 98
4.2.2 Timescale: Social Interactions Evolve Over "Time" 100
4.2.3 Complexity: The Parameter Space of Social Theories 101
4.3 Reflections and Conclusions 103

A Supplementary Information for Chapter 2 105
A.1 Room assignment task 105
A.2 Reading the mind in the eye 108
A.3 Screenshots of the instructions and comprehension check 109
A.4 Validity of participant's individual skill measure 114
A.5 Comparing participants in phase one and two 116
A.6 Performance as a function of the environment complexity 118
A.7 Group composition; Supporting tables 122
A.8 Out of sample prediction accuracy 124

B Supplementary Information for Chapter 3 127
B.1 Guess the correlation game 127
B.2 S1: Individual and collective error 131

List of Figures

1-1 Linking network structures and collective outcome. The green regions represent the feasibility domains (parameter space or values of confirmation bias rates compatible with the persistence of the belief in the community) of two network structures in a belief dynamics model. . . 33

1-2 Although the left domain (the efficient network) is larger than the right domain (inefficient network), what matters is the overlap between the feasibility domain and the characterization of the environment. 34

2-1 Schematic illustration of the experiment design. 45

2-2 The five task difficulty levels in phase two were characterized by the different number of students to be assigned, the number of dorm rooms available, and the number of constraints. Increasing the task difficulty (i.e., environment complexity) reduces the normalized score and increases the time it takes participants to submit an assignment. Data is combined across both individual and team conditions across all 6 blocks. Error bars indicate 95% confidence intervals. The effective normalized score of a feasible solution is 80% and the minimum time required for a solution to be submitted is one minute, hence the starting points of the Y axes. 58

2-3 Comparing performance across individuals, real teams, and nominal teams. Individual, real team, or nominal team data is combined across all 6 blocks and standardized within each task complexity level. Error bars indicate 95% confidence intervals. 61

2-4 Team composition and team performance. The effects of cognitive style diversity shown in the figure are for participants' cognitive styles in solving the room assignment task as defined by "optimizer" vs. "satisfier." Error bars indicate 95% confidence intervals. See Appendix A.7 for additional analyses on the effects of skill/cognitive style diversity. 62

2-5 Using linear regression (70% training, and 30% testing; randomized and repeated 5 times) to predict team's normalized score with team's skill level, skill diversity, social perceptiveness, cognitive style diversity, and the number of female team members. (A) Compares predictive performance for covariates regressed independently (i.e. in separate models; green symbols), and in a single model where covariates are added in order of increasing independent predictive performance (purple symbols). (B) Predictive performance for a single regression model where covariates are added in order of decreasing independent predictive performance. Error bars indicate 95%confidence intervals.. 66

3-1 An illustration of the overall experimental design.In study 1, the feedback level is fixed (i.e., full feedback) and network plasticity is manipulated (i.e., static network versus dynamic network). In study 2, plasticity is fixed (i.e., always dynamic network) and feedback is manipulated (i.e., no feedback, self feedback, and full feedback). 75

3-2 Illustration of the experimental conditions in study 1. Panel (A) depicts the Solo condition (i.e., no social information) where participants make independent estimates. This condition corresponds to the baseline wisdom of the crowd context. Panel (B) describes the Static network condition (i.e., social learning) where participants engage in a stage of interactive social learning, where they are exposed to the estimates of a fixed set of peers in real time. Panel (C) describes the Dynamic network (i.e., selective social learning) condition that adds the possibility for participants to choose who to follow and be influenced by in the next round. 76

3-3 Guess the Correlation Game. An illustrative examples of the scatter plots used in the experiment is shown in Panel (A). Task difficulty, therefore, could be varied systematically at the individual level by varying the number of points, linearity, and the existence of outliers. For any given round, all participants saw plots that shared an identical true correlation, but difficulty levels could differ among them as shown in Panel (B). Participants were not informed about the difficulty level they or other participants were facing. 78

3-4 Shock to the Information Environment. We provide a change in the environment after round 10 by changing the difficulty levels for the participants for the remainder of the experiment and thereby we simulate non-stationary distributions of information among participants. 79

3-5 Individual and collective outcomes. Groups connected by dynamic influence networks and provided with feedback incur substantially lower individual errors as shown in Panels (A) & (B); and collective errors in Panels (C) & (D). The reduction is notably larger and more significant in periods where networks had adapted to the information environment (i.e., rounds $[6, 10]$ and $[16, 20]$). Errors are normalized with respect to average errors in the *solo* condition within each study. Error bars indicate 95% confidence intervals. 81

3-6 Mechanisms promoting collective intelligence in dynamic networks. Panel (A) shows that the network becomes more centralized with time (Freeman global centralization—i.e., how far the network is from a star network). Panel (B) depicts the relation between performance (i.e., average error) and popularity (i.e., number of followers). Panel (c) shows the relationship between accuracy of initial estimate and confidence (i.e., resistance to social influence). Error bars indicate 95% confidence intervals. 82

3-7 An example of the network evolution in the experiment. The circle color represents performance. The size of each circle represents the number of followers (i.e., popularity). The dashed orange line is the distribution of estimates prior to social influence, the blue solid line is the distribution of post-social influence estimates, while the dashed vertical line is the true correlation. 83

3-8 Mean-variance trade-off. Mean and standard deviation of absolute errors incurred by *top-k* estimates during the adapted periods (rounds $\in [6, 10] \cup [16, 20]$). *Top-12* estimates correspond to the full-group mean, and *top-1* to the group's best individual. Within each condition, *top-k* trade-off curves first gain in both objectives, then trade off lower average error for higher variability, and finally regress in both objectives as $k \to 1$. Across conditions, for any $k \in [1, 12]$, groups in the *dynamic* condition outperformed groups in the *static* and *solo* conditions. Moreover, the full-group mean of *dynamic* networks averaged 28% lower error and 48% less variability than the best individual playing *solo* (*dynamic top-12* vs. *solo top-1*; $P < 10^{-2}$); and the best individual in the *dynamic* condition averaged 32% lower error and 38% less variability than her analogue in *solo* (*dynamic top-1* vs. *solo top-1*; $P < 10^{-2}$). Bars indicate 95% confidence intervals. 85

3-9 Traditional accounts of 'wisdom of crowds' phenomena assume unbiased and statistically independent signals among agents. In our model, we assume arbitrary (potentially biased) initial signals. 87

3-10 Evolution of collective error: wisdom of the crowd (WC) and wisdom of the dynamic network (WDN). Panel A) stationary distribution of information among agents. Panel B) non-stationary information environment, shocks to the information distribution introduced at $t = \{100, 200\}$ 90

3-11 As the nose level increases in the provided feedback, the collective performance degrades until it converges to the performance of the independent crowd. 91

3-12 Panel (**A**): Learning rates associated to different λ's, where colored bands show 95% confidence intervals. Panel (**B**): Effects of λ and ρ on collective error, where shades of orange indicate time-averaged collective error. Panel (**C**): Effects of λ and ρ on collective error, normalized per type of information environment (ρ column). 92

4-1 The Virtual Lab Framework. The figure illustrates the three conceptual dimensions for virtual lab experiments. 99

A-1 An illustration of the "room assignment" task used in phase one of the experiment. In this case, there are $N = 6$ students that need to be assigned to $M = 4$ rooms, while satisfying $Q = 2$ constraints. 106

A-2 An illustration of a more difficult "room assignment" task. In this case, there are $N = 18$ students that need to be assigned to $M = 8$ rooms, while satisfying $Q = 18$ constraints. 106

A-3 An illustration of phase two "room assignment" task that was done by a group of three individuals in phase two. 107

A-4 An illustration of the "Reading the Mind in the Eye" test used in phase one of the experiment. The participant is shown a pair of eyes and asked to choose the emotion that best describes what the individual in the picture is feeling or thinking of. 108

A-5 Participants who obtained a higher score on the two hard tasks in the phase one experiment (i.e., "high skill") outperformed participants who obtained a lower score on those two hard tasks (i.e., "low skill") on each single task instance. Error bars represent 95% confidence intervals. . 115

A-6 Comparing the distributions of phase one participants and phase two participants with respect to their skill (i.e., scores obtained in room assignment tasks) and social perceptiveness levels (i.e., scores obtained in RME tests). Left: comparison results for the pilot study; Right: comparison results for the main experiment. Gaussian kernels are used for kernel density estimation. 117

A-7 Varying the room assignment task difficulty vs normalized score. The five task difficulty levels were characterized by the different number of students to be assigned, the number of dorm rooms available, and the number of constraints. Data is analyzed separately for individuals and teams from each of the six blocks. Increasing the task difficulty reduces the normalized score for both individuals and teams of all skill levels and social perceptiveness. Error bars indicate 95% confidence intervals. 119

A-8 Varying the room assignment task difficulty vs duration. The five task difficulty levels were characterized by the different number of students to be assigned, the number of dorm rooms available, and the number of constraints. Data is analyzed separately for individuals and teams from each of the six blocks. Increasing the task difficulty increases the time it takes participants to submit an assignment for both individuals and teams of all skill levels and social perceptiveness. Error bars indicate 95% confidence intervals. 120

A-9 Varying the room assignment task difficulty vs efficiency. The five task difficulty levels were characterized by the different number of students to be assigned, the number of dorm rooms available, and the number of constraints. Data is analyzed separately for individuals and teams from each of the six blocks. Increasing the task difficulty reduces the efficiency for both individuals and teams of all skill levels and social perceptiveness. Error bars indicate 95% confidence intervals. 121

A-10 Out of sample predictions on the team's cumulative score. Predict the team's normalized score with the team's skill level, skill diversity, social perceptiveness, cognitive style diversity, and the number of female team members. Three models (i.e., linear regression, elasticNet, and random forests) are used. Models are first learned on 70% of the teams and then tested on the rest 30% of the teams. This procedure is then repeated 5 times. Error bars indicate 95% confidence intervals. In all models, the majority of the explained variance in team's normalized score can be attributed to the team's skill level. 124

A-11 Out of sample predictions on team's duration on tasks. Predict the team's duration on tasks with the team's skill level, skill diversity, social perceptiveness, cognitive style diversity, and the number of female team members. Three models (i.e., linear regression, elasticNet, and random forests) are used. Models are first learned on 70% of the teams and then tested on the rest 30% of the teams. This procedure is then repeated 5 times. Error bars indicate 95% confidence intervals. The set of independent variables can hardly be used to explain the variance in team's duration on tasks. 125

B-1 Participants in all conditions make independent guesses about the correlation of two variables independently. 128

B-2 Participants in the network condition engage in a an active social learning phase, where they are exposed to their ego-network's estimates in real time. 129

B-3 After each task round, participants in the feedback conditions see the appropriate level of feedback for the conditions. This figure illustrates the dynamic network condition with full feedback (i.e., as opposed to no-feedback or only self-feedback). In all of our experiments, the maximum number of outgoing connections is three. 130

B-4 Dynamic social influence benefits the performance of individuals in the crowd. (A) Kernel Density Estimate (KDE) of participants' individual performance (i.e., average error across all rounds) for the three experimental conditions. We find that participants in groups connected by dynamic influence networks (Dynamic condition) achieved 38% reduction in average error compared to participants in unconnected groups (Solo condition), and 12% reduction in average error compared to participants in groups connected by static influence networks (Static condition). Panel (B) compares the average performance of individuals across conditions. Two-sample t-tests show a significant difference between the average individual error of participants in the Solo and Static conditions ($P < 0.0001$), as well as between participants in the Static and Dynamic conditions ($P < 0.001$). Panel (C) compares the standard deviation of participant's individual performance across conditions, and shows that individual performance in groups connected by dynamic influence networks was, not only better on average, but also substantially more equal on its distribution among group members. . 131

B-5 Panel (A) shows individual errors in the full game and Panel (B) shows the error in the adapted period (i.e., periods [6-10] and [16-20]). The error for the initial guess in both panels is the same across conditions, however, the dynamic network condition incurs much lower errors in the adapted periods (as in Panel B). 132

B-6 Panel (A) shows the errors before the interactive estimation phase (i.e, pre-social learning). Panel (B) shows the errors after the participants revised their estimates in the static and dynamic network conditions (i.e., post-social learning). 133

B-7 The distribution of pre-social learning and post-social learning for the three conditions. 133

B-8 Dynamic social influence effect in individual rounds: Adaptive with time and reduces individual error. All error rates are post-social learning errors. 134

List of Tables

2.1 Main properties of the 5 room assignment tasks used in phase one of our experiment. 49

2.2 Summary of the six blocks that we used in phase two of our experiments. 54

2.3 Main properties of the 5 room assignment tasks used in phase two of our experiment. The order of tasks was randomized in the experiment. 54

2.4 Relation between team's average skill level and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's average skill level. All models included random effects for teams as intercept to account for dependence across tasks (i.e., random effects are clustered on each team, using team id as the identifier). Increasing a team's average skill significantly increases the team's score in solving CSOPs, but has no effect on duration or efficiency. 63

2.5 Relation between team's average social perceptiveness and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's average skill level. All models included random effects for teams as intercept to account for dependence across tasks (i.e., random effects are clustered on each team, using team id as the identifier). Increasing a team's average skill significantly increases the team's score in solving CSOPs, but has no effect on duration or efficiency. 64

2.6 Relation between team's skill diversity and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's average skill level. All models included random effects for teams as intercept to account for dependence across tasks (i.e., random effects are clustered on each team, using team id as the identifier). Increasing a team's average skill significantly increases the team's score in solving CSOPs, but has no effect on duration or efficiency. 64

2.7 Relation between team's cognitive style diversity and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's average skill level. All models included random effects for teams as intercept to account for dependence across tasks (i.e., random effects are clustered on each team, using team id as the identifier). Increasing a team's average skill significantly increases the team's score in solving CSOPs, but has no effect on duration or efficiency. 65

A.1 The relation between the team's cognitive style diversity (in terms of whether all team members are fast/slow problem solvers or both types exist in the team) and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's cognitive style diversity. All models include random effects for teams as well as the team's skill level category as an intercept to account for dependence across tasks. Increasing a team's cognitive style diversity has no effect on the team's score, but reduces duration 122

A.2 The relation between the team's cognitive style diversity (in terms of whether all team members have the same constraint violation tolerance or not) and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's cognitive style diversity. All models included random effects for teams as well as the team's skill level category as an intercept to account for dependence across tasks. Increasing a team's cognitive style diversity has no effect on the team's score, but reduces duration. 123

A.3 The relation between the team's cognitive style diversity (in terms of whether all team members are pragmatic/tenacious or both types exist in the team) and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's cognitive style diversity. All models include random effects for teams as well as the team's skill level category as an intercept to account for dependence across tasks. Increasing a team's cognitive style diversity has no effect on the team's score and duration. 123

Chapter 1

Introduction

A substantial body of work has shown that a group of individuals can often achieve higher levels of intelligence than their members working alone. For example, the classical concept of the "wisdom of crowds" articulates how—in a startlingly wide range of settings—the aggregate (e.g., average) estimate of a group is better than the estimate of the best-performing individual. Examples include financial markets, which provide a mechanism for revealing investors' private information in order to arrive at a global estimate of value, and democracy, which aggregates differences of opinion to reach a collective decision on who should lead us.

The use of these (e.g., teams, markets, polls, and votes) and related modern mechanisms is on the rise, finding applications in areas as diverse as problem-solving [136, 108] technological and economic forecasting [140, 215], crowdsourcing [100, 33, 194], product rating [192, 144], public policy design [137, 176], and mapping natural disasters [133, 70]—just to mention standouts. At the same time, there are many instances of collective failure—from market crashes to the spread of false and harmful rumors. Such collective decision systems are central to the way society organizes and allocates resources; hence, providing a sound understanding of and useful design guidelines for improving the performance of collective decision systems is of paramount importance.

Although recent availability of massive digital traces on human behavior and the ubiquity of computational approaches have both extended and changed classical social science inquiry [174] bringing the era of computational social science [117] and the

emergence of network science [22, 209] (see Section 1.1). These advances have allowed scientists to generate a tremendous number of studies on *collective intelligence*, but they have been much less successful at reconciling some of the many inconsistencies and contradictions amongst them. For instance, studies have shown that the same attribute of interest (e.g., social interaction via communication networks, cognitive style diversity of team members, etc.) can either promote the "wisdom of the crowd" or, conversely, lead to the "madness of the mob." In general, for the same social context being studied and for the same global feature of interest, different theories have disagreed on which attributes are most relevant, and empirical studies offered an overwhelming lack of consistent evidence (see Section 1.2 on the *incoherency problem*). However, I argue that many of the studies on this topic only consider, explicitly or implicitly, static and stable environments, offering at best a partial view of human collective decision making (see Section 1.3 on the need for an *environment-dependent framework*).

In this dissertation, I address the question of the determinants of collective intelligence using an illustrative example (see Section 1.3.1) and a series of human experiments and supporting simulations (see Chapters 2 and 3). The results show that *what is optimal* always depends on the *environment*, and that groups provided with appropriate learning mechanisms can adapt to biased and non-stationary information environments, significantly improving both individual and collective judgments. The findings presented in this thesis can help reconcile some previously conflicting claims from the collective intelligence literature and motivate a future research program to more systematically identify stable principles of collective performance (see Chapter 4).

1.1 Premise: A Unifying Theory

Many scientists continuously aspire to discover universal principles that are valid across many different systems. While the domain of physical systems has offered examples of such widely applicable "laws," social phenomena have tended to be less

fruitful in terms of generating such generalizations. This desire to build models of social phenomena that are as predictive as those in physics, as well as the pursuit of unifying principles and operationally meaningful theorems in the social sciences, has has been termed "physics envy" in the social sciences [123, 125, 47]. While physicists can explain most of all observable physical phenomena using Newton's three laws of motion, social scientists (probably) wish they had 99 laws that explain 3% of human behavior. It is not only that social science has one theory for one thing and another theory for another thing [94], but rather that it has many theories for the very same thing [210].

The unfavorable comparison of social sciences to the natural sciences (and physics in particular) has a long [141] and quite unproductive history (e.g., see Watts argument against it in [210]). However, is it possible that this state of affairs has changed with the study of *computational social science* and *complex networks* emerging into prominence?

The Era of Computational Social Science. Recent widespread adoption of electronic and pervasive technologies, the development of e-government, and open data movements have enabled the study of human behavior at an unprecedented level and helped uncover seemingly universal patterns underlying human activity. Lazer, Pentland et al. [117] formally introduced *computational social science* (CSS) as a new field of research that studies individuals and groups in order to understand populations, organizations, and societies using *big data*[1], i.e. phone call records [4, 5, 11, 10], GPS traces [104], credit card transactions [184, 52], web page visits [59, 7], emails [105, 16], and data from social media [152, 6, 12, 8]. Driven by the ubiquitous availability of data and inexpensive data storage capabilities, the concept of *big data* has permeated the public discourse and led to surprising insights across the sciences and humanities. Such understanding can answer epistemological questions on human behavior in a data-driven manner, and provide prescriptive guidelines for persuading people to undertake certain actions in real-world social scenarios. In particular, this availability of data over the past fifteen years has shed light on the important role

[1] I acknowledge that no one who works with such large-scale data likes the term *big data*.

networks play in human society.

The Emergence of Complex Networks. At least for half a century now, there has been a surge in data availability and the ability to represent different systems (e.g., physical, biological, social, and technological) as a collection of nodes (or entities) connected with each other according to specific link topologies. Examples range from the tiny intracellular system, which consists of different molecules signaling each other via chemical reactions, that determines our biological existence, to the enormous cosmic web composed of discrete galaxies held together by gravity that determines the fundamental structure of our universe. We also see networks between these two scales, from food webs that represent the who-eats-whom (or interdependence) between species in ecology to social actors—be they individuals, organizations, or nations—exchanging ideas and favors in a social system. Indeed, recent research efforts have revealed a number of distinctive structural properties that many networks seem to share across many domains. Such properties include the "small world" effect [211, 208], the right-skewed degree distribution [21], clustering [149], and community structures [80]. Considering the ubiquity of networks and their structural properties, much effort has been made to understand the relationship between network structures and a system's function. This is a topic that is of utmost relevance to the social sciences: *what is the role of social network structural properties in generating globally observable, dynamical features?* More relevant to the topic of this dissertation: what role could the group composition and communication structure between individuals play in generating collective intelligence?

While these advances have allowed social scientists to generate a tremendous number of studies and theories on many important topics, they have been much less successful at reconciling some of the many inconsistencies and contradictions amongst them (see Section 1.2 and [210]). I conjecture that a non-trivial portion of the recent scholarly work on collective intelligence is based on simplistic axioms from which one can derive seemingly mathematically rigorous universal principles, carefully calibrated simulations, and the very occasional (and narrowly conditioned) empirical tests of those theories. Thus, I will argue that we need a research agenda to inves-

tigate the relative importance of different theories under an environment-dependent framework. I hypothesize that under such a framework, we can reconcile those previously conflicting claims from the collective intelligence literature and motivate a future research program to identify stable principles of collective performance.

1.2 Reality: An Incoherency Problem

Over the past couple of decades, scientists have generated many studies on the topic of collective intelligence. Although many of these studies share similarities in empirical motivation and theoretical objectives (i.e., studying the same 'thing'), the prescriptive consequences of their findings are not only different, but are logically incompatible—that is, each makes assumptions and reaches conclusions that, if true, would render the other false [210]. Duncan J. Watts in his Nature Human Behavior article titled "Should social science be more solution-oriented?" highlighted that for any topic of which he has undertaken a great amount of studying—be it cooperation mechanism, organizational performance, collective action, network dynamics, systemic risk—one would likely encounter the problem of irreconcilable results [210]. In this dissertation, I argue that the topic of *collective intelligence* is no exception.

For instance, studies that focused on the patterns of social interactions on collective intelligence found that social influence can promote the "wisdom of the crowd" [26] and, conversely, lead to the "madness of the mob" [127]. Inefficient communication structures simultaneously enhance [116, 58] and hinder [136, 82, 26] collective performance. Weak bridging ties are advantageous for innovation [87, 165, 168], as well as the opposite—strong cohesive ties are more advantageous [202, 203, 164, 204]. Other studies have found that network structures can affect (i.e., promote or hinder) cooperation [162, 74, 38], while others report no relationship between network structure and cooperation levels [193]. Homogeneity of tie strengths have been found to be beneficial for coordination and also can derail it [38, 157, 158].

The issue of incoherency exists in many modeling and empirical settings, not just in those that focus on social network phenomena mentioned in the previous

paragraph. For instance, when it comes to the composition of the group (i.e., the attributes of the constituents), some studies have identified the individual ability of the group members as an important predictor of collective performance [187, 60, 27], while others report no or at best a weak relationship [217, 166, 154]. Skill diversity (i.e., variance in group members' ability) has been shown to both enhance [98] and handicap collective performance [18, 67, 60]. Similar inconsistencies arise for cognitive style diversity [118, 150, 3], social perceptiveness [217, 69, 72, 120], and even the relative performance of teams versus individuals [48, 190, 219].

In general, for the same social system being studied and for the same global feature (i.e., collective outcome) of interest, theories disagreed on which attributes are most relevant, and empirical studies have offered an overwhelming lack of consistent evidence for a direct effect of any such property. What is most troubling is not the coexistence of theoretical and empirical disagreements in the literature, but that such incoherence is barely noticed and little demand for reconciliation efforts has been put forward [210].

1.3 Resolution: An Environment-Dependent Framework

There are some notable efforts that have focused on providing a partial resolution to some of the inconsistencies in the literature [16, 24, 123, 15, 86]. The common theme of these attempts is the finding that *what is advantageous depends on the environments in which the social system is situated.* This dependence on environmental conditions is well established in other fields, like the modern investigations of ecological communities [40, 185, 191, 41, 37, 169, 172].

Therefore, I hypothesize that in order to systematically reconcile these contradictory results, we need to understand the relative importance of the determinants of collective dynamics under an environment-dependent framework. If the actors are the subjects, then the environment is the object (i.e., the stimulus). Therefore, the

environment itself is system-dependent (e.g., may vary from one social system to another). In the case of problem-solving and collective intelligence, the environment may be characteristics of the task (e.g., complexity, type, information distribution across agents, etc.); in a product diffusion setting the environment may be characterized by the thing being diffused (e.g., product characteristics); and in cooperation or coordination games the environment may be characterized by the incentives (the payoff matrix, the rate of interaction, mutation, etc).

So far, most theories of collective outcomes have been largely silent on the relevance of those environmental conditions. For example, within the context of collective intelligence, to what extent does the optimal communication structure depend on the characteristics of the task being performed? Different studies usually adopt different types of tasks; therefore, it is possible that the discrepancy in the results is partly due to varying task characteristics across studies. Typically, a large number of degrees of freedom involved in these analyses (e.g., choice of task parameters) limits the generalizability of the results. Thus, if studies focus on a specific set of task characteristics in order to infer the general importance of a structural attribute, it would lead to inconsistent conclusions. Also, all of these studies only consider, explicitly or implicitly, stable environments, offering at best a partial view of human collective outcomes.

I want to highlight that I do recognize that the environment is only one of the possible sources that contribute to the inconsistencies in this literature. For example, some theoretical constructs can be vague (e.g., what do you mean by "structural diversity"?) or ambiguous (e.g., how do you operationalize tie strength?), potentially causing different studies ostensibly about the same phenomenon (e.g., the impact of network diversity on innovation) 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).

Additionally, studying the relationship between attributes of interest and collective outcomes usually suffers from less than ideal empirical conditions. This method-

ological limitation also can give rise to inconsistencies. For instance, many of the empirical observational studies are correlational (i.e., no exogenous manipulations) and cannot account for self-selection or homophily. Experimental studies, on the other hand, offer a great degree of control and allow for the identification of causal effects, but they suffer from many constraints such as short duration, high cost, small scale, homogeneous participants, simplistic design, unrealistic and static tasks, etc. Such constraints limit the ability to explore the parameter space of social theories (i.e., systematically manipulating the environment) as well as the external validity of their findings.

I concede that it is not clear to me how much each of these possible factors contributes to the problem of incoherency. Therefore, in this dissertation, I will focus on the role of the environment and assume that the studies referenced above are not suffering from any of those other issues.

1.3.1 Illustrative Example

To illustrate how paying little attention to the environmental conditions can lead to inconsistent conclusions, I will use a structural stability approach common to the study of ecological communities [169, 40]. The idea of structural stability, as René Thom puts it, is a

> natural condition to place upon mathematical models for processes in nature because the conditions under which such processes take place can never be duplicated; therefore, what is observed must be invariant under small perturbations and hence stable.

As such, the structural stability approach focuses on studying the following question: how structurally stable is a system vis-à-vis environmental changes? In other words, does the qualitative behavior of a dynamical system change as a function of the parameters of the system itself?[2].

[2]I learned about structural stability from discussions with the members of the Structural Ecology group at MIT: Serguei Saavedra, Simone Cenci, and Chuliang Song.

While Thom explicitly refers to *mathematical models*, the same argument goes for lab experiments: social theories are rarely precise enough to estimate exact parameter values from empirical data (i.e., we can never duplicate nature in the lab). Thus, a robust test of even a single theoretical claim may require many experiments, each corresponding to a different set of parameters.

For illustrative purposes, let us take a simple belief dynamics model as an example of relating network structure (i.e., the attribute of interest) to a collective outcome. This particular toy model was developed to resemble a Lotka-Volterra (LV) system, in order to easily and directly associate the structure of the feasibility domain (i.e., the parameter region where the desired collective outcome is achieved—we will define this formally later) with the network structure. Our simple belief dynamics model is as follows:

$$Y_{t+1} = Y_t(X - \mathcal{A}Y_t)$$

In this model, the strength of beliefs in a community about some topic (e.g., the existence of supernatural agents) is represented by the n-dimensional vector Y_t, where $y_{i,t}$ corresponds to the strength/level of belief of individual i at time t. The temporal evolution of beliefs (e.g., how individuals update their beliefs in the next time step, Y_{t+1}) is a function of the beliefs at any given point Y_t, the vector of intrinsic attributes of individuals X (i.e., confirmation bias, which is the rate at which individuals increase/decrease their belief independently about the topic), and the interaction matrix $\mathcal{A}$ that captures the structure of social influence (i.e., the attribute of interest). Note that the confirmation bias *rates* are inherently linked to environmental conditions—in other words, the events that the individuals encounter in their environment. If we take our measure of collective outcome to be the persistence of the belief in the community (i.e., there are no non-believers at equilibrium), then this implies $Y_t^* = \mathcal{A}^{-1}(X - 1) > 0$. We can see that this condition will be satisfied as long as the vector of confirmation bias rates falls inside a feasibility domain constrained

by the interaction matrix [173, 40]. Formally, this domain is defined by:

$$D_F(\mathcal{A}) = \{X = y_{t^*,1}\alpha_1 + ... + y_{t^*,n}\alpha_n > 0\},$$

where α_i is the i^{th} column of the interaction matrix $\mathcal{A}$. Now, for simplicity (and to be able to depict the system graphically), let us assume that we have two types of individuals in this community (i.e., individuals can have one of two possible rates of confirmation bias[3]). Then we can view the system from the lens of its parameter space in Figure 1-1. The axes of Figure 1-1 represent the 2-dimensional parameter space of confirmation bias rates. The points e_1, e_2, and e_3 are three choices of confirmation bias parameter values. The colored regions correspond to the set of confirmation bias rates compatible with positive beliefs about the topic in the community (the necessary condition for the persistence of the belief). The size and shape of this region depend upon network structure (structures $\mathcal{A}_1$ and $\mathcal{A}_2$). In mathematical ecology, these regions are usually called the feasibility domain of a community [126].

In Figure 1-1, it is easy to see how three different investigations can reach different conclusions about the role of social network structure in the persistence of beliefs in a community. For instance, if the first investigator sets the confirmation bias values to e_1 then the conclusion that will be reached is that social structure $\mathcal{A}_1$ is superior to $\mathcal{A}_2$ when it comes to the persistence of beliefs. On the other hand, another investigator that sets the confirmation bias to e_2 will reach exactly the opposite conclusion. A final investigator choosing e_3 (or implicitly assuming no confirmation bias i.e., $e_{noBias} = [0, 0]$) will find no relationship between network structure and the persistence of beliefs.

In this example, one might be tempted to come to the general conclusion that network structure $\mathcal{A}_1$ is superior to $\mathcal{A}_2$ when it comes to persistence of beliefs because of the relative sizes of the feasibility domains of these two networks[4]—i.e., $volume(D_F(\mathcal{A}_1)) > volume(D_F(\mathcal{A}_2))$. This can be interpreted as follows: if the en-

[3] I want to highlight that this approach can be applied to any combination of structures/attributes of interests, environments (with an arbitrary number of dimensions), and models (including nonlinear functional responses) as demonstrated in [39].

[4] In some cases, the size and shape of the feasibility domain can be analytically investigated [171].

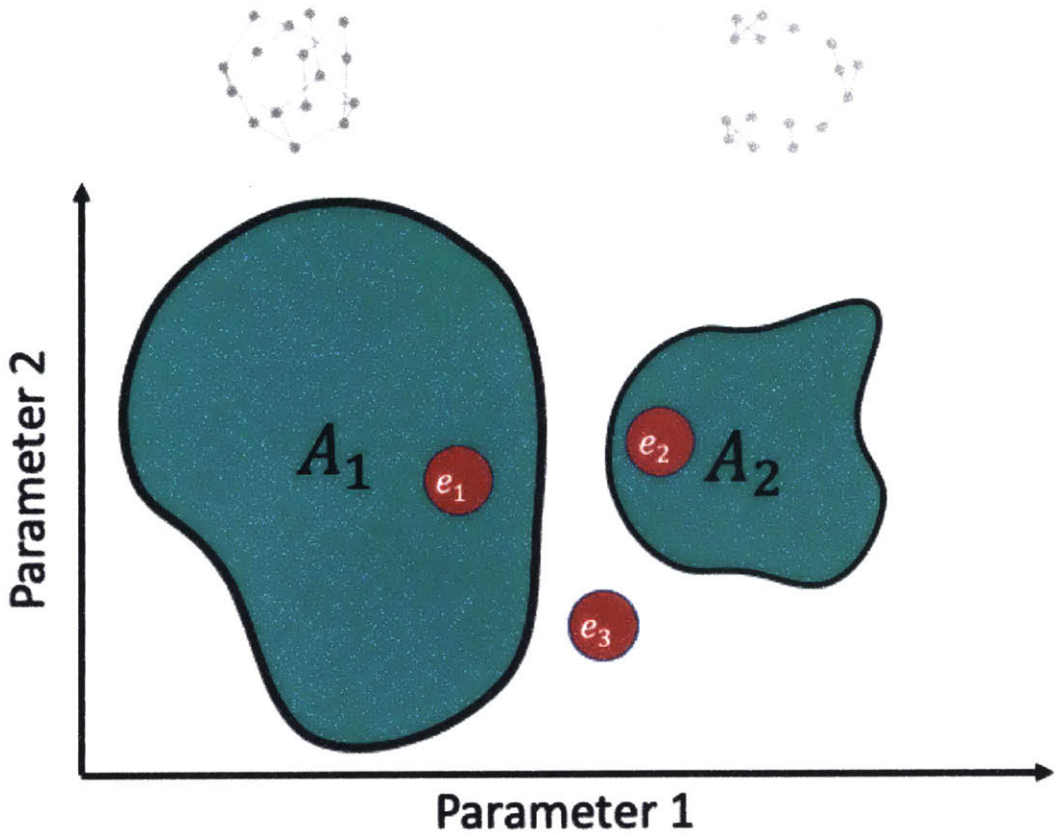


Figure 1-1: Linking network structures and collective outcome. The green regions represent the feasibility domains (parameter space or values of confirmation bias rates compatible with the persistence of the belief in the community) of two network structures in a belief dynamics model.

vironment values were uniformly sampled from the parameter space, then it is more likely to achieve the desired collective behavior under network structure $\mathcal{A}_1$ than $\mathcal{A}_2$.

However, this conclusion has no conceptual support, as the environmental conditions we care about are usually characterized by a distribution (e.g., set of environmental conditions over a period of time), rather than any particular point in the parameter space—that is, in the field, it is virtually impossible to measure the environment exactly. Therefore, what we care about is the overlap between the environmental conditions in a given setting/time and the feasibility domains defined by social network structures. In Figure 1-2, we can see that under different environmental conditions, what network structure is "best" can vary. Therefore, independent of the environment, there is no conceptual support of either a positive, negative,

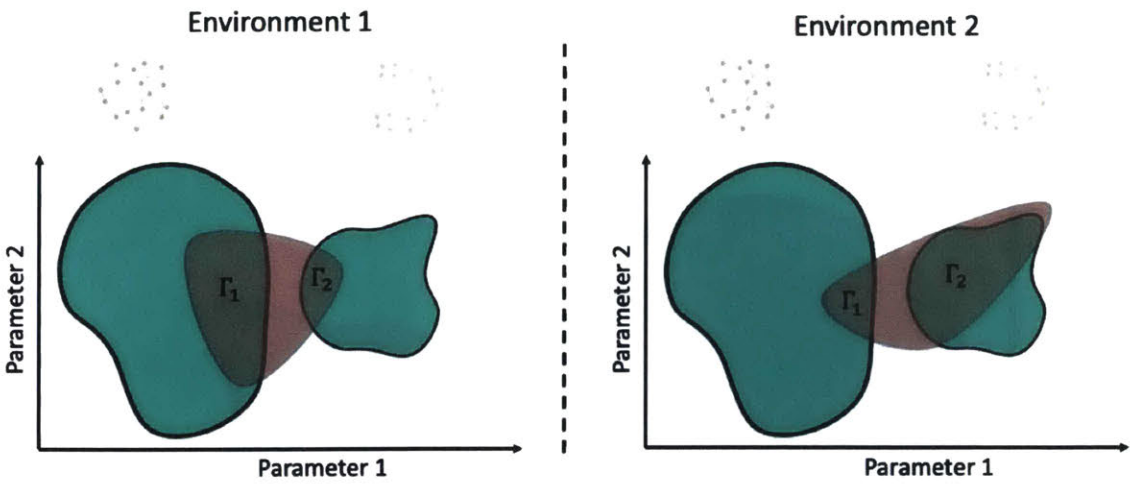


Figure 1-2: Although the left domain (the efficient network) is larger than the right domain (inefficient network), what matters is the overlap between the feasibility domain and the characterization of the environment.

or no association between network structure and function, even when we can fully characterize the feasibility domains.

1.3.2 Conceptual Reflections and Dissertation Organization

Overall, this simple conceptual analysis demonstrates that the association of a given attribute of interest (i.e., network structure, in this case) with global outcome depends on the environment. Therefore, without an environment-dependent framework from which to draw hypotheses and tune our intuitions, it is difficult to distinguish results that are unusual and interesting from results that are unusual and probably irrelevant (i.e., wrong or not generalizable).

Additionally, in some areas where there is a premium on slick studies with surprising results, 'surprising' should occur with reference to the particular region in the parameter space, and not in absolute terms. For instance, if a counterintuitive result can only emerge under very specific environmental conditions that are narrow and not representative of the conditions we care about or only occupy a small region in the parameter space that we rarely encounter, then how much does this result matter?

In the rest of this dissertation, I will continue to illustrate the importance of the

environment-dependent approach by conducting human experiments and simulations where we explicitly manipulate the environment (i.e., task characteristics in Chapter 2 and information distribution in Chapter 3). In order to conduct those studies, we built an experimentation platform that allows for conducting behavioral experiments of a scale, duration, and realism that far exceed what is possible in brick-and-mortar facilities and blur the line between lab and field experiments (see Chapter 4).

One of the main contributions of those studies is having a framework where the environment in which groups are situated is explicitly defined and manipulable (e.g., to simulate the non-stationarity of the environment). For instance, by manipulating the environment (and keeping everything else fixed), we can show how some of the seemingly contradictory results in the literature can be obtained as a function of the environmental conditions (i.e., whether nominal teams are better than real teams as in Chapter 2; or whether efficient network structures are more advantageous as in Chapter 3). Therefore, this allows us to reevaluate the importance of some of these attributes from the point of view of the environment (i.e., the conditions under which those attributes are of relative importance).

In particular, In the first study (i.e., Chapter 2), we focused on how different individual level attributes and group compositions (e.g., skill, cognitive style diversity, social perceptiveness) can affect collective performance, and examined whether those effects are robust to environments (i.e., tasks) of variable complexity. In this study, we asked two main research questions: 1) Do groups perform worse than comparable individuals on simple tasks but better on complex tasks?; 2) Do the effects of group composition on group performance vary with task complexity?

In the second study (i.e., Chapter 3), we focused on the role of dynamic communication structures on promoting collective intelligence. In recent years, both theoretical and experimental work has been limited mainly to frameworks where agents are placed in static social structures in stable environments. Yet, it is increasingly recognized that most natural and social systems are best described as "dynamic" networks, with links existing only intermittently in response to environmental variations (i.e., the different environments the group can be situated in). In Chapter 3, we shift the

focus on the role of communication networks from the purely structural aspects of the topology to the role environmental changes play in determining the dynamical processes defined on it. This would mean changing the usually ill-defined question "Which network structure is best to promote collective intelligence?" to "What mechanisms should we provide the social system to enhance its ability to adapt in a changing environment?" In the context of collective intelligence and group problem-solving, this dissertation overcomes some of the common limitations of prior studies by considering dynamical social influence networks where individuals can actively choose and dynamically rewire their social connections in non-stationary environments, which narrows the gap between stylized experiments and real-world social contexts.

In order to operationalize the "environment" in our lab settings, we built an experimentation platform (Empirica.ly). The platform forces the investigator to explicitly define the space of the environment in which the group of participants is situated, and therefore, the exploration of the interactions between the environments and the attributes of interest becomes more systematic (as opposed to having isolated and non-comparable studies). It is necessary to acknowledge that this remains a simplistic view of real social system environments. In real-world social systems (e.g., health, education, inequality, cultural norms, economic policies) environments are high-dimensional and interact in much more complicated ways in order to produce particular individual and group outcomes [210].

1.3.3 On the Shoulders of Giants

Samuel Taylor Coleridge, in *The Friend* (1828), wrote:

> The dwarf sees farther than the giant, when he has the giant's shoulder to mount on.

Indeed, academic advancements rarely happen in a vacuum, but transpire as we build on ideas and tools from others: a path-dependent wisdom of crowds. The work of this dissertation is inspired by and built upon a few common patterns that have emerged in several different research programs. Here, in addition to the structural stability

approach used for the illustrative example earlier[5], I will briefly mention several—non-exhaustive[6]—strands of academic research that lend support to the environment-dependent framework presented in this thesis.

Representative Design

> There is little technical basis for telling whether a given experiment is an ecological normal, located in the midst of a crowd of natural instances, or whether it is more like a bearded lady at the fringes of reality, or perhaps like a mere homunculus of the laboratory out in the blank. [34]

Egon Brunswik developed an innovative methodological framework called *representative design*, where he wondered why the logic we demand for generalization (i.e., sampling theory) over the subject side[7] is ignored when we consider the object side (i.e., conditions, stimulus, input, or environment)? In particular, he highlighted that one may only generalize the results of observations and experiments to those environmental conditions (or objects) that have been sampled in the experiment—in the same way that scientists apply this principle to the subjects (i.e., the participants) [61]. That is, to study the agent $\times$ environment relations, the environmental conditions should be sampled from the agent's natural environment in order to be representative of the population of environments to which it has *adapted* and to which the experimenter could generalize. Therefore, Brunswik called for an explicit theory of the environment in experimental psychology [85]. Similar to the discussion (in Section 1.3.1) on feasibility regions, Brunswik argued that experimenters should avoid oversampling highly improbable conditions (or conditions that do not exist in the population), because even if the results from those conditions are interesting, are they really relevant?

[5] I have already demonstrated how tools borrowed from the study of ecological communities [40, 173, 169] are valuable for viewing communities through the lens of their environmental variations.

[6] Interested readers should also refer to the No Free Lunch Theorem [102, 95, 216] and the Contingency Theory of Organizations [63].

[7] In actuality, as Henrich [93] pointed out, most participants in psychological experiments are WEIRD; also see Chapter 4.

Since Brunswik's time, the idea that human behavior is shaped by the environment structure has been generally accepted (e.g., ecological rationality, evolutionary psychology/game theory, ecological psychology) and the concerns with the limited generalizability of research findings have been expressed periodically. Nonetheless, there were two mainstream criticisms of the *representative design* approach that I want to highlight here. First, there are concerns regarding the associated difficulty of implementing representative designs [50]. How can one possibly sample situations? However, I would argue that the difficulty of sampling situations[8] can be overcome with modern technologies, such as the Web, to effectively reproduce and explore environments (see Chapter 4 on *high-throughput social science using virtual labs*).

The second objection argues that, even if we could define and sample the environment, there is no need to do so. After all, the goal of the social scientist is not to generalize the results from the experiment to situations 'outside' the experiment, but to test hypotheses and advance particular theories. This criticism is brought on by the strong emphasis on ensuring *internal validity* for the sake of replicability, at the expense of *external validity*. In other words, this objection presupposes that the purpose of social science experiments is not to *solve practical problems* in the real world.

Solution-Oriented Social Science

Duncan Watts has argued in a recent article that social science should be more "solution-oriented" in order to reconcile the competing claims in the literature (i.e., the incoherency problem in Section 1.2). That is, the research community needs to place more emphasis on solving practical problems—the sort with direct engineering analogues [210]—rather than the advancing of particular theories. For instance, in the article Watts suggests asking questions like:

- "How do I maximize the impact of my advertising spending?"

[8] I am referring to formal situational sampling [84], which focuses on the formal properties of the environment (i.e., defining the universe of possible environments), irrespective of its operationalization.

- "How do I increase productivity in my organization?"

- "How do I increase pro-social behaviour in my community?"

I want to argue that Watts's perspective is akin to the environment-dependent framework proposed in this dissertation. In all of these questions, the locust of activity (e.g., "in my organization") is limiting the generality of the answer to the objective (e.g., "how to increase productivity"). In other words, the answer to the first part of the sentence is dependent on the conditions specified in the second part. Hence, the answer will be relative, not absolute, which—I will argue—will lead to reliable and coherent results, not falsely conceived as universally valid.

Adaptive Market Hypothesis

Andrew Lo [123, 124] applies the principles of biological evolution (i.e., competition, adaptation, and natural selection) to financial markets. In particular, the approach focuses on explaining how emergent market attributes (e.g., prices) are related to the interaction of distinct groups of market participants within a specific environmental conditions (e.g., regulations, number of competitors, magnitude of profit opportunities).

In particular, the Adaptive Market Hypothesis asserts that market behavior adapts to a given financial environment, and an efficient market (the dominant theory of markets) is merely the steady-state limit of a market in a *static* financial environment; an idealized market is unlikely to ever exist in practice.

The Adaptive Market Hypothesis is specifically studying the individual-level investor (i.e., economic agent) as well as the larger market (i.e., macroeconomy). However, I think the adaptiveness to environmental conditions approach applies to other collective social phenomena, more generally, and for the same reasons (i.e., evolutionary processes working in a non-static environment). In this work, we see how similar ideas can expand beyond the domain of financial markets.

Ecological Rationality

Ecological rationality [198, 183, 81]—proposed by the German psychologist Gerd Gigerenzer of the Max Planck Institute for Human Development—in contrast to rational choice theory, maintains that the rationality of a particular decision depends on the context of circumstances in which it takes place. Therefore, what is considered rational under the rational choice account that focuses on agent characteristics (e.g., preference consistency) might not be considered rational under the 'ecological rationality' account, which also considers the structure of the environment.

This approach to decision-making is inspired by earlier work by Herbert A. Simon on heuristics and bounded rationality [181]. In particular, he explored how heuristics (a decision strategy that partially ignores available information) in appropriate *context* can achieve higher intelligence than other more complex approaches. The ecological rationality focuses on individual-level decision making, while in this dissertation we investigate the emergent phenomenon of collective behavior.

Chapter 2

Varying Environmental Complexity

Recent work on teams has emphasized the counterintuitive claim that the absolute skill level of team members matters less to collective performance than other factors such as skill diversity, cognitive style, and social perceptiveness. Through a novel two-phase experiment (phase one $N = 1200$, phase two $N = 828$; pre-registered[1]) in which individual on-task skill, cognitive style, and social perceptiveness were measured ex-ante and then systematically varied in team composition, we show that the effect of skill on team score is larger than all other factors across environments (i.e., tasks) of widely varying complexity. More importantly for practical applications, skill predicts twice as much out-of-sample variance as all other factors combined. We also show that while teams outperform comparable individuals on average, when compared with the best member from a same-sized group of individuals, teams score worse but compensate with faster completion time and higher efficiency when the task environment is complex. Our results help to clarify inconsistencies in the existing literature on the relationships between team construction and performance; they highlight the value of online experiments capable of supporting large sample sizes and complex, multifactorial designs; and they motivate a future research program to identify stable principles of collective performance (see Chapter 4).

[1]All of the data, analysis code and the pre-registration plan are publicly available at the Open Science Framework (OSF) repository. Our main hypotheses, experimental design, and analyses were pre-registered before the collection of the data (AsPredicted #13123). The study was reviewed and approved by the Microsoft Research Ethics Advisory Board (Approval#: 0000019).

2.1 Environments of Widely Varying Complexity

As organizations have moved inexorably to more team-based structures, the problem of improving team performance through judicious selection of team members has preoccupied management scientists and managers alike [115, 2, 101, 197, 145]. Previous research has found a variety of intriguing results regarding the impact of skill diversity [98], cognitive style diversity [68], and social perceptiveness [110, 217] on team performance. However, the existing literature exhibits two important limitations that undermine the practical relevance of these findings. First, the difficulty of executing large-scale experiments with complex multifactorial designs means that individual studies typically focus on single effects rather than comparing multiple effects directly; thus, it remains unclear which of many hypothesized relationships matter most in practice. Second, definitional ambiguity of quantities of interest (e.g., skill, cognitive style, collective performance) combined with researcher freedom to select among possible definitions create inconsistencies across published results for a given effect. For example, on the one hand, meta-analyses of lab studies conducted between the late 1960s and early 2000s find that average individual ability is the most consistent predictor of team performance [60, 187, 27]. On the other hand, more recent studies have argued strongly that average ability is less relevant to collective performance than other factors such as social perceptiveness (aka emotional intelligence) [217, 69, 120] and diversity [98, 13, 3]. Similar inconsistencies arise for more fundamental questions about the value of being in a team. For example, there is little consensus on whether teams always outperform independent individuals (i.e., the relative performance of teams versus individuals) [48, 219, 190].

Reading this literature, a hypothetical manager wishing to construct a team for some task (or environment) would have difficulty deciding whether for a particular task would team be less/more effective than their members, which of potentially many individual-level attributes to measure, how to optimally combine individuals with those attributes, and how that combination might depend on the difficulty of the task at hand. Moreover, because the effects of different combinations of attributes

are typically expressed in terms of regression coefficients, not their ability to predict the outcome of interest, it is unclear how much control over performance the manager could expect to exert in practice [122, 96, 220]. Here we address these limitations by using a novel two-phase experiment to answer two main questions: 1) Under what conditions, if any, do teams perform better than individuals? 2) Which of the four widely studied attributes of teams—average skill level, skill diversity, cognitive style diversity, and social perceptiveness—individually and collectively dominate team performance (effect size and predictive power) and does it vary with task complexity?

Our experimental design exhibits five important features that address limitations with previous studies and speak directly to the hypothetical manager's problem outlined above:

1. By varying the difficulty of the task (i.e., the environmental conditions) over a wide range (from "easy" to "super hard") without changing the nature of the task (see Sections 2.2.1, 2.2.1), we determine how, or if, the relative importance of different attributes changes with task difficulty (e.g., does being in a team, or the level of social perceptiveness, or skill diversity, matter more for the most difficult tasks than for easy tasks?).

2. The separation of the experiment into two "phases" eliminates confounding between individual and team measures of performance [48]. In phase one we measure all relevant attributes for individual workers (see Section 2.2.2); then in phase two we use this information to construct teams with desired combinations of individual attributes (e.g. "high skill but low social perceptiveness," "mixed skill but high social perceptiveness," etc.; see Section 2.2.3). In this way, all individual attributes are measured before assignment to teams or individuals.

3. Rather than relying on generic metrics for ability and cognitive style we measure them on the task itself, thereby improving reliability.

4. The combination of the two-phase design and the ability to recruit large samples online facilitates a relatively complex multifactorial design which in turn improves our ability to compare multiple effects directly (see Chapter 4).

5. Finally, we directly address the practical problem outlined above by expressing our team-level results in terms of their individual and combined predictive accuracy.

2.2 Experiment: Optimal Team Construction for a Complex Task

The goal of our experiment was to examine which of several factors (e.g., skill level, skill diversity, social perceptiveness level, cognitive style diversity, etc.) predicts team performance. To answer this question, we used a novel "two-phase" experimental design in which we recruited the same group of participants (recruited from Amazon's Mechanical Turk; see Section 2.2.1) twice to solve a sequence of Constraint Satisfaction and Optimization Problems (CSOPs)—a class of complex problems that are widely studied in artificial intelligence and operations research as abstractions of various real-world resource allocation and scheduling problems (see Section 2.2.1).

2.2.1 Experimental Setup

Specifically, participants were asked to solve a "room assignment" problem in which they had to assign N "students" to M "rooms" where each student had a specified utility for each room. Participants' objective was to maximize total student utility while also respecting Q constraints (e.g., "Students A and B may not share a room or an adjacent room"). Task difficulty (or the "environment complexity," therefore, could be varied systematically by changing the number of students (N), the number of rooms (M), and the number of constraints (Q). After completing five such tasks, each participant also completed a standard "Reading the Mind in the Eyes" (RME) test [25], which is commonly used as a measure of social perceptiveness (see Section 2.2.1).

On the high-level, phase one (see Section 2.2.2) was used for gathering ex-ante measurements of each participant's skill level on the room assignment problem, social

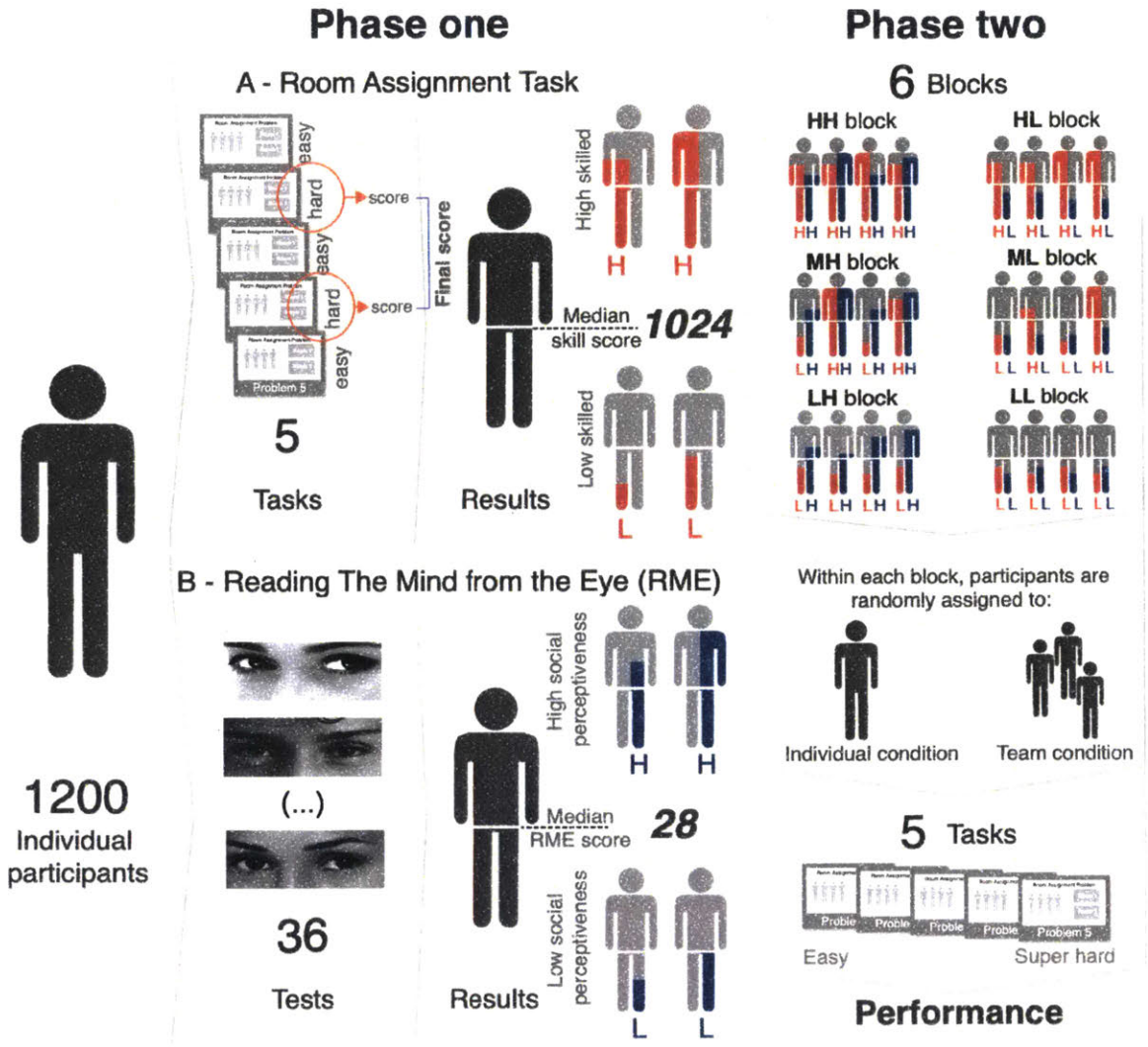


Figure 2-1: Schematic illustration of the experiment design.

perceptiveness level, and cognitive style. Then, in phase two (see Section 2.2.3), we deployed a block randomization scheme to randomly assigning participants into one of six blocks based on their phase one measurement results. Within each block, we randomized whether the participants will work as individuals or in teams (i.e., teams of three randomly selected participants) to solve another set of room assignment problems. Teams were also provided with a chat box, enabling them to communicate freely with each other and scores were now assigned to teams not individuals. See Figure 2-1 for an illustration of the experiment design

Finally, we used the ex-ante measurements from phase one to construct the inde-

pendent variables (i.e., whether the participants were assigned to individual or team condition as well as different influencing factors of team performance) and used the actual performance in phase two as the dependent variables, which together allowed us to examine the performance effects of being in a team versus an individual and the effect of different team compositions (see Section 2.2.4).

The experiment was developed using Empirica (https://empirica.ly/), an open-source "virtual lab" framework and a platform for running multiplayer interactive experiments and games in the browser [156].

The source code for the Room Assignment Tasks can be found at here, and the source code for the Reading the Mind in the Eye Test can be found at here .

Room Assignment Problem

In our experiments, we asked participants to solve room assignment problems, first individually and then within a team. A room assignment problem is a type of Constraint Satisfaction and Optimization Problem (CSOP, that is, an optimization problem on top of a constraint satisfaction problem) [79, 200]. We chose this task for three reasons. First, CSOPs are an abstraction of many resource allocation and optimization problems; thus, they capture important features of real-world team problem solving exercises without requiring participants to have specialized skills. Second, the payoff function for CSOPs can be described as a "rugged landscape" characterized by many locally optimal but globally suboptimal solutions. Correspondingly, CSOPs are amenable to potentially many solution strategies and styles, where no single strategy is universally superior. Third, the complexity of CSOPs can be systematically varied by adjusting a few key parameters; in our case, by changing the number of students N, the number of rooms M, and the number of constraints Q.

In our operationalization of this problem, participant(s) were tasked with assigning each of N "students" to one of M "rooms," while also respecting Q constraints on their choices (e.g., students A and B must be neighbors, must not share a room, etc.). In each room assignment problem, a "utility table" was presented, providing participant(s) with the information on students' ratings (between 0 and 100) to each

of the M rooms indicating how satisfied they would be if being assigned to the room. The participant(s) was then asked to find a room assignment plan that maximized satisfaction across all students without violating any constraints.

To incentivize the search for an optimal solution (i.e., the optimal room assignment plan) we provided participant(s) with additional bonuses based on how good their submitted solutions for the problem were. In particular, we defined the "score" of a room assignment plan as the following:

$$\begin{aligned} Score = {} & \text{The sum of students' ratings of their assigned rooms} \\ & - 100 \times \text{the number of violated constraints} \end{aligned}$$

By submitting a complete plan (that is, each student got assigned to one room) with a positive score in a room assignment problem, participant(s) could earn a "performance-based bonus" using a 500 points:\$1 USD conversion rate to exchange scores into payments. Moreover, if the submitted plan was indeed the optimal one, an additional \$0.5 USD "optimal assignment bonus" would be given [2]. We determined these values for the payments by conducting a series of pilot studies and observing how participant behavior responded to different payment schemes. For screenshots of the task, see Appendix A.1 and A.3.

Reading the Mind in the Eyes (RME)

Each participant also completed the revised version of the "Reading the Mind in the Eyes" test [69], a widely used test for measuring Social Perceptiveness/Emotional Intelligence. In this test, participants are shown 36 pairs of eyes. For each pair of eyes, they are provided with four words describing emotions. The participant is asked to select one of the four words that best describe what the person in the picture is thinking or feeling. See Appendix A.2 for an illustration of the test.

[2] These bonus rates are for phase one experiment. In phase two, the performance-based bonus conversion rate is 1000 points:\$1, while the optimal assignment bonus is \$0.7. We set these bonus rates to maintain a similar level of hourly payment between the phase one and two experiment.

Participants Recruitment

All participants were recruited on Amazon Mechanical Turk (MTurk[3]), which is an online labor market with a large and diverse pool of people ready to promptly perform tasks for pay (called human intelligence tasks, or HITs) [9]. We recruited our participants by posting a HIT for the experiment, entitled "Play games and get up to $17 in total pay," a neutral title that was accurate without disclosing the purpose of the experiment. The study was reviewed by the Microsoft Research Ethics Advisory Board and approved by the Microsoft Research Institutional Review Board (Approval#: 0000019). All participants provided explicit consent to participate in this study and MSR IRB approved the consent procedure. All data collected in the experiment could be associated only with the participant's Amazon Worker ID on MTurk, not with any personally-identifiable information. All participants remained anonymous for the entire study. In each phase of the experiment, participants first read instructions and could start the experiment only after they had correctly answered a set of questions testing their comprehension of the instructions (see Appendix A.1 screenshots and examples).

2.2.2 Design of Phase One Experiment

In phase one of the experiment, participants were asked to complete a sequence of 36 "Reading the Mind in the Eyes" (RME) test questions as well as a sequence of 5 room assignment tasks. More specifically, CSOP and RME were implemented as two distinct web apps, each of which appeared as a separate link in the MTurk iframe. The order of the links was randomized for each participant but they could choose to click on them in whatever order they wished. For the RME questions, participants were shown in each question a pair of eyes and were asked to select one of the four words that best describe the emotions shown by the eyes (See Appendix A.2 for an illustration of the test).

For the room assignment task part, we first introduced participants to the problem

[3] http://mturk.com

and each completed one practice task (as per our pre-registration, it is not included in the analysis), in which $N = 8$ students need to be assigned to $M = 5$ rooms while respecting $Q = 4$ constraints. Each participant was then given a sequence of five room assignment tasks, to be completed independently, where the maximum amount of time a participant could spend on a task was 5 minutes. Table 2.1 summarizes the main properties of the five task instances used in phase one.

Table 2.1: Main properties of the 5 room assignment tasks used in phase one of our experiment.

Task Order	N	M	Q	Max possible score	Difficulty
1	6	4	2	343	Easy
2	9	6	8	554	Hard
3	6	4	2	323	Easy
4	9	6	8	564	Hard
5	6	4	2	325	Easy

As shown in the table, we intentionally included 3 easy task instances and 2 hard task instances in the sequence. We did not randomize the order of the task instances in phase one to minimize the noise in the measurement of individual skill due to random ordering effects. We included more easy task instances than hard task instances in phase one to minimize potential self-selection in phase two of our experiments (i.e. where only participants who did well in phase one would return for phase two[4]), which turned out to be very effective (see Appendix A.5 for more details).

When working on a room assignment task, a participant was presented with a graphical interface where each student was represented as a person icon and each room was shown as a box (see Appendix A.1 for examples of the interface). The participant could then drag the icons of students and drop them to different boxes to adjust the room assignment plans. Assistive information such as the score of the current room assignment plan, the list of violated constraints, and the amount of time left in the task was also displayed and updated on the interface while the participant changed the solution. At any time during the allotted 5-minute period for a task,

[4]For example, participants who performed well in phase one may be more likely to participate the phase two experiment, implying possible self-selection biases; by having more easy task instances in phase one, most participants may feel they performed well thus bias is attenuated.

the participant could push a button to submit her solution and move on to the next task (or to the end of the room assignment task sequence), or the participant would be automatically redirected to the next task when the 5-minute timer was up. After the participant solved all five room assignment tasks in phase one, she was asked to complete an exit survey, in which we asked her to self-report the following information:

- Age
- Gender
- Highest Education Received
  - High School
  - US Bachelor's Degree
  - Master's or higher
  - Other
- Were the instructions clear?
- Was the pay fair?
- Was the time limit per task reasonable?
- Did you encounter any problems with the user interface?
- If you had assigned all students to rooms and had no conflicts, which of the following would you be most likely to do?
  - Submit your solution and move on the next task
  - Try to increase your score by moving students around as long as you didn't generate any new conflicts
  - Try to increase your score by moving students around even if it meant generating new conflicts

- If you had assigned some (but not all) students to rooms and had encountered one or more conflicts, would you:
  - Put off resolving the conflict(s) until all students had been assigned?
  - Stop assigning students to rooms until conflict(s) had been resolved?
  - Continue assigning students as long as no more than one conflict were present?
- When assigning a student to a room, did you focus more on
  - Which room had the highest score?
  - Which room(s) would avoid generating conflicts?
- Any other feedback?

At the end of phase one, we obtained a number of measurements for each participant:

- *Skill*: defined as the sum of the participant's score on the two hard room assignment tasks. We only use participant's scores on the hard tasks as hard tasks are more discriminative and scores on hard tasks have higher variability, but we note that a participant's score on the two hard tasks highly correlate with the participant's score on each of the five room assignment tasks (see Appendix A.4 for validity check).
- *Social perceptiveness* level: defined as the number of RME questions the participant correctly answered.
- Cognitive style: operationalized in four different ways:
  1. *speed (fast vs. slow)*, which is decided by whether the total amount of time the participant spent on solving the hard instances of phase one room assignment tasks is below or above the median;

2. *problem-solving style (pragmatic vs. tenacious)*, which is decided by the participant's self-reported answer for the exit-survey question "If you had assigned all students to rooms and had no conflicts, which of the following would you be most likely to do?": pragmatic (i.e., the participant chose "submit your solution and move on the next task" or "try to increase your score by moving students around as long as you did not generate any new conflicts") or tenacious (i.e., the participant chose "try to increase your score by moving students around even if it meant generating new conflicts");

3. *constraint violation tolerance (low vs. high)*, which is decided by the participant's self-reported answer for the exit-survey question "If you had assigned some (but not all) students to rooms and had encountered one or more conflicts, what would you do?": low (i.e., the participant chose "stop assigning students to rooms until conflict(s) had been resolved") or high (i.e., the participant chose "put off resolving the conflict(s) until all students had been assigned" or "continue assigning students as long as no more than one conflict were present"); and

4. *problem-solving focus (optimizer vs. satisficer)*, which is decided by the participant's self-reported answer for the exit-survey question "When assigning a student to a room, what did you focus more on?": optimizer (i.e., the participant chose "which room had the highest score") or satisficer (i.e., the participant chose "which room(s) would avoid generating conflicts").

Although our measurements of each participant's skill and social perceptiveness level are continuous, to facilitate the block randomization scheme that we would adopt in phase two of our experiment, we further used a median split to categorize each participant into the high or low class on both measurements. For example, a participant whose skill was above the median skill while social perceptiveness was below the median level would be categorized as "high skill, low social perceptiveness." We note that in our analysis we use the original (continuous) scores for individuals

that we obtained from phase one (where a team's score is the average of the team members' scores), not the block labels, to differentiate high-skilled/low-skilled (or high social perceptiveness/low social perceptiveness) teams. See Section 2.2.4 for more details.

2.2.3 Design of Phase Two Experiment

As per our pre-registration, we included the first 1200 participants who completed our phase one experiment into the second phase of our experiment. Among these 1200 participants, there were 313 "high skill, high social perceptiveness" (HH) individuals, 284 "high skill, low social perceptiveness" (HL) individuals, 249 "low skill, high social perceptiveness" (LH), and 354 "low skill, low social perceptiveness" (LL) individuals.

During a pilot study we conducted prior to our main experiment, we deployed a simple randomization scheme and had individuals of different levels of skills and social perceptiveness to form teams of three members at random in phase two. The majority of the teams formed in this way contained a mixture of high/low skill (or high/low social perceptiveness) individuals. As a result, the variance of a team's skill or social perceptiveness level (defined as the average skill or social perceptiveness level of members in that team) across different teams was limited. Practically, this implies that a large sample size would be needed to detect any statistically significant performance effect of team composition.

To address this problem, we adopted a block randomization scheme in phase two of our main experiment. Specifically, prior to the start of phase two, we created six qualifications on Amazon Mechanical Turk, with each qualification corresponded to a "block." Participants of one particular block could only find and work on the HIT corresponding to their block, but not the other five HITs. Table 2.2 provides a summary of these six blocks.

For each individual of a particular type (e.g. "high skill, low social perceptiveness" or HL), with 50% probability we assigned her to the block in which all individuals were of the same type (e.g., the "HL" block), and with 50% probability we assigned her to the block in which all individuals had the same social perceptiveness label as her, but

Table 2.2: Summary of the six blocks that we used in phase two of our experiments.

Block name	# assigned to this block	# showed up in this block
HH	155	100
MH	285	213
LH	122	90
HL	147	97
ML	310	221
LL	181	107

Table 2.3: Main properties of the 5 room assignment tasks used in phase two of our experiment. The order of tasks was randomized in the experiment.

Task ID	N	M	Q	Max possible score	Difficulty Level
1	6	4	2	340	Easy
2	8	5	5	441	Medium
3	9	6	8	672	Hard
4	12	7	12	673	Very Hard
5	18	8	18	996	Super Hard

may have different skill labels (e.g., the "ML" block, meaning "mixed skill levels, low social perceptiveness"). Within each block, we further randomly assigned participants either to the individual condition (31% of the time) or to the team condition (69% of the time). The individual condition was identical to phase one except that the five room assignment tasks were different (and generally more difficult) and that the maximum time allotted per task was ten rather than five minutes. Table 2.3 summarizes the main properties of the 5 task instances we used in our phase two experiment (the task sequence used in the individual condition is the same as that used in the team condition). In the team condition, participants worked in teams of three members from the same block.

The main effect of the block randomization scheme was to oversample statistically less frequent combinations (e.g., all team members had high skills or high social perceptiveness), which helped us to increase the statistical power of our experiments (a secondary benefit was that it allowed us to match the distributions of participant types in phases one and two; see Appendix A.5). To illustrate, the frequency of HH individuals in the population is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ hence under simple random assignment the

expected frequency of all HH teams would be $\frac{1}{4}^3 = \frac{1}{164}$. Of the 1,200 participants who were qualified for phase two, 828 participants entered the experiment and 237 of them placed in the individual condition (the data for 3 of them was incomplete; hence the effective number of individuals is 234) and 591 placed in the team condition. Of the 197 teams formed, the data for 1 team was incomplete, hence the effective number of teams is 196. In the absence of block randomization, therefore, we would expect to have $196/64 = 3$ All-HH teams. With block randomization, we guaranteed at least 22 All-HH teams (because of random assignment in the MH block it is possible that one or more additional All-HH teams would result). Put another way, to generate 22 All-HH teams with simple random assignment we would have required $22 \times 64 = 1408$ teams or over 4,000 participants just for the teams condition (6,000 in total). In summary, the number of teams formed in HH, MH, LH, HL, ML, LL blocks were 22, 55, 18, 21, 56, 24, respectively. Note that we did not block on participants' cognitive styles, as doing so would require a much larger sample size.

During the experiment, each individual (or team) first completed one practice task ($N = 9$, $M = 6$, $Q = 8$). Then, they could proceed to complete the sequence of room assignment tasks of various levels of difficulty; each task had a maximum time limit of 10 minutes (unlike phase one, which had a time limit of 5 minutes), and the task order was randomized (to account for any ordering effects). While participants in a an individual condition were presented with a set up that is identical to phase one, participants assigned team condition were presented with an interface where all team members can drag any icon of students to any room cell simultaneously as they wish (see Appendix A.1 for an example of task interface). To avoid conflicts, when one team member was moving a student icon, that particular student icon was "locked" and other team members could not move it until it was released. We provided a chatbox on the task interface, enabling team members to communicate freely with each other during the tasks. We also presented an event log on the task interface to help team members make sense of all movements that had been made within the current task. At any time during a task, each team member could indicate whether she was satisfied with the current solution using a toggle button. Once all

three members of a team indicated they were satisfied with the solution, the team would move on to the next task (or to the end of the experiment). If the team had never unanimously suggested they were satisfied with the solution, the team would automatically be redirected to the next task when the 10-minute timer was up.

At the end of phase two of the experiment, while participants in the individual condition were asked to were asked to complete an exit survey that is identical to the one in phase one, participants in team condition were asked the following:

- How would you describe your strategy in the game?
- Do you feel the pay was fair?
- How satisfied are you with the outcome of the game?
  - Extremely satisfied (1) — Extremely dissatisfied (7)
- Do you think your team worked well together?
  - Strongly agree (1) — Strongly disagree (7)
- How valuable do you think your perspective was to the end results?
  - Extremely valuable (1) — Extremely invaluable (7)
- How comfortable were you in sharing your perspective with the team through the chat?
  - Extremely comfortable (1) — Extremely uncomfortable (7)
- Feedback, including problems you encountered.

2.2.4 Details of Analysis

In this work, we are interested in comparing the effect of being in a team (i.e., teams vs individuals) as well as examining the several factors (e.g., skill level, skill diversity, social perceptiveness level, cognitive style diversity, etc.) that determine the team performance. In the case of the first question (i.e., team vs individual) or independent

variable is a binary indicator that specifies whether the observation in phase two is generated by an individual or a team. For the second question (i.e., different team compositions), we defined a number of measures as our independent variables to capture various possible influencing factors of team performance:

- *(Team-level) skill*: the average value of three team members' skills (recall each member's skill was measured in phase one experiment as the sum of scores obtained on the two hard tasks)

- *(Team-level) social perceptiveness level*: the average value of three team members' social perceptiveness level (recall each member's social perceptiveness level was measured in the phase one experiment as the number of RME questions correctly answered)

- Skill diversity: the variance of the three team members' skills

- Cognitive style diversity: Given an operationalization of cognitive style, we label the team as homogeneous or diverse on that cognitive style by checking whether the three team members in the team belong to the same type ("homogeneous") or not ("diverse").

The main dependent variable is each team's or individual's performance in the second phase of our experiment. As per our pre-registration, we measured team performance in two ways:

- *Normalized score*: the score a team obtained in a room assignment task divided by the maximum score of that task, i.e., normalized score $= \frac{\text{score on Task } T}{\text{max score for task } T}$

- *Duration*: the amount of time a team spent on solving a room assignment task

- *Efficiency* (not pre-registered): Acts as a useful summary of the two other metrics, i.e., efficiency $= \frac{\text{normalized score on task } T}{\text{duration on task } T}$

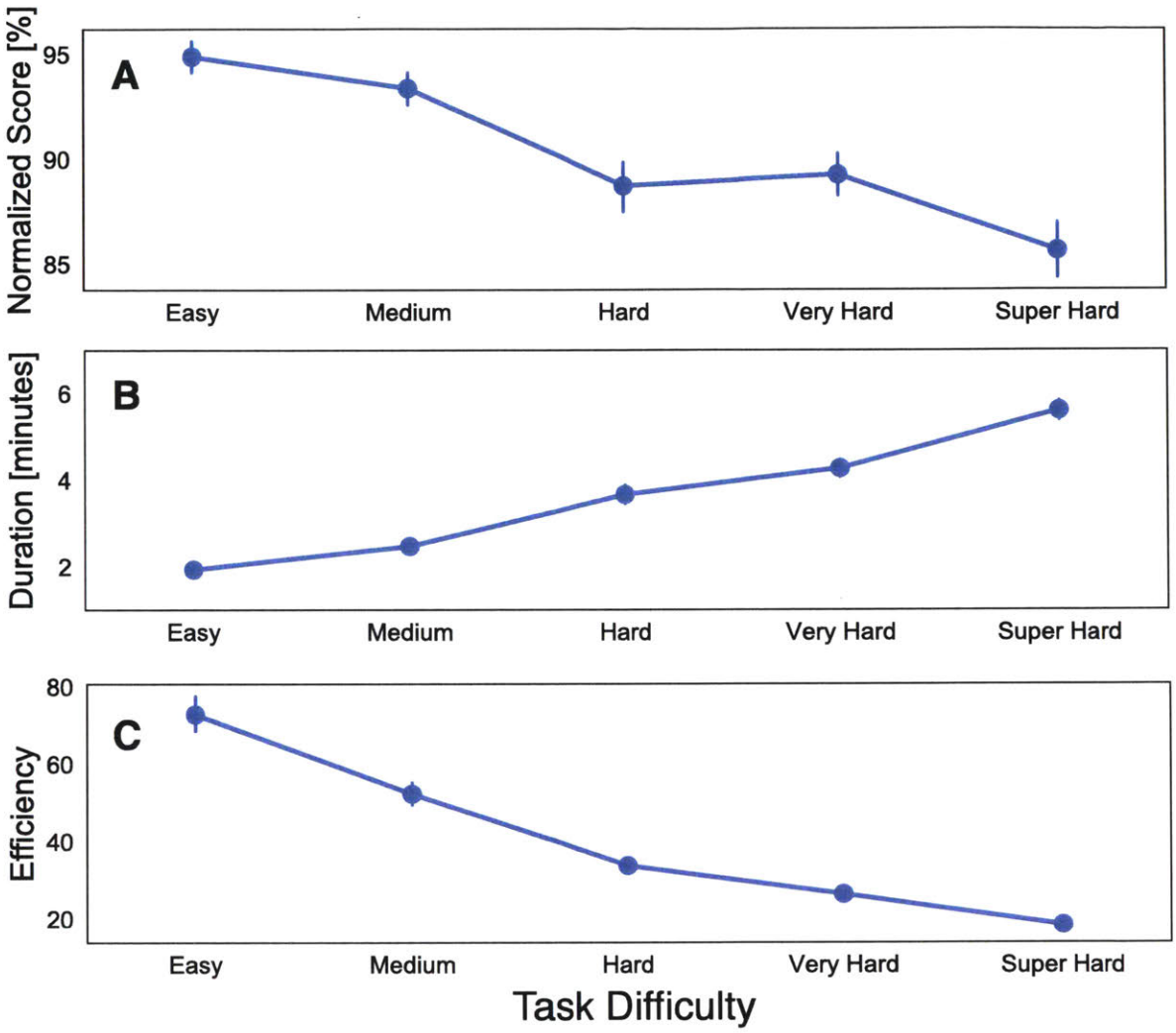


Figure 2-2: The five task difficulty levels in phase two were characterized by the different number of students to be assigned, the number of dorm rooms available, and the number of constraints. Increasing the task difficulty (i.e., environment complexity) reduces the normalized score and increases the time it takes participants to submit an assignment. Data is combined across both individual and team conditions across all 6 blocks. Error bars indicate 95% confidence intervals. The effective normalized score of a feasible solution is 80% and the minimum time required for a solution to be submitted is one minute, hence the starting points of the Y axes.

2.2.5 Results

Performance as a Function of the Environment Complexity

Fig 2-2 shows how overall performance varied as a function of task complexity, where we use two independent definitions of performance. First, we define performance as normalized score (i.e., $\frac{\text{score on Task } T}{\text{max score for task } T}$) thereby allowing us to compare performance across tasks of different complexity which may have widely varying maximum possible scores. Second, we also define performance as duration (i.e., time elapsed from the start of a task until a solution is submitted)[5]. In addition, Fig 2-2 also shows our third metric, *efficiency*, which acts as a useful summary of the two other metrics (i.e., efficiency $= \frac{\text{normalized score on task } T}{\text{duration on task } T}$).

Fig 2-2 shows that higher complexity led both individuals and teams (see also Appendix A.6) to score a lower fraction of the maximum possible score (2A) and work for longer (2B). Efficiency, therefore also decreased with task complexity (2C). Although the direction of these results is unsurprising, the large and roughly linear dependency of two separate performance measures on complexity validates our design, in which overall complexity is manipulated by varying one or more task/environment parameters (N, M, Q).

Moreover, the ability to vary human-experienced complexity by such substantial margins (on average, individuals and teams spent roughly three times as much work time on "super hard" as "easy" tasks, but obtained normalized scores that were roughly ten percentage points lower) allows us to test for interaction effects between optimal team composition and task complexity where theories of collective performance have been largely silent, i.e., to what extent does the optimal composition depend on the characteristics of the task being performed? Alternatively, one can view varying complexity as a robustness check on findings obtained for any single task [23]. In other words, systematically varying task/environment complexity is informative with respect to our main research questions regardless of whether optimal team composition depends on it.

[5]where we note that all tasks timed out at 10 mins regardless of complexity

Groups are Superior; Only when the Environment is Complex

Fig 2-3 compares overall team performance with that of "comparable individuals," which we define in two ways: first, a randomly drawn individual from the same block; and second, by constructing a "nominal team," drawing three individuals randomly and without replacement from the same block, and then choosing the individual with the highest score from phase one. Nominal teams, therefore, simulate a situation in which teams simply nominate their best performer to do all the work while the others contribute nothing. For all levels of task complexity, Fig 2-3A shows that teams score higher than randomly selected individuals but lower than nominal teams, consistent with longstanding findings that nominal teams outperform real teams under various circumstances [197].

Interestingly, however, Fig 2-3B shows that teams complete the most complex tasks—but not simpler ones—faster than either random individuals or nominal teams, suggesting that for tasks with many components (students and rooms) and many constraints the benefits of distributing work to a team outweigh the process losses (e.g., motivation loss, coordination cost) associated with groups [109]. Finally, Fig 2-3C shows that for complex tasks the gains in speed exceed the deficits in score, resulting in a striking interaction between task complexity and configuration with respect to efficiency: for easy tasks teams are considerably less efficient than either random individuals or nominal teams, yet they are considerably more efficient than either for the most complex tasks. This result is reminiscent of group decision making among social insects where a study have found that colonies outperform individuals when the discrimination task is difficult but not when it is easy [177].

Skill Accounts for 4 Times as Much as Everything Else

Fig 2-4 shows the absolute and relative effects of all pre-registered independent variables on collective performance, which is quantified as score (Fig 2-4A), duration of completion (Fig 2-4B), and efficiency (Fig 2-4C) respectively (and all three metrics are standardized within each task complexity level as per our pre-registration). Across

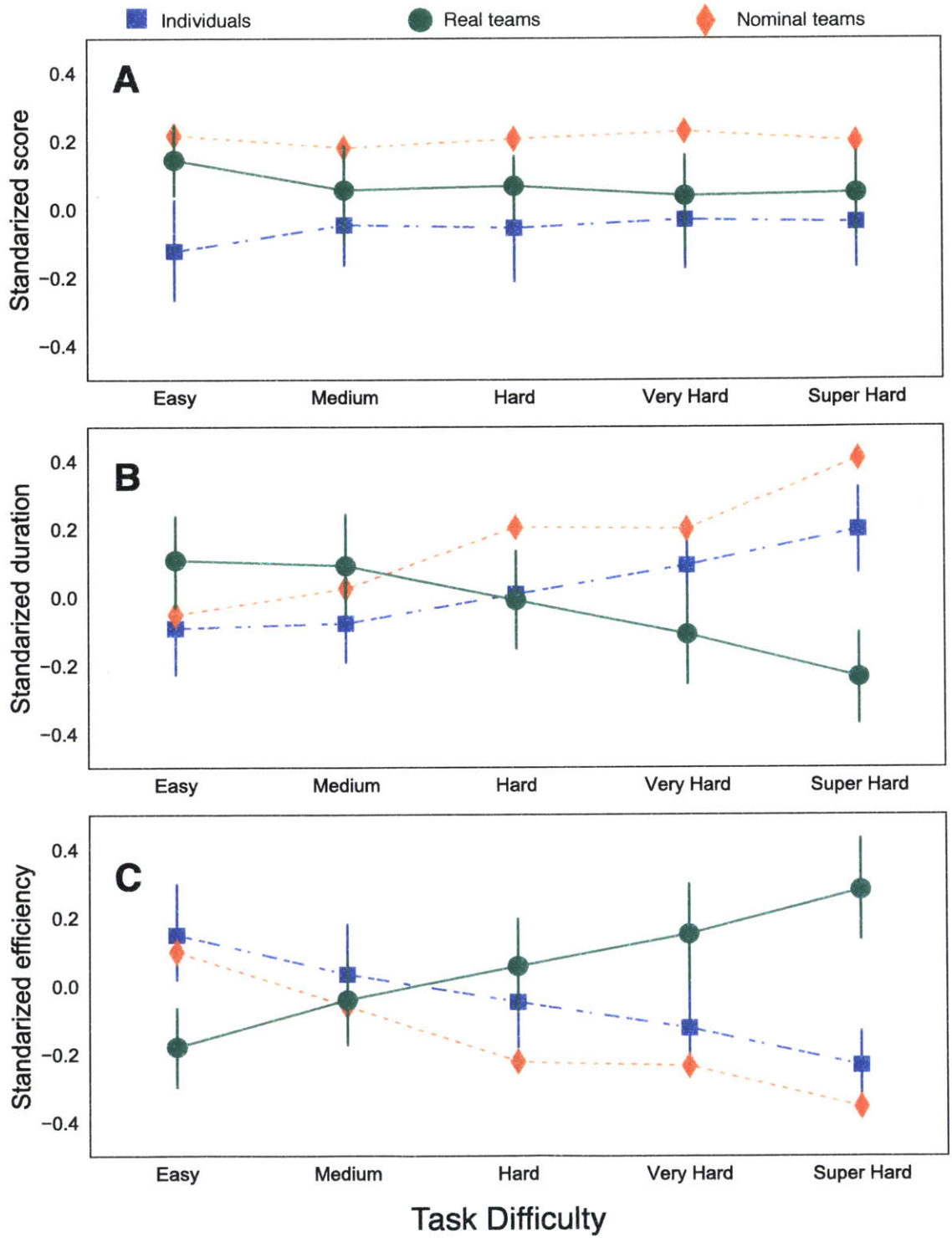


Figure 2-3: Comparing performance across individuals, real teams, and nominal teams. Individual, real team, or nominal team data is combined across all 6 blocks and standardized within each task complexity level. Error bars indicate 95% confidence intervals.

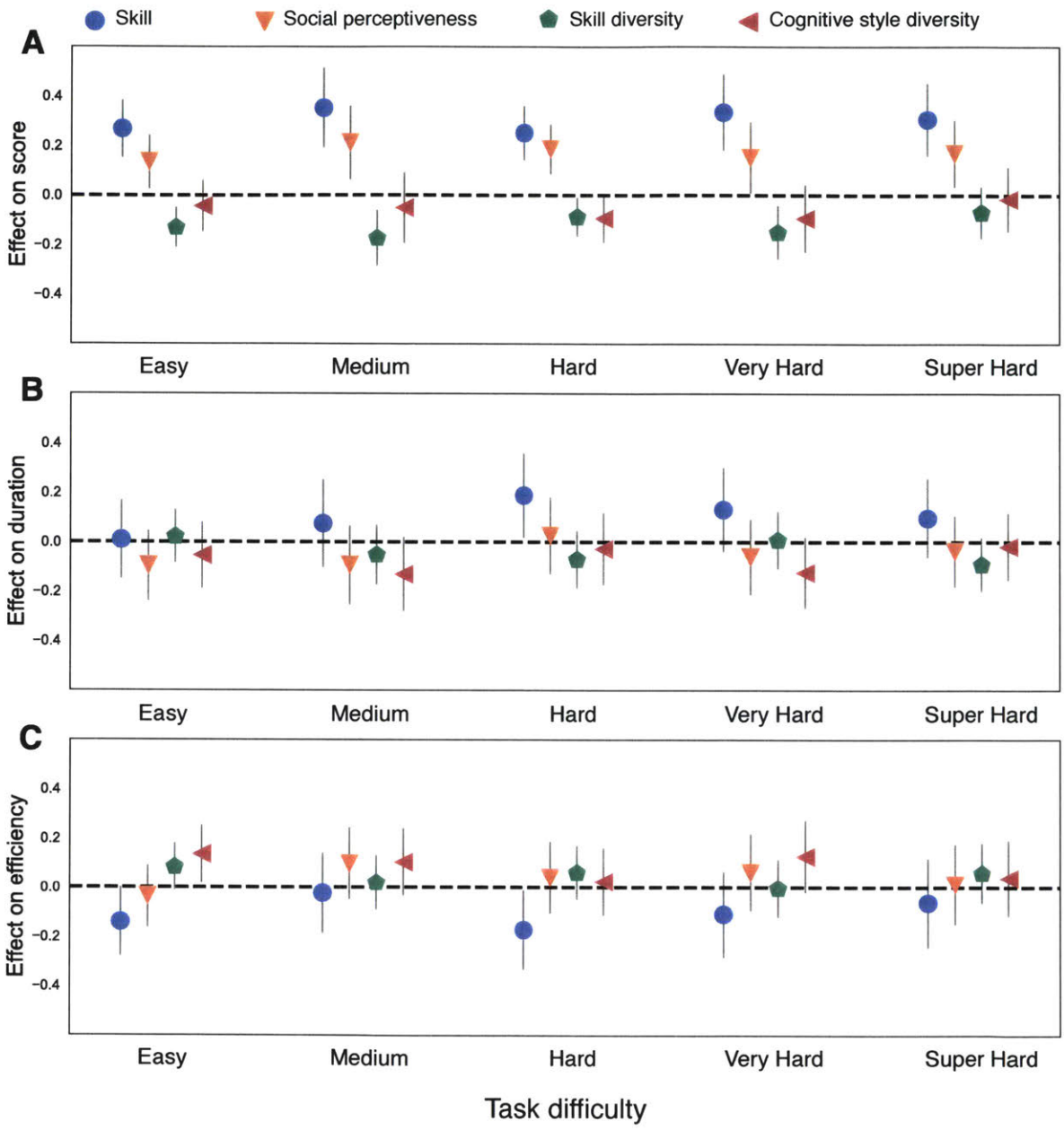


Figure 2-4: Team composition and team performance. The effects of cognitive style diversity shown in the figure are for participants' cognitive styles in solving the room assignment task as defined by "optimizer" vs. "satisfier." Error bars indicate 95% confidence intervals. See Appendix A.7 for additional analyses on the effects of skill/cognitive style diversity.

Table 2.4: Relation between team's average skill level and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's average skill level. All models included random effects for teams as intercept to account for dependence across tasks (i.e., random effects are clustered on each team, using team id as the identifier). Increasing a team's average skill significantly increases the team's score in solving CSOPs, but has no effect on duration or efficiency.

	Score	Duration	Efficiency
(Intercept)	0.081*	−0.028	0.050
	(0.035)	(0.056)	(0.051)
Skill level	0.303***	0.102	−0.100
	(0.041)	(0.066)	(0.061)
N	980	980	980
$team_{id}$	196	196	196

Significance: $*** \equiv p < 0.001$; $** \equiv p < 0.01$; $* \equiv p < 0.05$

all complexity levels Fig 2-4A shows that average skill had the largest effect on teams' scores, and was both positive and highly significant (Table 2.4). In addition, the effect of skill is consistently and significantly larger than that of social perceptiveness (Wald chi-square test; $\chi^2 = 6.35$, $P = 0.012$; see Appendix A.7 for additional "relative importance" analysis), which was also positive and significant (Table 2.5). In contrast, skill diversity (i.e., variance in team members' ability) has consistently and significantly negative effects on the score (see Table 2.6) while no measure of cognitive style diversity has any consistent and significant effect (see Table2.7 and Appendix A.7). Compared with team score, the effects of skill, social perceptiveness, and diversity on duration (Fig 2-4B) and efficiency (Fig 2-4C) are small and not significant at the $p < 0.05$ level.

Effect sizes are important for testing theories, but in practice, it is also important to consider predictive accuracy [96, 210, 220]. To illustrate, recall our hypothetical manager who wishes to compose a team for some task, and who has prior information about the skill, cognitive style, and social perceptiveness of prospective team members. In essence, the manager's task is to predict which combination of traits will yield the best collective performance. More specifically, the manager cares about two

Table 2.5: Relation between team's average social perceptiveness and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's average skill level. All models included random effects for teams as intercept to account for dependence across tasks (i.e., random effects are clustered on each team, using team id as the identifier). Increasing a team's average skill significantly increases the team's score in solving CSOPs, but has no effect on duration or efficiency.

	Score	Duration	Efficiency
(Intercept)	0.068	−0.030	0.053
	(0.038)	(0.056)	(0.051)
Social perceptiveness	0.171***	−0.051	0.036
	(0.040)	(0.060)	(0.055)
N	980	980	980
$team_{id}$	196	196	196

Significance: $*** \equiv p < 0.001$; $** \equiv p < 0.01$; $* \equiv p < 0.05$

Table 2.6: Relation between team's skill diversity and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's average skill level. All models included random effects for teams as intercept to account for dependence across tasks (i.e., random effects are clustered on each team, using team id as the identifier). Increasing a team's average skill significantly increases the team's score in solving CSOPs, but has no effect on duration or efficiency.

	Score	Duration	Efficiency
(Intercept)	0.082	−0.010	0.026
	(0.203)	(0.062)	(0.057)
Skill diversity	−0.072*	−0.035	0.046
	(0.030)	(0.045)	(0.041)
N	980	980	980
$team_{id}$	196	196	196

Significance: $*** \equiv p < 0.001$; $** \equiv p < 0.01$; $* \equiv p < 0.05$

Table 2.7: Relation between team's cognitive style diversity and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's average skill level. All models included random effects for teams as intercept to account for dependence across tasks (i.e., random effects are clustered on each team, using team id as the identifier). Increasing a team's average skill significantly increases the team's score in solving CSOPs, but has no effect on duration or efficiency.

	Score	Duration	Efficiency
(Intercept)	0.070	−0.031	0.053
	(0.039)	(0.056)	(0.051)
Cognitive style divesrity	−0.060	−0.070	0.087
	(0.039)	(0.056)	(0.051)
N	980	980	980
$team_{id}$	196	196	196

Significance: $*** \equiv p < 0.001$; $** \equiv p < 0.01$; $* \equiv p < 0.05$

related questions. First, what is the predictive accuracy of his or her "model" (i.e., how much 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 team performance depends on the observed individual traits (versus unobserved traits, factors external to the individuals, and random noise), and hence to what extent it can be "engineered" at all. The answer to the second question indicates which of the observed variables to prioritize, and how much, when selecting team members. The latter is particularly important when there is a cost associated with the measurement of the relevant variables.

Addressing the first question, Fig. 2-5A shows the out-of-sample R^2 for a simple linear regression model where the dependent variable is the total normalized score (i.e. summed over all tasks), and all observed independent variables are included first independently (i.e., separate, univariate regressions; green symbols) and then cumulatively (purple symbols) in order of increasing independent explanatory power (i.e., the R^2 of the corresponding univariate regression). Overall, the R^2 was approximately 0.24, meaning that the model "explained" about 24% of the observed variance in held-

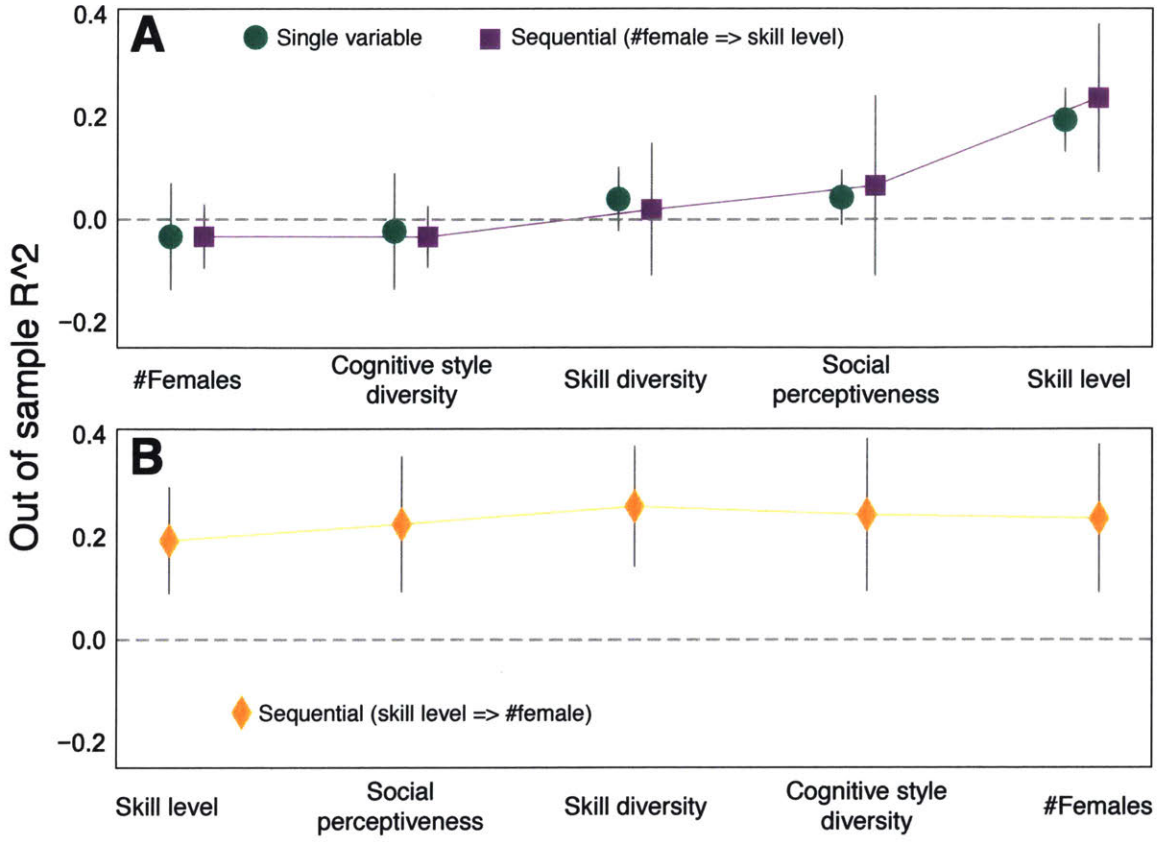


Figure 2-5: Using linear regression (70% training, and 30% testing; randomized and repeated 5 times) to predict team's normalized score with team's skill level, skill diversity, social perceptiveness, cognitive style diversity, and the number of female team members. (A) Compares predictive performance for covariates regressed independently (i.e. in separate models; green symbols), and in a single model where covariates are added in order of increasing independent predictive performance (purple symbols). (B) Predictive performance for a single regression model where covariates are added in order of decreasing independent predictive performance. Error bars indicate 95%confidence intervals..

out data (more complex machine learning models scored similarly, see Appendix A.8). The obtained overall $R^2 \approx 0.24$ is a figure that is intermediate between recent attempts to predict individual life-course outcomes [167] (i.e., $0.03 \leq R^2 \leq 0.23$) in the Fragile Families predictive challenge [128, 111, 121] and attempts to predict the size of Twitter cascades ($R^2 = 0.4$) [135].

Addressing the second question, Fig. 2-5B shows cumulative R^2 for the same model but starting with the most explanatory variable (i.e., skill level) and adding variables in order of decreasing explanatory power. Although social perceptiveness and skill

diversity do visibly increase out-of-sample R^2, these improvements are even smaller than one would surmise from the corresponding regression coefficients in Fig. 2-4A: skill alone corresponds to an R^2 of 0.19, or 80% of all explained variance. In other words, our hypothetical manager could predict her team's performance almost as well knowing only skill as she could with variables together. In contrast, predicting duration is a much harder task: almost no variance can be explained either by skill or by any combination of measured attributes (see Appendix A.8).

2.3 Discussion and Chapter Reflections

These results provide mixed support for previous studies and also help to clarify some inconsistencies between them. First, our results help to reconcile conflicting prior findings regarding the effectiveness of teams vs. individuals: whereas we find that teams clearly outperform comparable individuals selected at random, consistent with [219], we also find that teams score worse than the best individual selected from a nominal team of the same size, consistent with [109, 197]. Interestingly, even as teams underperform nominal groups in terms of score, for the most complex tasks—but not for simpler tasks—they attain higher efficiency by completing their work faster.

Second, our finding that the effects of average individual skill and social perceptiveness are positive and highly significant is consistent both with the aforementioned meta-analytical studies that favored ability [60, 187, 27], and also with the more recent experiments that emphasized social perceptiveness work [217, 120, 69, 110]. However, our ability to compare effect sizes and predictive performance across multiple effects resolves the apparent inconsistency between the two sets of results: skill dominates social perceptiveness by an order of magnitude.

Third, our findings of that skill diversity is negatively associated with team performance is consistent with [18] but directly contradicts [98]. Even if the latter claim is interpreted as implicating cognitive diversity more generally rather than skill per se, we find no evidence that any of several measures of skill or cognitive style diversity is positively associated with performance. Naturally, teams can be diverse with

respect to attributes other than skill and cognitive style (e.g., demographics, political ideology, worldview etc.) and diversity can affect outcomes other than performance on task (e.g., satisfaction, legitimacy, social equity, etc.); thus our results should not be construed as finding any effect of diversity writ large. Nevertheless, they do reinforce recent research [53, 66] which also concludes that unambiguously positive effects of diversity are more difficult to detect in carefully controlled empirical studies than what would be expected from theory [98].

Indeed, team composition and team performance are multifarious constructs each of which can be operationalized in many ways; moreover, the relationship between the two may be contingent on numerous other mediating variables related to the nature of the task [186, 138] and the environment [123, 35]. Finally, the literature on team performance comprises a mixture of simulation, observational, and experimental studies; thus it is hardly surprising that it exhibits inconsistencies. In this paper we have introduced an approach to studying team performance that leverages a unique combination of (a) class of tasks with variable environmental complexity (i.e., the complexity parameters, N, M, and Q) to increase the robustness of our results and allows us to test for interaction effects; (b) two-phase design which allows us to measure individual on-task skill and cognitive style as well as social perceptiveness prior to team assignment; (c) large sample size and block randomization to increase power; and (d) a pre-registered analysis plan to constrain researcher degrees of freedom [180].

In conclusion, our results show that on-task skill of team members far outweighs other factors, such as skill diversity, cognitive style diversity, and social perceptiveness, that have been emphasized in recent years, accounting for roughly three-quarters of explained variance. Although this result is robust to task complexity, which we varied widely, a major limitation is that we only studied one type of task. We, therefore, hope that future work will apply a similar approach to qualitatively different tasks as well as varying other parameters of interest (e.g., team size, communication patterns, division of labor, leadership, etc.). Naturally a research program that explores many parameters while still running large-N samples is logistically challenging; however, we propose that "virtual lab" experiments of the sort that we have described

here, in combination with emerging "open science" practices such as pre-registration, open data and code, replication, and "many-labs" style collaborations [112], offer a promising route forward. Finally, our emphasis on predictive accuracy seeks to move studies of team performance away from tests of theoretical conjectures (e.g., "does X correlate with performance?") and toward tests of practical significance (e.g., "how much observed variance can be explained an in terms of what?").

Chapter 3

Non-Stationary Information Environments

Social networks continuously change as people create new ties and break existing ones. It is widely noted that our social embedding exerts strong influence on what information we receive, and how we form beliefs and make decisions. However, most studies overlook the dynamic nature of social networks, and its role in fostering adaptive collective intelligence. It remains unknown (1) how network structures adapt in non-stationary environments, and (2) whether this adaptation promotes the accuracy of individual and collective decisions. Here, we answer these questions through a series of behavioral experiments and supporting simulations. Our results reveal that social network plasticity (i.e., dynamic networks) when provided with feedback can adapt to biased and non-stationary information environments. Moreover, we show that groups in dynamic networks when provided with feedback can significantly outperform their best-performing member, and that even the best member's judgment substantially benefits from group engagement. Thereby, our findings substantiate the role of social network plasticity and feedback as adaptive mechanisms for refining individual and collective judgments.

3.1 Adaptive Systems and Environmental Conditions

Adaptive systems, both natural and artificial, rely on feedback, empirical learning, and reorganization [213, 201]. Such systems are widespread, and can often be viewed as networks of interacting entities that dynamically evolve over time. Cell reproduction, for example, relies on protein networks to combine sensory inputs into gene expression choices adapted to environmental conditions [71]. Neurons in the brain dynamically rewire in response to environmental cues to enable human learning [83]. Eusocial insects modify their interaction structures in the face of environmental hazards as a strategy for collective resilience [191]. Human social network plasticity and feedback have been shown to promote human cooperation [161, 77], and culture transmission networks over generations enabled human groups to develop technologies above any individual's capabilities [89, 147]. In the artificial realm, prominent machine learning algorithms rely on similar logic, where dynamically updated networks guided by feedback integrate input signals into useful output [30, 146]. Across the board, the combination of environmental feedback (e.g., survival, payoff, reputation etc) and network dynamics represent a widespread strategy for collective adaptability in the face of environmental changes; providing groups with an effective and easy-to-implement mechanism of response to external and internal disturbance [97, 119, 191].

In our view, the information processing capabilities of interacting human groups are no exception. People's behavior, opinion formation, and decision-making are deeply rooted in cumulative bodies of social information [20], accessed through social networks formed by choices of whom we friend [11, 206], follow [195], call [65, 155], imitate [222, 179], trust [46, 205], and cooperate with [207, 161, 75]. Moreover, peer choices are frequently revised, most often based on notions of environmental cues and feedback such as: success and reliability, or proxies such as reputation, popularity/prestige, and socio-demographics [142, 107, 214, 90, 77].

3.2 Conventional Wisdom on the Wisdom of Crowds

It is widely noted, however, that social influence strongly correlates individuals' judgment in estimation tasks [144, 127, 26, 82], compromising the first of two assumptions underlying common statistical accounts of 'wisdom-of-crowds' phenomena [194]: namely, that (i) individual estimate are uncorrelated, or negatively correlated, and (ii) individuals are correct in mean expectation [78, 82].

In recent years, numerous studies have offered conflicting findings, showing that social interaction can either significantly benefit group and individual estimates [142, 18, 26, 148], or, conversely, lead them astray by inducing social bias, herding, and group-think [144, 82, 127]. There are some notable efforts that focused on providing a partial resolution to the conflict between the 'wisdom' and 'madness' of interactive crowds and found that these divergent effects are moderated by whether well-informed individuals are placed in prominent positions in the network structure [82, 26], how self-confident they are [18, 114, 129, 106], ability to identifying experts [142, 36], dispersion of skills [142, 16, 130, 29] and quality of information [103], diversity of judgments among group members [49, 29], and social learning strategies being deployed [24, 199] as well as the complexity/difficulty of the task being performed [199, 130]. In other words, what is advantageous for the group *depends on the environment* in which the group is situated in. Because people often do not know their environment (or the environment is non-stationary) it is advantageous to find easy-to-implement mechanism that perform well across shifting environments.

3.3 The Role of the Environment, Again

Notably, both theoretical and experimental work on collective intelligence (including the reconciliation effort mentioned above) has been predominantly limited to frameworks where the communication network structure is exogenous, where agents are randomly placed in static social structures —dyads [18, 114], fully-connected groups [142, 127, 217, 148], or networks [82, 26].

Unlike what is explicitly or implicitly assumed in most existing work, the social networks we live in are not random or imposed by external forces, but emerge shaped by endogenous social processes and gradual evolution within a particular environmental conditions. The present study builds on the observation that agent characteristics, such as skill and information access, are not randomly located in network structure. Intuitively, groups can benefit from awarding centrality to and amplifying the influence of well-informed individuals. Therefore, the distribution of agents is often the outcome of social heuristics that form and break ties influenced by social and environmental cues [107, 214, 32, 90], and therefore, the emergent structure cannot be decoupled from the environment.

Here, we hypothesize that dynamic social influence networks guided by feedback may be central to human collective intelligence, acting as core mechanisms by which crowds, which may not initially be wise, evolve into wisdom, adapting to biased and potentially non-stationary information environments.

3.4 Experiment: Guess the Correlation Game

3.4.1 Experimental Design

To test these hypotheses, we developed two web-based experiments (i.e., $S1$ and $S2$) that allow us to identify the role of dynamic networks and feedback in fostering an adaptive 'wisdom of crowds.' In both studies, Participants ($N_{S1} = 719; N_{S2} = 480$) from Amazon Mechanical Turk engaged in a sequence of 20 estimation tasks. Each task consisted of estimating the correlation of a scatter plot, and monetary prizes were awarded relative to performance. Participants were randomly allocated to groups of 12, and each group was randomized to one of three treatment conditions in $S1$ or four treatment conditions in $S2$. In study 1, the feedback level is fixed (i.e., full feedback) and network plasticity is manipulated (i.e., static network versus dynamic network). In study 2, plasticity is fixed (i.e., always dynamic network) and feedback is manipulated (i.e., no feedback, self feedback, and full feedback). Fig. 3-1 illustrates

the overall experimental design.

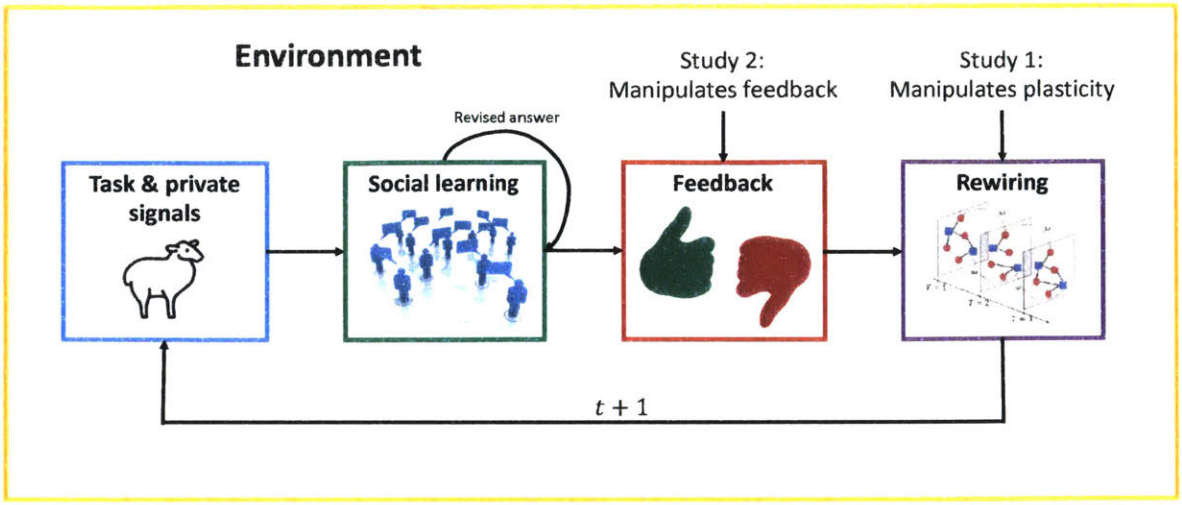


Figure 3-1: An illustration of the overall experimental design.In study 1, the feedback level is fixed (i.e., full feedback) and network plasticity is manipulated (i.e., static network versus dynamic network). In study 2, plasticity is fixed (i.e., always dynamic network) and feedback is manipulated (i.e., no feedback, self feedback, and full feedback).

Study 1: Manipulates network plasticity; full feedback

In $S1$, each group was randomized to one of three treatment conditions:

- *solo* condition: each participant solved the sequence of tasks in isolation (i.e., no social information). This condition corresponds to the traditional 'wisdom of the crowds' context [78, 194, 160]. See Figure 3-2A.

- *static* condition: participants were randomly placed in static communication networks. That means, participants will engage in a stage of active social learning, where they are exposed to their ego-network's estimates in real time. See Figure 3-2B. This context is analogous to that studied by work at the intersection of the 'wisdom of crowds' and social learning, such as [127, 129, 82, 55, 136].

- *dynamic* condition: participants at each round were allowed to select up to three peers to to follow (i.e., get the ability to communicate with) in subsequent rounds. See Figure 3-2C. This condition is novel to the work of this dissertation.

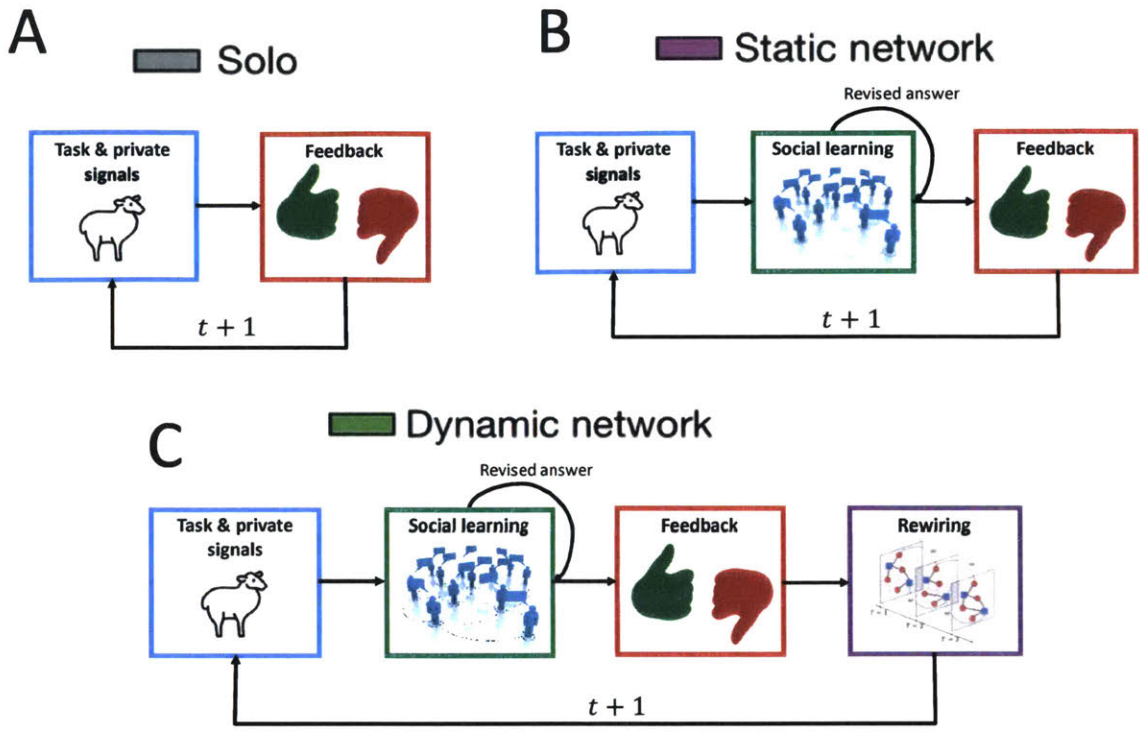


Figure 3-2: Illustration of the experimental conditions in study 1. Panel (A) depicts the Solo condition (i.e., no social information) where participants make independent estimates. This condition corresponds to the baseline wisdom of the crowd context. Panel (B) describes the Static network condition (i.e., social learning) where participants engage in a stage of interactive social learning, where they are exposed to the estimates of a fixed set of peers in real time. Panel (C) describes the Dynamic network (i.e., selective social learning) condition that adds the possibility for participants to choose who to follow and be influenced by in the next round.

Note that in the *social learning* stage (i.e., in the static and dynamic conditions; see Figure 3-2) participants observe in real-time the estimates of the other participants that they are connected to and can update their estimates multiple times before they submit their final estimate. It is up to the participant to decide how to update their guess to accommodate the information and experiences, the opinions and judgments, the stubbornness and confidence, of the other players.

After submitting a final guess, participants in all conditions were given performance feedback. That included how much they earned, what was the correct correlation, what was their guess.

Study 2: Manipulates feedback; dynamic network

In $S2$, each group was randomized to one of four treatment conditions:

- *solo* condition, where each individual solved the sequence of tasks in isolation.
- *no feedback* condition, in which participants were not shown performance feedback.
- *self feedback* condition, in which participants were shown their own performance feedback.
- *full feedback* condition, in which participants were shown scores of all participants (including their own)

Participants in all conditions (except solo, our baseline) were allowed to revise which peers to follow in subsequent rounds (i.e., similar to the 'dynamic network' condition in study 1). To further assist with reproducibility of our study, we pre-registered our $S2$ main research questions and analysis plan (AsPredicted.org #16474), and made all data and code available at OSF.io.

Estimation Task: Guess the correlation game

Participants were prompted to estimate the correlation from a scatter plot and were awarded a monetary prize based on the accuracy of their final estimate. We call this task, 'Guess the Correlation Game' [151].

This estimation task is designed to expose the mechanisms that allow intelligent systems to adapt to changes in their information environment. We can influence the performance level of participants by implementing three difficulty levels (i.e., varying the number of points, and adding outliers or non-linearities): easy, medium, and hard. see Figure 3-3.

At every round, all plots seen by participants shared an identical true correlation, but difficulty levels could differ among them [143]. The allowed the simulation of a shock to the distribution of information among participants. Specifically, each

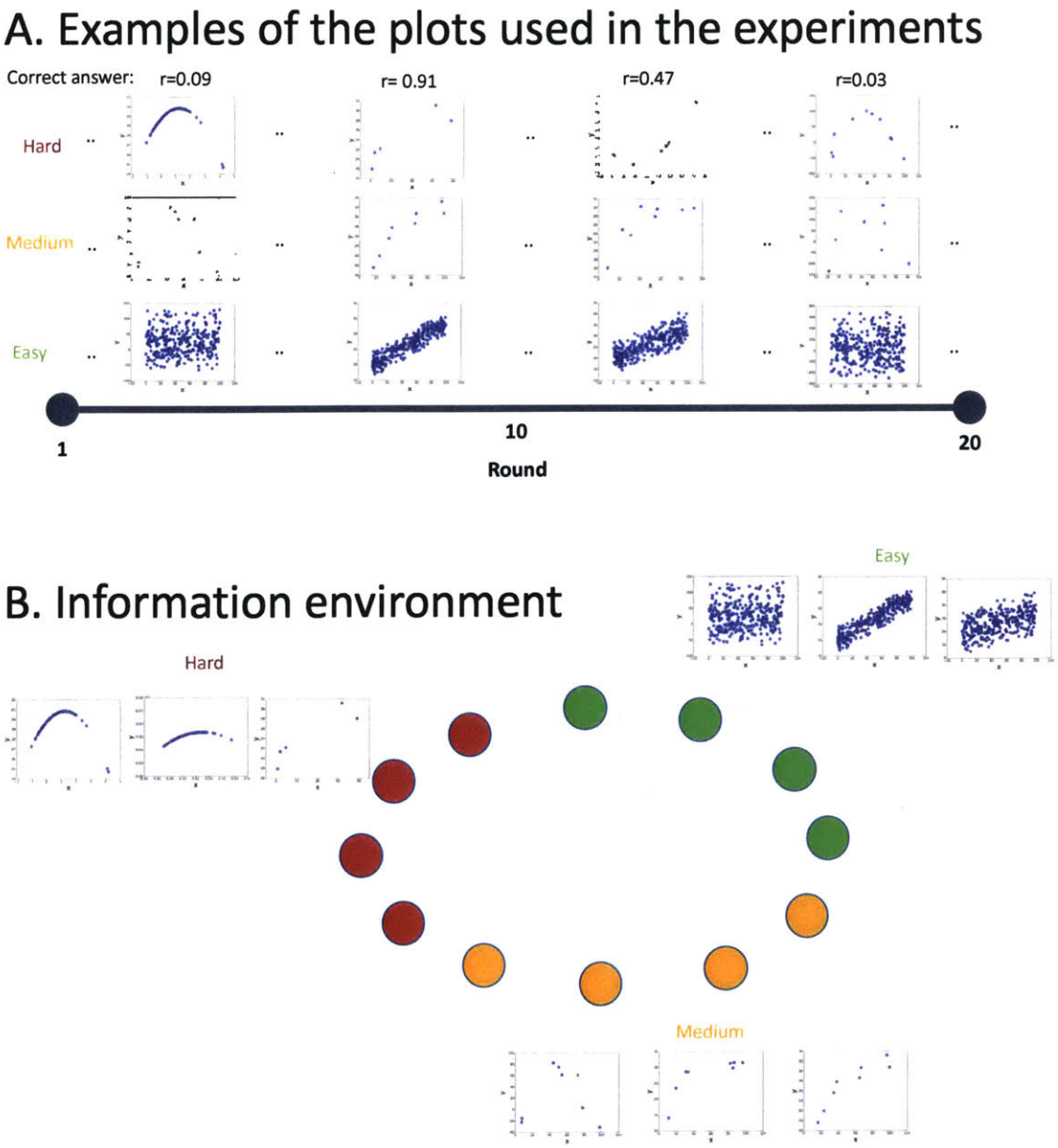


Figure 3-3: Guess the Correlation Game. An illustrative examples of the scatter plots used in the experiment is shown in Panel (A). Task difficulty, therefore, could be varied systematically at the individual level by varying the number of points, linearity, and the existence of outliers. For any given round, all participants saw plots that shared an identical true correlation, but difficulty levels could differ among them as shown in Panel (B). Participants were not informed about the difficulty level they or other participants were facing.

participant experienced a constant difficulty level across the first ten rounds; then, at round eleven, we introduced shocks by reshuffling difficulties to new levels that remained constant thereafter (see Figure 3-4). Participants were not informed about

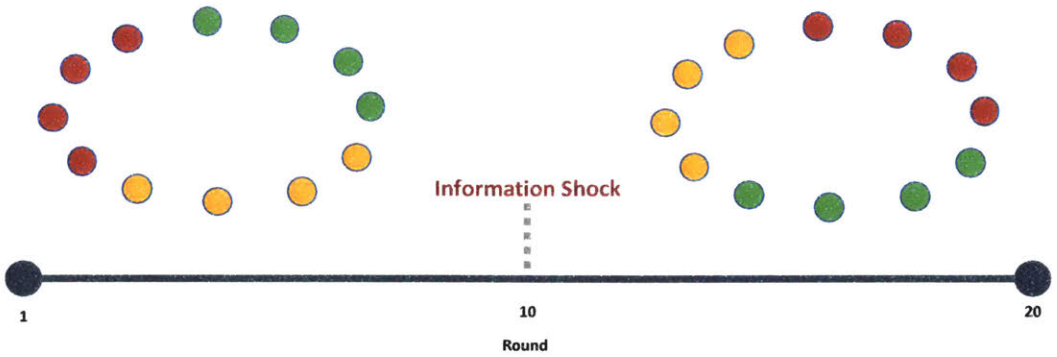


Figure 3-4: Shock to the Information Environment. We provide a change in the environment after round 10 by changing the difficulty levels for the participants for the remainder of the experiment and thereby we simulate non-stationary distributions of information among participants.

the difficulty levels they or their peers faced. See Appendix B.1.

Participant Recruitment

Amazon's Mechanical Turk is an online labor market with a large and diverse pools of people ready to promptly perform tasks for pay (called human intelligence tasks, or HITs). Typical tasks include image labeling, sentiment analysis, or classification of URLs. Additionally, Mechanical Turk is increasingly becoming a popular tool for behavioral scientists as well. Studies from across the social sciences have systematically replicated classic results from psychology and economics with data obtained from such online labor markets and deemed online experiments to be as reliable as that obtained via traditional methods [17, 45, 99, 28, 9]. Accordingly, we posted each of our experimental sessions as an external HIT (a URL of our web application, is displayed in a frame in the Worker's web browser).

All participants were recruited on MTurk by posting a HIT for the experiment, entitled "Guess the correlation and win up to $10", a neutral title that was accurate without disclosing the purpose of the experiment. The study (Approval#: 1509172301) was reviewed and approved by the Committee on the Use of Humans as Experimental participants (COUHES) at MIT. All participants provided an explicit consent to participants in this study and COUHES approved the consent procedure. All data collected in the experiment could be associated only with participant's Amazon

Worker ID on MTurk, not with any personally-identifiable information. All players remain anonymous for the entire study. At the beginning of a session, participants read on-screen instructions for the condition they are randomly assigned to their conditions. Participants could start the experiment only once they have completed a set of comprehension questions.

3.4.2 Experimental Results

Individual and Collective Outcomes

We first compared individual- and group-level errors across conditions. Evolutionary reasoning suggests that people's propensity to imitate follows from its direct benefits to the individual, but it may, nonetheless, induce benefits to the population as a whole [32]. Our first result is that networked collectives across studies significantly outperformed equally sized groups of independent participants, which is consistent with prior work on search [136, 57] as well as estimation tasks [130]. Fig. 3-5 shows the individual and group error rates—using the arithmetic mean as group estimate—normalized with respect to baseline errors in the solo condition. Overall, we find that participants in dynamic networks with feedback achieved the lowest error rates. The performance edge was larger in periods where networks had adapted to their information environment (rounds $[6, 10] \cup [16, 20]$, the 'adapted periods').

In particular, in $S1$ dynamic networks averaged 33% lower individual error ($P < 10^{-5}$), and 34% lower group error, compared to participants in static networks ($P < 10^{-4}$). In the adapted periods, dynamic networks reduced error by 47% ($P < 10^{-10}$) compared to groups that lacked plasticity (i.e., connected by static networks).

In $S2$, participants with full feedback averaged 47% lower individual error ($P < 10^{-10}$), and 54% lower group error, compared to participants in the no-feedback condition ($P < 10^{-4}$). Additionally, participants with full feedback averaged 42% lower individual error ($P < 10^{-4}$), and 42% lower group error, compared to participants in the self-feedback condition ($P < 10^{-3}$). Overall, the differences between the self-feedback and no-feedback conditions are not significant. However, in the adapted

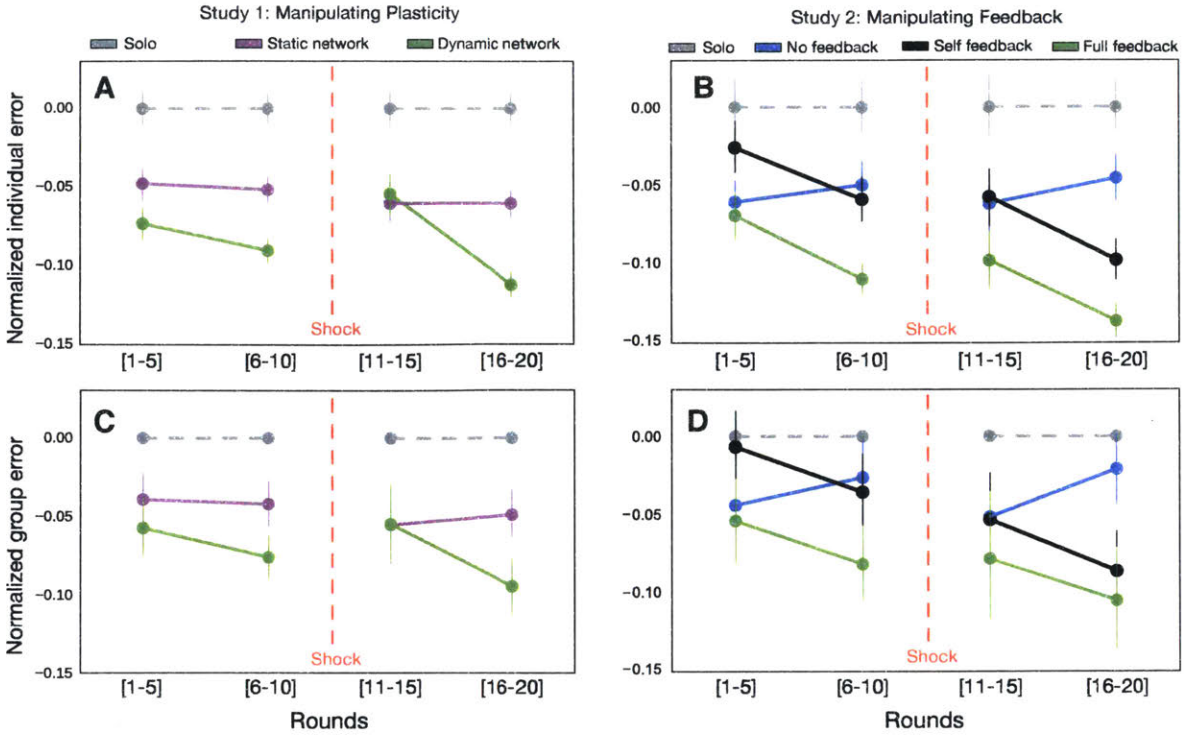


Figure 3-5: Individual and collective outcomes. Groups connected by dynamic influence networks and provided with feedback incur substantially lower individual errors as shown in Panels (A) & (B); and collective errors in Panels (C) & (D). The reduction is notably larger and more significant in periods where networks had adapted to the information environment (i.e., rounds $[6, 10]$ and $[16, 20]$). Errors are normalized with respect to average errors in the *solo* condition within each study. Error bars indicate 95% confidence intervals.

periods, participants in the self-feedback condition achieved 60% lower group error than the no-feedback condition ($P < 10^{-3}$).

Hence, these results from both studies support our primary hypothesis that adaptiveness through feedback and network plasticity can benefit both individual and collective judgment. Additional analyses on individual and group level errors are presented in Appendix B.2.

Adaptive Mechanisms

Two social mechanisms underlie the favorable performance of networked groups. First, dynamic networks adaptively centralized over high-performing individuals. This behavior was predicted by abundant evidence from cognitive science and evolutionary

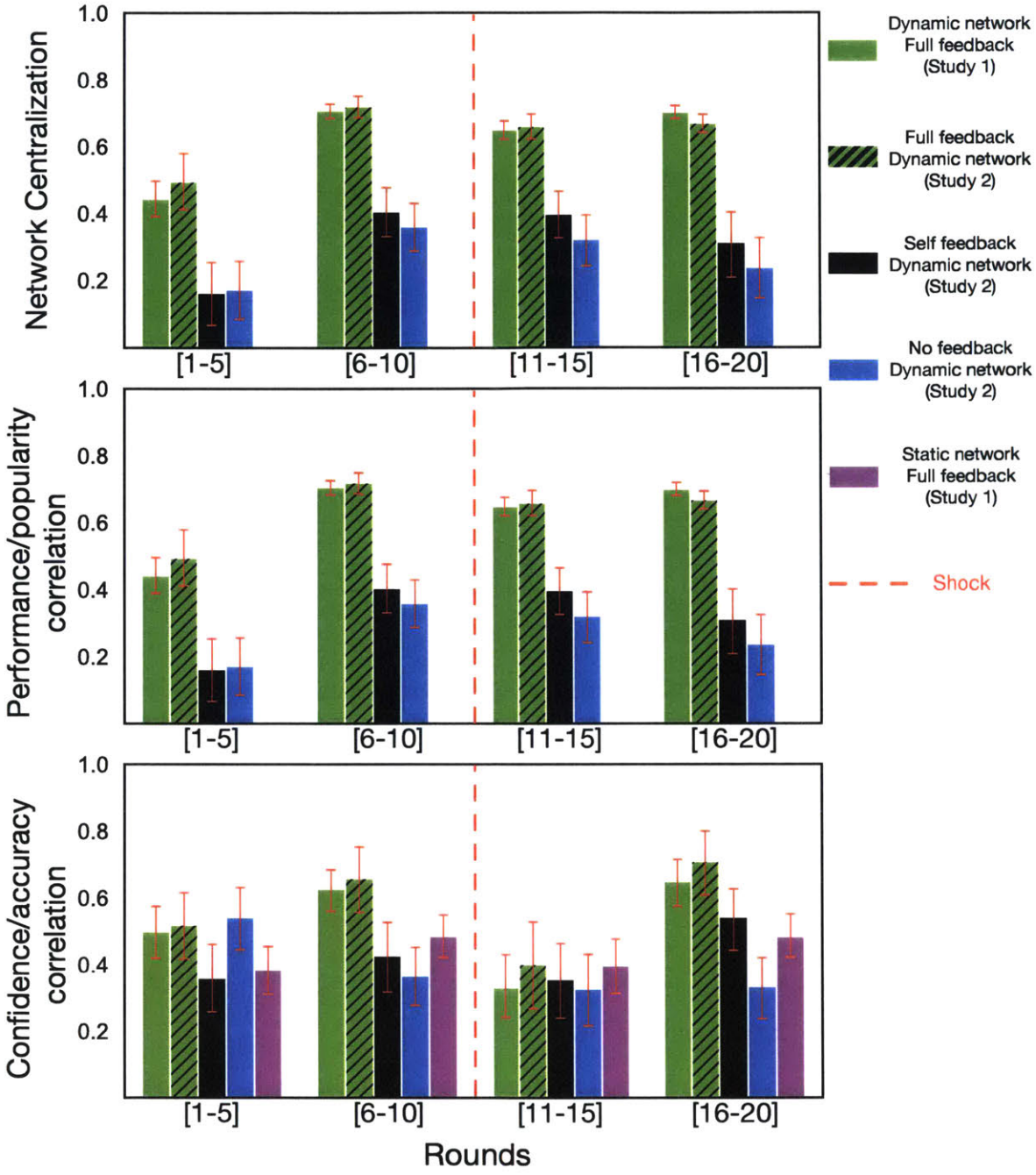


Figure 3-6: Mechanisms promoting collective intelligence in dynamic networks. Panel (A) shows that the network becomes more centralized with time (Freeman global centralization—i.e., how far the network is from a star network). Panel (B) depicts the relation between performance (i.e., average error) and popularity (i.e., number of followers). Panel (c) shows the relationship between accuracy of initial estimate and confidence (i.e., resistance to social influence). Error bars indicate 95% confidence intervals.

anthropology, which indicate that people naturally engage in selective social learning [32, 214, 89]—i.e., the use of social cues related to peer competence and reliability to choose whom we pay attention to and learn from selectively. Figs. 3-6A and 3-6B show that participants in dynamic networks consistently used peers' past performance information as success cues to guide their peer choices. As rounds elapsed, performance information accrued, and social networks evolved from fully distributed into networks that amplified the influence of well-informed individuals. Upon receiving an information shock, the networks slightly decentralized, entering a transient exploration stage before finding a configuration adapted to the new distribution of information among participants (see Fig. 3-7).

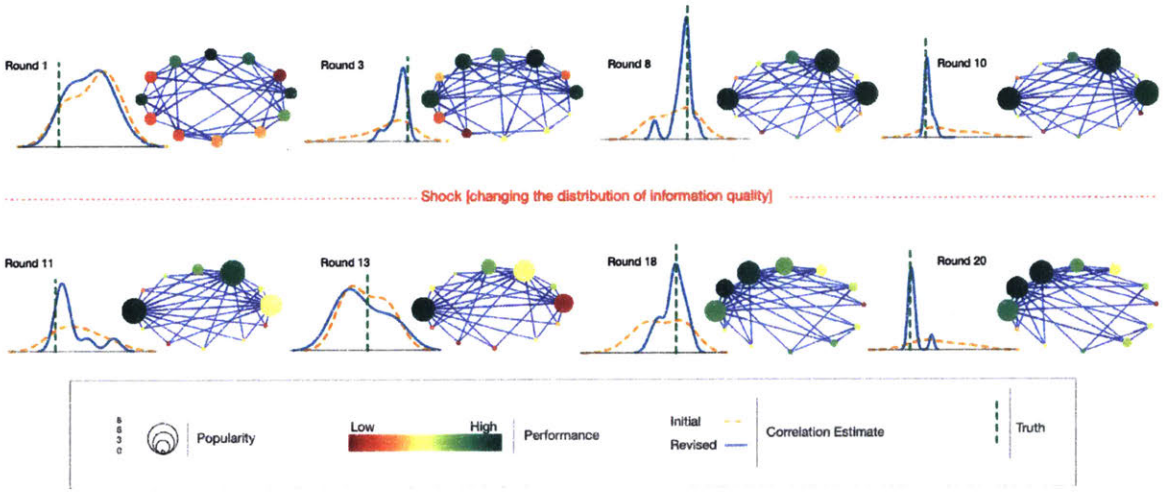


Figure 3-7: An example of the network evolution in the experiment. The circle color represents performance. The size of each circle represents the number of followers (i.e., popularity). The dashed orange line is the distribution of estimates prior to social influence, the blue solid line is the distribution of post-social influence estimates, while the dashed vertical line is the true correlation.

A centralization mechanism alone could suggest that group members may merely follow and imitate the best individual among them, hence bounding collective performance by that of the group's top performer. However, research on the *two-heads-better-than-one* effect indicates that, in the simpler case of dyads, even the best individual can benefit from social interaction [18, 114]; and that the critical mechanism enabling this effect is a positive relationship between individuals' accuracy and their confidence. Fig. 3-7C shows that participants in dynamic networks had, overall, a

positive correlation between the accuracy of their initial estimates and their self-confidence (measured in terms of resistance to social influence). Participants were likely to rely on private judgments whenever these were accurate and likely to rely on social information otherwise. Fig. 3-7C also shows that, as rounds elapsed, participants used task feedback to calibrate their accuracy-confidence relation gradually, and were able to re-adapt gradually upon the shock. Consistent with prior literature [26, 129], a positive correlation of confidence and accuracy was found in all networked conditions (i.e., including static networks in study 1), explaining their favorable performance compared to unconnected groups in both studies.

Mean-Variance Trade-off

The joint effect of centralization and confidence mechanisms explains the adaptive advantage of dynamic networks with feedback. Moreover, it suggests that their collective performance may not be bounded by that of the best individual, and that even the best individual may benefit from network interaction. To test these implications, we generalize the use of group means as collective estimates, common in 'wisdom of crowds' studies, and analyze the performance of *top-k* estimates—that is, collective estimates where only the guesses of the k best-performing group members are averaged. Top-k subsets within each group were computed based on ex-post individual performances across all rounds. In particular, *top-12* estimates correspond to the group mean, and *top-1* to estimates of the group's best-performing individual. Fig. 3-8 reports the mean and standard deviation of estimation errors incurred by *top-k* estimates during the adapted periods. Ideal estimates would minimize both mean error and variability, approaching the lower left end of the trade-off space.

The shape of *top-k* curves reveals that, as we remove low-performing individuals (from $k = 12$ to $k = 1$), estimates initially improve in both mean and standard deviation. Then, as we further curate the crowd beyond $k = 6$, *top-k* estimates trade off between decreasing mean error and increasing variability, and finally regress in both objectives as $k \to 1$. Comparison across conditions shows that, for any $k \in [1, 12]$, dynamic influence networks improved estimation errors in terms of both

mean and standard deviation. In particular, Fig. 3-8 shows that the full-group average in dynamic networks got 28% lower error and 48% less variability than the best individual in the *solo* groups (*dynamic top-12* vs. *solo top-1*; $P < 10^{-2}$). Moreover, even the best individual derived substantial benefits from social interaction, averaging 32% lower error and 38% less variability when forming and revising social connections rather than working in isolation (*dynamic top-1* vs. *solo top-1*; $P < 10^{-2}$).

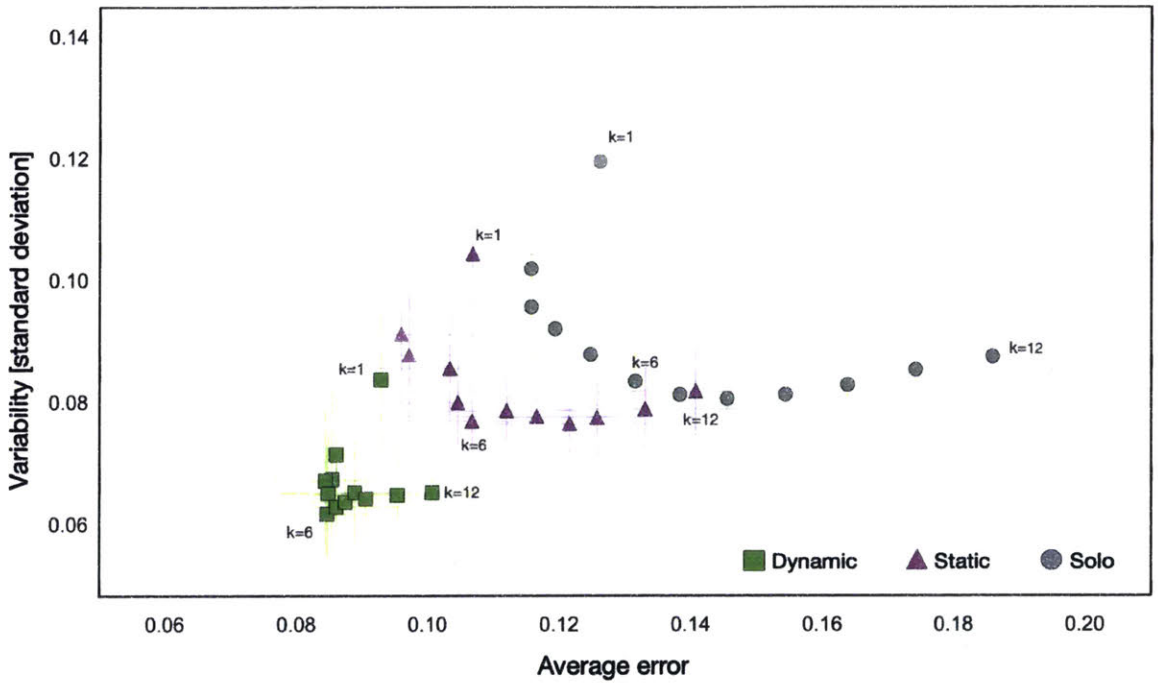


Figure 3-8: Mean-variance trade-off. Mean and standard deviation of absolute errors incurred by *top-k* estimates during the adapted periods (rounds $\in [6, 10] \cup [16, 20]$). *Top-12* estimates correspond to the full-group mean, and *top-1* to the group's best individual. Within each condition, *top-k* trade-off curves first gain in both objectives, then trade off lower average error for higher variability, and finally regress in both objectives as $k \to 1$. Across conditions, for any $k \in [1, 12]$, groups in the *dynamic* condition outperformed groups in the *static* and *solo* conditions. Moreover, the full-group mean of *dynamic* networks averaged 28% lower error and 48% less variability than the best individual playing *solo* (*dynamic top-12* vs. *solo top-1*; $P < 10^{-2}$); and the best individual in the *dynamic* condition averaged 32% lower error and 38% less variability than her analogue in *solo* (*dynamic top-1* vs. *solo top-1*; $P < 10^{-2}$). Bars indicate 95% confidence intervals.

3.5 Numerical Model and Simulations

We implemented numerical simulations to further assess the extent to which the interaction between the quality of performance feedback and the adaptability of dynamic networks. Therefore, we focus on two conditions: 1) traditional wisdom of crowds (i.e., independent actors with individual feedback); 2) adaptive wisdom of crowds (i.e., dynamic networks and full feedback). In order to follow the properties of our framework (see Figure 3-1), we need to operationalize models of its human components, i.e., the social learning and network rewiring heuristics. We model the former as a DeGroot process[54], and propose a performance-based preferential detachment and attachment model for the latter.

3.5.1 Model Specifications

Notation. Let $N = \{1, 2, ..., n\}$ represent a group of agents that participate in a sequence of tasks, indexed by discrete time t. Let $G(N, E^{(t)})$ be a sequence of directed graphs representing the influence network at each period t. Let $e_{ij}^{(t)} \in [0, 1]$ denote the edge weight of (i, j) at time t, and $M^{(t)}$ the row-normalized stochastic matrix associated with $E^{(t)}$, i.e., $M_{ij}^{(t)} = \frac{e_{ij}^{(t)}}{\sum_{h \in N} e_{ih}^{(t)}}$.

Agents receive private signals $s_i^{(t)} \in [0, 1]$, for $i \in N$, regarding the true state of the world $\omega^{(t)} \in [0, 1]$. Similarly, we denote agents' post-social learning beliefs by $p_i^{(t)} \in [0, 1]$, for $i \in N$.

Private Signals. We depart from the commonly made assumption of collective unbiasedness of agents' private signals [194, 127, 55, 82, 1], allowing agents' signals to be distributed with arbitrary means and skewness. Let $\mu_i = E[s_i]$ denote the mean of agent i's signal, and $\bar{\mu} = \frac{1}{n} \sum_i \mu_i$ be the collective mean of private signals; we are interested on the more general setting of information environments where $\bar{\mu} \neq \omega$, i.e., where the collective distribution of initial signals is not centered on the truth. Figure 3-9 illustrates the difference between unbiased and biased information environments.

Social Learning Process. Social learning is modeled as a DeGroot process [54],

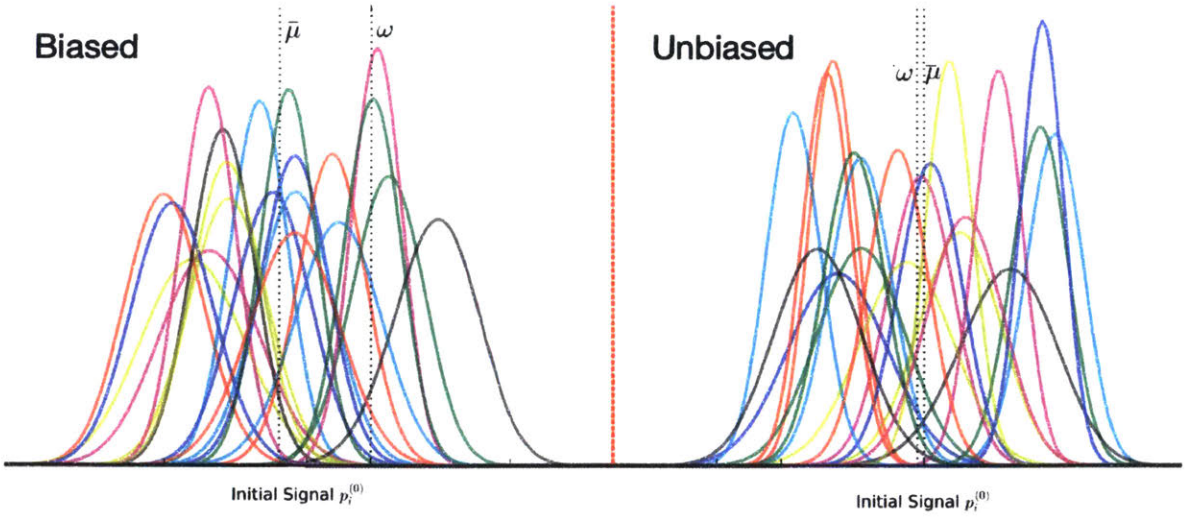


Figure 3-9: Traditional accounts of 'wisdom of crowds' phenomena assume unbiased and statistically independent signals among agents. In our model, we assume arbitrary (potentially biased) initial signals.

where each agent updates her belief by taking weighted averages of her own belief (i.e., private signal) and the beliefs of neighboring agents. DeGroot averaging as social learning heuristic has been well studied empirically and theoretically [55, 82], and shown to robustly describe real-world belief updating better than more optimal rational Bayesian models [43]. In particular, we model post-social learning beliefs as the result of a two-stage DeGroot process on private signals, given by

$$p^{(t)} = \left(M^{(t)}\right)^2 s^{(t)} \tag{3.1}$$

Individual performance is evaluated based on the errors of post-social influence estimates. Individual cumulative error is defined by:

$$\epsilon_i^{(t)} = \frac{1}{\lambda+1} \sum_{r \in [0,\lambda]} \left| p_i^{(t-r)} - \omega^{(t-r)} \right|,$$

where λ controls the number of retrospective periods that performance information is averaged across.

Agents assess performance of other agents relative to the performance of the best agent in the group. We define relative error of agent i as $\pi_i^{(t)} = \epsilon_i^{(t)} - \epsilon_{min}^{(t)}$, and

denote the set of performance information available to agent i at time t by vector $\Pi_i^{(t)} \in [0,1]^n$ with elements

$$\pi_{ij}^{(t)} = \begin{cases} \pi_j^{(t)} & \text{for } j \neq i \\ \pi_{s_i}^{(t)} & \text{for } j = i \end{cases}$$

where $\pi_{s_i}^{(t)}$ is the relative error of agent i's private signal.

Collective error. We are interested on the wisdom of the dynamic network (WDN) error, ϵ_{wdn}, which captures collective error after selective social learning according to the interaction network. We compare ϵ_{wdn} against the wisdom of the crowd (WC) baseline, ϵ_{wc}, which captures collective error of the simple averaging of agents' initial signals.

$$\epsilon_{wdn}^{(t)} = \left|\omega^{(t)} - \frac{1}{n}\sum_i p_i^{(t)}\right| \tag{3.2}$$

$$\epsilon_{wc}^{(t)} = \left|\omega^{(t)} - \frac{1}{n}\sum_i s_i^{(t)}\right| \tag{3.3}$$

Influence Rewiring Process. Individuals connect by weighted influence links that are revised over time. We model influence rewiring heuristics that strengthen links when a neighbor exhibits high performance and weaken or break links when a neighbor performs poorly. Agents can distribute attention among a limited number of peers, captured by parameter κ, which represents cognitive or infrastructure constraints (e.g., limits on our ability to keep track of social information and relations [64]). In particular, agents dynamically allocate $\kappa \in N$ shares of their attention to other agents. Let $e_{ijk} \in \{0,1\}$, for $k \in \{1,2,...,\kappa\}$, indicate that i places attention share k on j, then $e_{ij} = \sum_k e_{ijk}$ and $e_{ij} \in \{0,1,2,...,\kappa\}$.

Probability of detachment. Probability that agent i detaches from j is a positive function of i and j's errors, and given by equation 3.4. For example, if i's error is among the lowest of the group ($\pi_i^{(t)} \approx 0$), i is unlikely to rewire her local network. Conversely, if i's error is significant (e.g., $\pi_i^{(t)} \approx 1$), i detaches from j with

probability dependent on j's error [1].

$$\beta_{ij}^{(t)} = \left(\pi_i^{(t)}\, \pi_{ij}^{(t)}\right)^{\frac{1}{2}} \tag{3.4}$$

Probability of Attachment. High-performing agents are more likely to be followed. Analogous to generalized preferential attachment [21], probability that agent i attaches to j is inversely proportional to j's error, and given by

$$\alpha_{ij}^{(t)} = \left(\frac{1-\pi_{ij}^{(t)}}{n-\sum_j \pi_{ij}^{(t)}}\right)^2 c \tag{3.5}$$

where c is a normalization constant.

Network Evolution. Define i.i.d. random variables $b_{ijk}^{(t)} \sim Bernoulli(\beta_{ij}^{(t)})\ \forall(i,j,k)$, then random variables $b_{ij}^{(t)} = \sum_k b_{ijk}^{(t)} e_{ijk}^{(t)} \sim Binomial(e_{ij}^{(t)}, \beta_{ij}^{(t)})$ indicate the amount of attention shares that i detaches from j in period t. Define n-dimensional random vectors $a_i^{(t)} \sim Multinomial(\sum_j b_{ij}^{(t)},\ \alpha_i^{(t)})$, where $\alpha_i^{(t)}$ is i's vector of attachment probabilities. Elements $a_{ij}^{(t)} \in \{0, 1, ..., \kappa\}$ indicate the amount of shares that i attaches to j in period t , and network evolution is given by

$$e_{ij}^{(t+1)} = e_{ij}^{(t)} - b_{ij}^{(t)} + a_{ij}^{(t)} \qquad \forall i,j \tag{3.6}$$

3.5.2 Simulation Results

Using the above-described model, we performed Monte Carlo simulations of a group of twenty agents who participate in a sequence of estimation tasks, where agents can follow and be influenced by a maximum number of five peers ($\kappa = 5$). We intentionally chose parameter values that differ from our experiments in order to examine the robustness of our findings under different parameter values. Indeed the results of these simulations corroborate our main experimental results that relates adaptability in non-stationary information environments to plasticity and feedback. Figure 3-

[1] Analogous to the Watts-Strogatz network rewiring model [211], the expression's exponent can be parameterized to reflect context-dependent costs of rewiring edges, due to infrastructure or cognitive constraints.

10 highlights how the dynamic agent network responds to environmental shocks. Figure 3-10B compares the dynamics of collective intelligence for WC and WDN under a non-stationary environment, where shocks to the joint distribution of private signals $p_s^{(t)}$ and truth $\omega_s^{(t)}$ are introduced at $t = \{100, 200\}$. The dynamic network adapts to post-shock distributions by shifting influence to agents with better information in the post-shock environment, driving collective error $\epsilon_{wdn}^{(t)}$ significantly below $\epsilon_{wc}^{(t)}$. Difference in means tests showed that $E[\epsilon_{wdn}^{(t)}] < E[\epsilon_{wc}^{(t)}]$ for all $t \in \{[3, 99] \cup [140, 199] \cup [225, 300]\}$ with a 95% confidence level. Therefore, dynamic networks indeed adapted to shocks by shifting influence weight to agents with better information, substantially decreasing individual and group error.

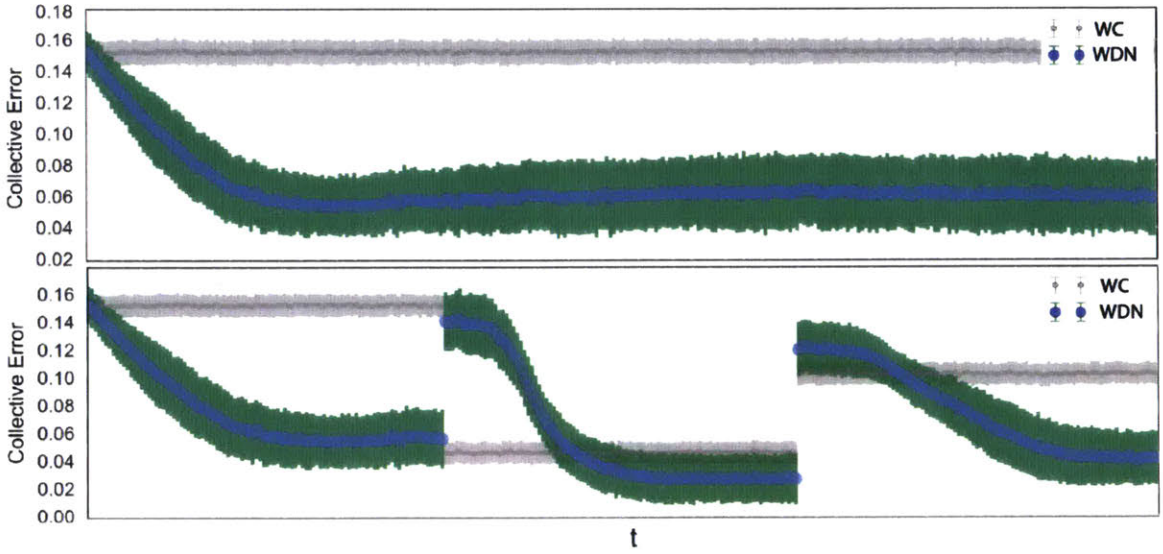


Figure 3-10: Evolution of collective error: wisdom of the crowd (WC) and wisdom of the dynamic network (WDN). Panel A) stationary distribution of information among agents. Panel B) non-stationary information environment, shocks to the information distribution introduced at $t = \{100, 200\}$

Moreover, simulation results show that accurate peer performance information (i.e., high quality feedback) is necessary for enabling beneficial group adaptation by means of social rewiring. Figure 3-11 shows that, as we add increasing noise of peer performance information, the collective performance of adaptive networks deteriorates until converging to that of the simple wisdom of crowds[2].

[2]This is a realistic assumption as usually the environmental cues about performance can be noisy in many cases.

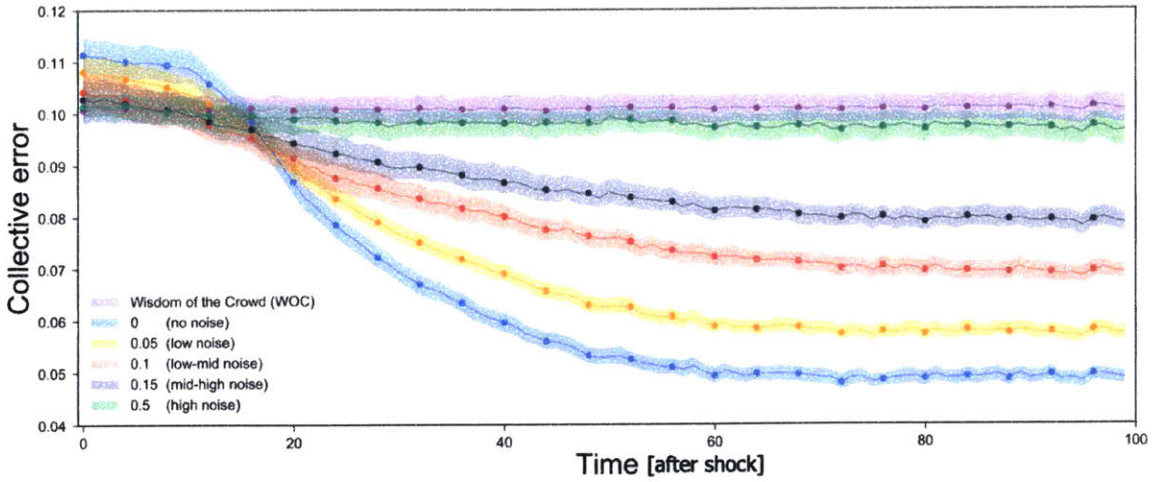


Figure 3-11: As the nose level increases in the provided feedback, the collective performance degrades until it converges to the performance of the independent crowd.

Lastly, we explore through simulation the interaction between network learning rates—a network's sensitivity to changes in agents' performance, parameterized by λ—and the arrival rate of environmental shocks 3-12. Networks with faster learning rates could adapt to environments with frequent information shocks. Conversely, networks with slower learning rates could leverage longer learning periods, eventually achieving lower error rates in environments with infrequent shocks. This short-term versus long-term accuracy trade-off implies that optimal network learning rates depend on the pace at which the information environment changes, analogous to notions of optimal learning rates in natural systems and artificial intelligence algorithms [30, 113].

3.6 Chapter Summary and Reflections

Social networks have a strong influence on how people form judgments and make decisions. We address the question whether the structure of such networks can adapt to leverage the skills of individuals and promote collective intelligence. In our experiment, groups of participants were embedded in social networks and asked to solve a series of estimation tasks. We show, for the first time to our knowledge, that groups in dynamic networks—where network structure can change by forming and breaking

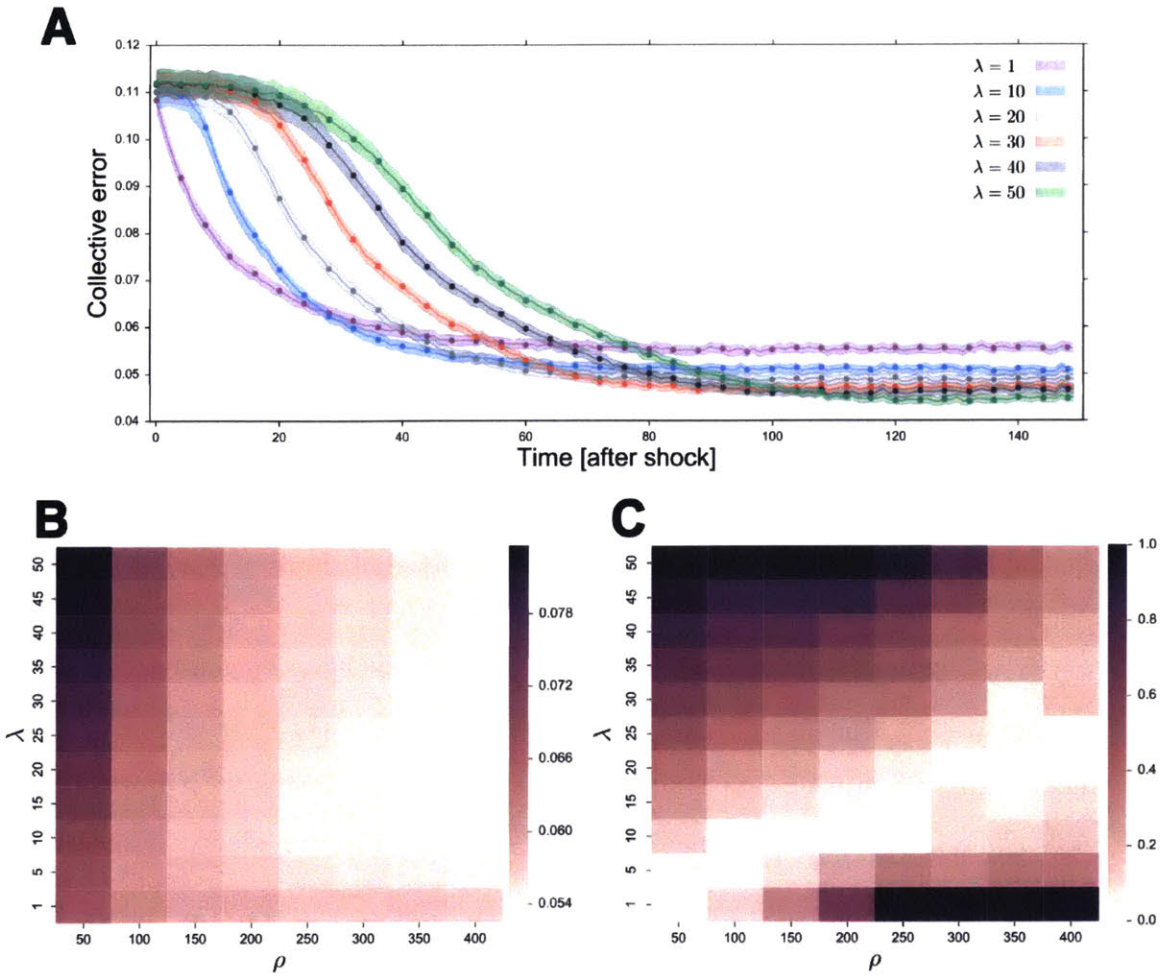


Figure 3-12: Panel (**A**): Learning rates associated to different λ's, where colored bands show 95% confidence intervals. Panel (**B**): Effects of λ and ρ on collective error, where shades of orange indicate time-averaged collective error. Panel (**C**): Effects of λ and ρ on collective error, normalized per type of information environment (ρ column).

ties in response to peers' performance—improve individual and collective performance substantially (compared to static networks and unconnected groups), and far outperform even their best-performing member in isolation. Such findings highlight the role of adaptive social networks as prime mechanisms for refining individual judgments and inducing the collective 'wisdom of the network.'

We acknowledge that the results of laboratory experiments and numerical models do not translate directly into the real world, the evidence presented here suggests that details of interpersonal communications—both in terms of the structure of the social interactions and the mechanism of its evolution—can have an effect on the ability

of the system to promote collective intelligence. This is an evidence that dynamism of the network has profound effects on the processes taking place on them, allowing them to be more efficient, stable against perturbations, and able to adapt to non-stationary environments. This can help/motivate the design of field experiments in a real system, which would narrow the gap between stylized experiments and real-world social contexts.

The insights here provided suggest design guidelines germane to real-world collective intelligence mechanisms, in contexts such as commodity markets, social trading platforms, crowdfunding, crowd work, citizen science, prediction markets, and online education (e.g., MOOCs). We expect the adaptive systems view on collective intelligence to further sprout connections with fields such as evolutionary biology and artificial intelligence, advancing an interdisciplinary understanding and design of social systems and their information affordances.

Chapter 4

The Virtual Lab: High-throughput Social Science

Behavioral labs have long played an important role in all social science disciplines. The "Behavioral Lab" approach is a useful tool that offers a great degree of control and allows for the identification of causal effects. However, many social phenomena of interest to researchers and policy makers alike involve large populations interacting in complex non-stationary environments over extended periods of time. By contrast, behavioral experiments have historically been restricted to small N samples of WEIRD[1] subjects interacting in highly simplified environments over very short (i.e., up to 1 hour) intervals. Consequently, the results of even well designed and run lab experiments suffer from severe external validity problems.

Another related problem, which is even more relevant to the framework of this dissertation, is that social theories are rarely precise enough to estimate exact parameter values from empirical data; thus, a robust test of even a single theoretical claim may require many experiments, each corresponding to a different set of parameters. Unfortunately, the costs and logistics involved in running lab experiments typically restrict researchers to exploring a tiny fraction of the relevant parameter combinations, thereby leading to fragile and inconsistent findings that in turn lead to the

[1]Henrich's research [92, 91, 178, 93] demonstrated that people with a Western, Educated, Industrialized, Rich, and Democratic background — the WEIRD people — are not representative of humans at large, but rather outliers.

problem of incoherency discussed in Chapter 1 (in particular, Section 1.2).

We propose to address both sets of problems by dramatically scaling up and speeding up the current state of the art in virtual lab technology. Specifically, we intend to ease three main bottlenecks to existing research capabilities:

- **Size.** Alleviate scaling and replication difficulties by recruiting and maintaining a large and diverse pool of subjects (see Section 4.2.1).

- **Duration.** Relieve the constraints on location in order to enable the design of experiments with human interactions at different time intervals (see Section 4.2.2).

- **Complexity.** Diminish the technological cost associated with building "immersive" environments and more realistic tasks that are characterized by a large set of parameters (i.e., beyond the 2×2 payoff-matrix[2]; see Section 4.2.3).

Achieving these goals will require virtual lab software (e.g., Empirica.ly [156]) that is optimized for reducing the overhead associated with building and running experiments (see Section 4.1), and in return, it will allow research teams to coordinate research designs by recruiting a community of researchers to collaborate on a single research program (e.g., optimal team composition, increasing cooperation in real-world scenarios, influence maximization on networks) to explore the parameter space of social theories and maximize cumulative knowledge (see Section 4.2).

4.1 The Interactive Environment

The idea of web-based "virtual labs" to create experiments with rich interactive environments and crowdsourced labor from the internet has started to gain popularity. Although the number of synchronous online experiments is on the rise [175, 132, 133], most of them make use of customized implementations, leaving a large number of

[2]Stylized economic experiments like prisoner's dilemma shows how two rational individuals with two possible actions (i.e., cooperate or defect) might not cooperate, even if it appears that it is in their best interests to do so.

open methodological challenges yet to be solved [19]. So far, researchers trying to pursue this approach get side-tracked mostly by the tedious logistics of randomization, storing data, managing participants, synchronization, waiting rooms, setting up an infrastructure, etc. It can be frustrating for the researchers to put all that effort in before even getting to the experiment. This is a lot of work, and this work is redundant. It also encourages bad practices like copy-pasting boilerplate from someone else's code without understanding it and slowing down the pace of science by building ad-hoc experiments that are not reusable by others.

While the virtual lab is a promising methodology and a growing area of research, the field is still young, and it lacks established software for conducting synchronous online experiments. There are a few platforms that allow researchers to run online experiments. For example, commercial products include Testable.org and Gorrila.Sc.; while non-commercial (open-source or free) products include jsPsych [51] and PsyToolkit [189, 188]. However, these are designed to support questionnaires and single-participant reaction-time experiments (i.e., no group interactions or multiplayer types of games). Experimental software like Breadboard [139], Z-tree [73], and oTree [44] do support group interactions, but they are originally designed for sequential interactions (not continuous) and for an insufficient number of highly constrained settings (e.g., stylized economic games or studies of social networks). Finally, nodeGame [19][3] and TurkServer [131] are flexible and support real-time group interactions; however, they require relatively substantial programming expertise.

Therefore, we decided to build our own platform, Empirica (https://empirica.ly/), which is a free, open-source framework that allows researchers to conduct behavioral experiments of a scale, duration, and realism that far exceed what is possible in brick and mortar facilities. The goal is to address the problem of long development cycles required to produce software to conduct online experiments. It handles all the logistics and allows researchers to go straight to their research questions. Also, it provides data and experiment design layers that allow different researchers to coordinate designs.

[3]This is probably one of the most flexible tools currently available for developing synchronous games.

Empirica has modular source code and defines an API (application programming interface) through which experimenters can create new strategic environments and configure the platform. The deployment of the experiment happens from a live web interface and allows the researcher to watch the progress in real time with the ability to create one-way mirrors to observe the behavior of participants. With no installation required on the participants' part, experiments can run on a great variety of devices, from desktop computers to laptops, smartphones, and tablets. This software has been used to run the experiments in Chapters 2 and 3.

4.2 Towards Expanding the "Lab Experiment" Design Space

As we have discussed in the beginning of this chapter, the "virtual lab" approach is intended to lift the historical barriers along three major conceptual dimensions: (i) size, (ii) time, and (iii) complexity (see Figure 4-1).

4.2.1 Size: In Complex Systems, Large is Different

Morris Zelditch in 1969 argued in his paper "Can you really study an army in the laboratory?" that it is neither possible nor necessary to study large human organizations in the lab [221]. Zelditch asserted that it is sufficient to test social theories in small-group experiments and then use theory to generalize the results to the larger group. Today, we know that this perspective is inadequate. In complex systems, the collective behavior is not merely the sum of the individual components. We cannot understand how an ant colony operates by studying individual ants. In the same way we cannot understand social systems—such as markets, organizations, institutions—by studying individuals or small groups in the lab setting. Groups behave differently at different scales [212, 196]. For example, Elinor Ostrom[4] found empirically a non-linear relationship between community size and its ability to protect the commons [159]. Also,

[4]The first—and so far, only—woman to win a Nobel Memorial Prize in Economic Sciences.

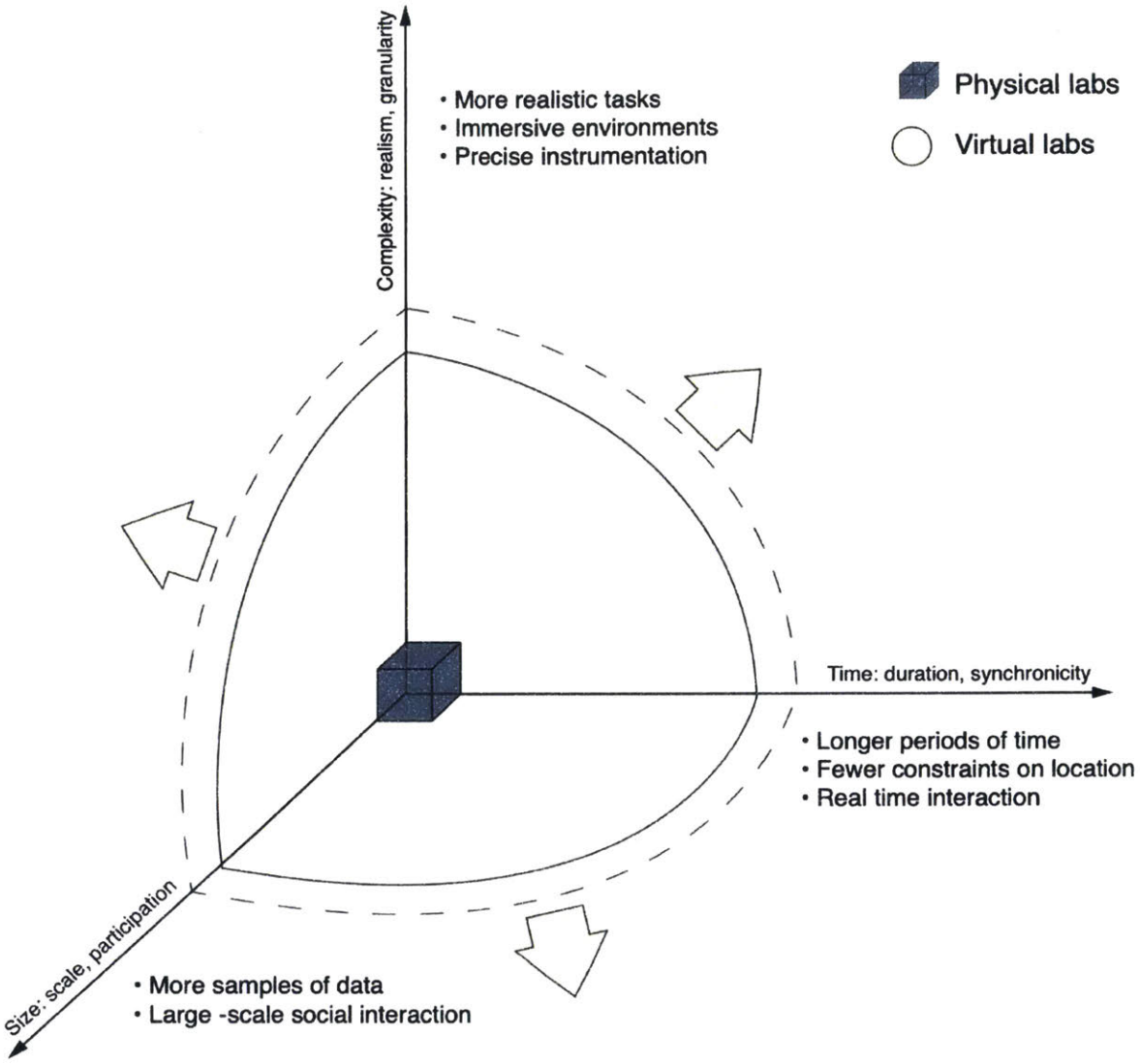


Figure 4-1: The Virtual Lab Framework. The figure illustrates the three conceptual dimensions for virtual lab experiments.

experimental work has shown some interesting relationships between group size and collective performance such as generating complex cultural artifacts [56][5], mapping disasters [133], and disrupting science and technology [218]. This explains why the municipal is different from the national and why the United States is not just $\times 1,000$ Singapores.

Therefore, when it comes to social theories, the size of a group is an important parameter that can change the qualitative behavior of interacting individuals. It

[5] Although this result is controversial [14].

seems that we do need to study the army in the lab, after all.

Unlike the physical lab, which is constrained by the behavioral lab space at a university, the virtual lab—in theory—has no limit to the number of participants. In particular, shifting the meaning of a "large group" from a couple of dozens to hundreds of participants has been enabled by the availability of a large and cheap labor market for research. The ability to crowdsource participants has had a large impact on human subjects research, from the computational sciences to the behavioral sciences. For instance, crowdsourcing labor from Amazon Mechanical Turk (MTurk) is an increasingly popular tool for conducting behavioral studies. There have been efforts to systematically replicate classic results from the social sciences [17, 99, 28], with outcomes that in many cases appear to be as reliable as data obtained via traditional methods.

The use of crowdworkers has had a profound influence on the nature and pace of data collection and has opened new avenues to cost-effective replication and extension of familiar research paradigms. This shift has had special impact in areas such as studies of cooperation and conflict, person perception, intergroup attitudes and stereotypes, and group behavior, where cumbersome interactive multi-participant experiments can be conducted much more easily via online platforms [88, 9].

The advances in scale offered by online labor markets for crowdsourcing participants, though significant, are not without limitations. In particular, there are at least three challenges yet to be solved [134]: 1) recruiting simultaneous participants (i.e., availability at the same time to study interaction between participants); 2) participants' uniqueness (i.e., avoiding learning affects [163, 42], where prior exposure to a task affects subsequent experimental results on similar tasks); and 3) large sample size (i.e., at any given time there are *only* around 2K active high-effort workers on MTurk [62]).

4.2.2 Timescale: Social Interactions Evolve Over "Time"

Social interactions between individuals evolve over time (as we have discussed in Chapter 3), and the nature of *time*—be it simultaneous or sequential; real-time or

offline; continuous or discrete; fast or slow; one-shot or repeated—can fundamentally change group-level outcomes [182, 170, 31].

A notable example is the dynamics of cooperation (i.e., paying a personal cost for a shared benefit). In Prisoner's Dilemma, where defection is the prevailing action in one-shot interactions, cooperation can be sustained when interactions are repeated and participants can remember previous actions of their peers [153]. Moreover, cooperation can be further sustained in long-run experiments (i.e., lasting for 20 consecutive weekdays [132]) and real-time interactions (i.e., going from discrete-time to continuous-time increases cooperation from 40% to 90% [76]).

Virtual lab experiments grant great flexibility to experimenters to study human interactions at different time intervals, from one-shot real-time (i.e., seconds) to very long sequential (weeks and months) interactions— thereby allowing for more "immersive" environments.

Despite the opportunities virtual labs can bring to studying human behavior, behavioral research online has so far remained largely limited to offline decision-making tasks[6] or one-shot interactions with simultaneous decisions. This is partly because of the lack of an established software that is widely used for conducting synchronous online experiments, which leaves a large number of open methodological challenges yet to be solved by the experiment designer.

4.2.3 Complexity: The Parameter Space of Social Theories

Throughout this dissertation, I have argued that the importance of a result needs to be evaluated through the lens of the environment—that is, the qualitative behavior of an attribute changes as a function of the combinations of parameter values chosen by the researcher (cf. Section 1.3).

For instance, the collective intelligence literature can be mapped into a very high-dimensional parameter space, where each dimension represents what the investigator thinks is a relevant attribute. A non-exhaustive list of potentially relevant variables for

[6] E.g., using the strategy method, where decisions for each possible information set are collected, then interactions are emulated post hoc.

team composition and performance is: scale (2 individuals - 10^4 individuals), problem dimensionality (estimation tasks - complex problems), problem difficulty (easy that anyone can solve; hard, or no one can solve alone), interdependence between task components, means of communication (slider, language, cursor, price, wager, probability judgment), communication structure (fully connected - chain - nominal), information access and feedback, goal (forecasting, searching, estimating, acting, transferring information), timescale (one-shot, iterative), etc. The parameter space of these theories has been sparsely investigated with ad hoc experiments, each solving specific problems with custom operationalization and often reaching simplistic conclusions that are in contradiction with each other and hard to reconcile (cf. Section 1.2 on the *incoherency problem*). A common language between researchers is missing and the results remain isolated rather than cumulative. With a virtual lab framework, researchers could solve the problem of exploring the parameter space[7] empirically in at least two ways:

The Exhaustive Exploration. Due to the lower cost associated with running virtual lab experiments, researchers could test multiple combinations of parameters of behavioral theories and generate empirical "phase diagrams," where the collective outcome of interest is observed against all the possible combinations of model parameters[8].

The Coordinated Exploration. The exploration of this massive space can potentially proceed in a distributed and guided fashion[9]. The following would be a possible iteration, where a community of researchers collaborate on a single research program (e.g., optimal team composition, increasing cooperation in real world scenarios, influence maximization on networks) coordinate research designs (e.g., define the relevant variables and outcomes) to maximize cumulative knowledge:

- The researcher asks the "framework" what experiment to run, specifying what parameters to hold fixed (e.g., task characteristics) and which are allowed to vary (e.g., team compositions).

[7]The exploration of parameter space is usually done through computational or analytical models.
[8]In Section 1.3.1 we illustrate this using an analytical model, instead of empirical; see Figure 1-1
[9]This is based on discussions with Niccolo Pescetelli, Duncan J. Watts, and James A. Evans

- The platform suggests experiment parameters to be evaluated (e.g., scale=50 individuals, means of communication=language, structure='teams', skill diversity= high, etc.).
- The researcher implements the experiment using the provided parameters.
- The results are fed back to the framework, which uses them to update its parameters and proceeds to the next iteration of experimentation.
- Code and results (that are compatible and comparable) are made publicly available and published in peer-reviewed journals.

4.3 Reflections and Conclusions

In conclusion, group attributes (e.g., network structure, team composition, individual-level attributes) and collective performance are multifarious constructs each of which can be operationalized in many ways. Moreover, the relationship between the two may be contingent on numerous other mediating variables related to the nature of the environment. Therefore, without an environment-dependent framework from which to draw hypotheses and tune our intuitions, it is difficult to distinguish results that are unusual and interesting from results that are unusual and probably irrelevant.

Although the idea that the environment shapes human behavior has been generally accepted, there are two mainstream criticisms of this approach. First, there are concerns regarding the associated difficulty of manipulating and measuring environment (i.e., how can one possibly sample situations?). In this dissertation, we showed that focusing on the formal properties of the environment (i.e., defining the universe of possible environments) and the use of modern virtual lab technologies can effectively overcome this limitation. The second objection argues that, even if we could define and sample the environment, there is no need to do so. After all, the goal of the social scientist is not to generalize the results from the experiment to 'outside' situations but to test hypotheses and advance particular theories. However, recent movements in social science have argued that social science should be more

"solution-oriented" to reconcile the competing claims in the literature. That is, the research community needs to place more emphasis on solving practical problems—the sort with direct engineering analogs [210]—rather than the advancing of particular theories. In this dissertation, we have followed a "solution-oriented" approach by advancing our fundamental understanding of collective intelligence in the course of solving applied problems.

In future work, we hope to apply the same approach to qualitatively different environments by varying other parameters of interest (e.g., group size, communication patterns, division of labor, leadership). Although such a program would be logistically challenging, "virtual lab" experiments of the sort that we have described here, in combination with emerging "open science" practices such as pre-registration, data availability, open code, and "many-labs" style collaborations, offer a promising route forward. In order to operationalize the "environment," we built an experimentation platform (Empirica.ly). The platform forces the investigator to explicitly define the space of the environment in which the group of participants is situated, and therefore, the exploration of the interactions between the environments and the attributes of interest becomes more systematic (as opposed to having isolated and non-comparable studies). We hope that our emerging ability to conduct *virtual lab experiments*—of a scale, duration, and realism that far exceed what is possible in brick-and-mortar facilities—will blur the traditional boundary between "lab" and "field" experiments and revolutionize our understanding of human behavior, not only for the design of social science experiments but for rebuilding society as a whole.

Appendix A

Supplementary Information for Chapter 2

A.1 Room assignment task

Payoff

Rooms	101	102	103	104
Student A	10	47	32	20
Student B	78	65	46	37
Student C	35	59	41	53
Student D	40	65	12	43
Student E	18	39	40	78
Student F	22	51	57	40

Constraints

- B and E must be neighbors.
- C and F can't live in the same room or be neighbors.

Figure A-1: An illustration of the "room assignment" task used in phase one of the experiment. In this case, there are $N = 6$ students that need to be assigned to $M = 4$ rooms, while satisfying $Q = 2$ constraints.

Timer 09:51 Score 0

Rooms	101	102	103	104	105	106	107	108
Student A	27	51	31	25	47	45	41	54
Student B	44	40	11	62	46	22	27	53
Student C	24	17	38	42	32	64	58	41
Student D	50	26	57	28	54	11	17	32
Student E	47	38	51	12	16	34	59	48
Student F	37	63	48	61	24	28	57	39
Student G	77	61	12	72	66	27	53	35
Student H	68	44	35	58	61	11	24	45
Student I	31	51	16	47	38	58	35	66
Student J	54	55	61	68	24	46	70	42
Student K	66	29	63	74	35	60	43	50
Student L	15	63	19	43	51	28	34	40
Student M	43	50	36	21	17	56	39	47
Student N	33	50	68	44	29	43	54	36
Student O	17	32	45	68	54	16	33	53
Student P	30	34	17	44	49	25	16	18
Student Q	49	29	61	37	39	20	62	48
Student R	19	25	28	50	54	37	42	20

Constraints

- A and B must be neighbors.
- A and J can't live in the same room or be neighbors.
- B and F must be neighbors.
- ✕ B and H must live in the same room.
- B and P can't live in the same room or be neighbors.
- ✕ C and E must be neighbors.
- C and J can't live in the same room.
- D and F can't live in the same room or be neighbors.
- E and Q can't live in the same room or be neighbors.
- F and I can't live in the same room.
- ✕ G and J must be neighbors.
- I and N can't live in the same room or be neighbors.
- J and K can't live in the same room or be neighbors.
- ✕ K and R must be neighbors.
- L and Q can't live in the same room or be neighbors.
- M and O must live in the same room.
- N and R must live in the same room.
- O and P must be neighbors.

Figure A-2: An illustration of a more difficult "room assignment" task. In this case, there are $N = 18$ students that need to be assigned to $M = 8$ rooms, while satisfying $Q = 18$ constraints.

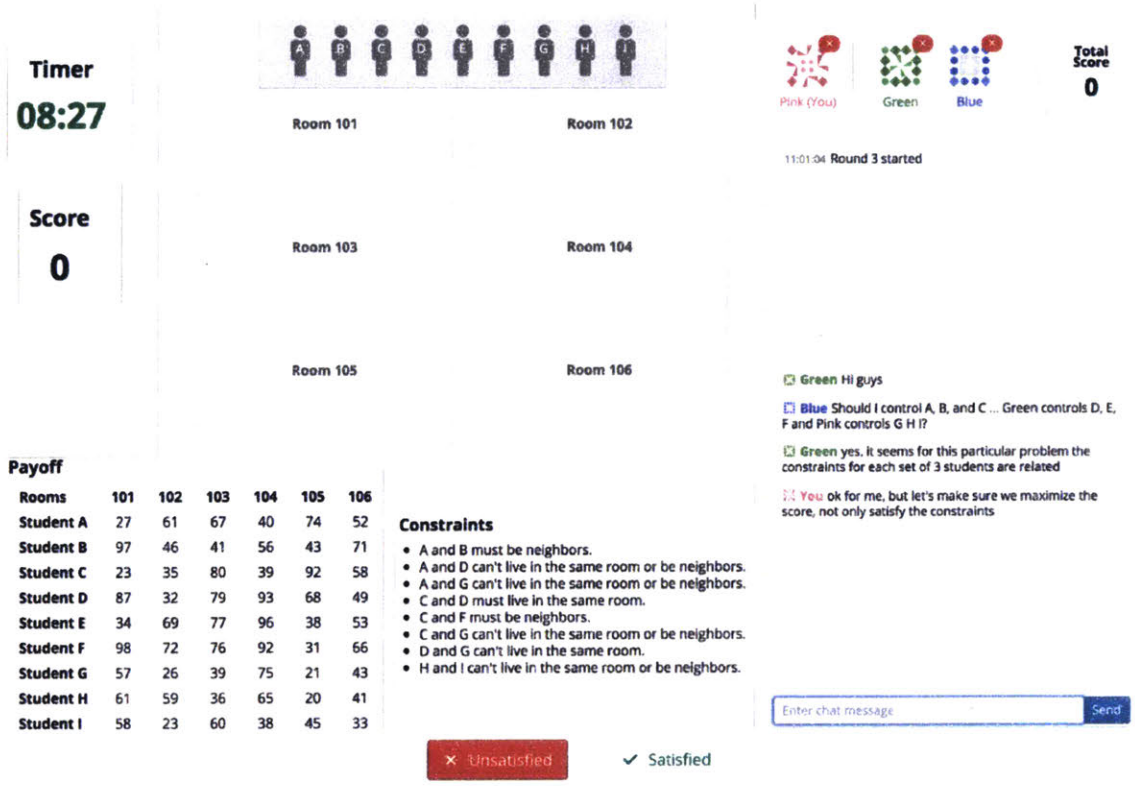


Figure A-3: An illustration of phase two "room assignment" task that was done by a group of three individuals in phase two.

A.2 Reading the mind in the eye

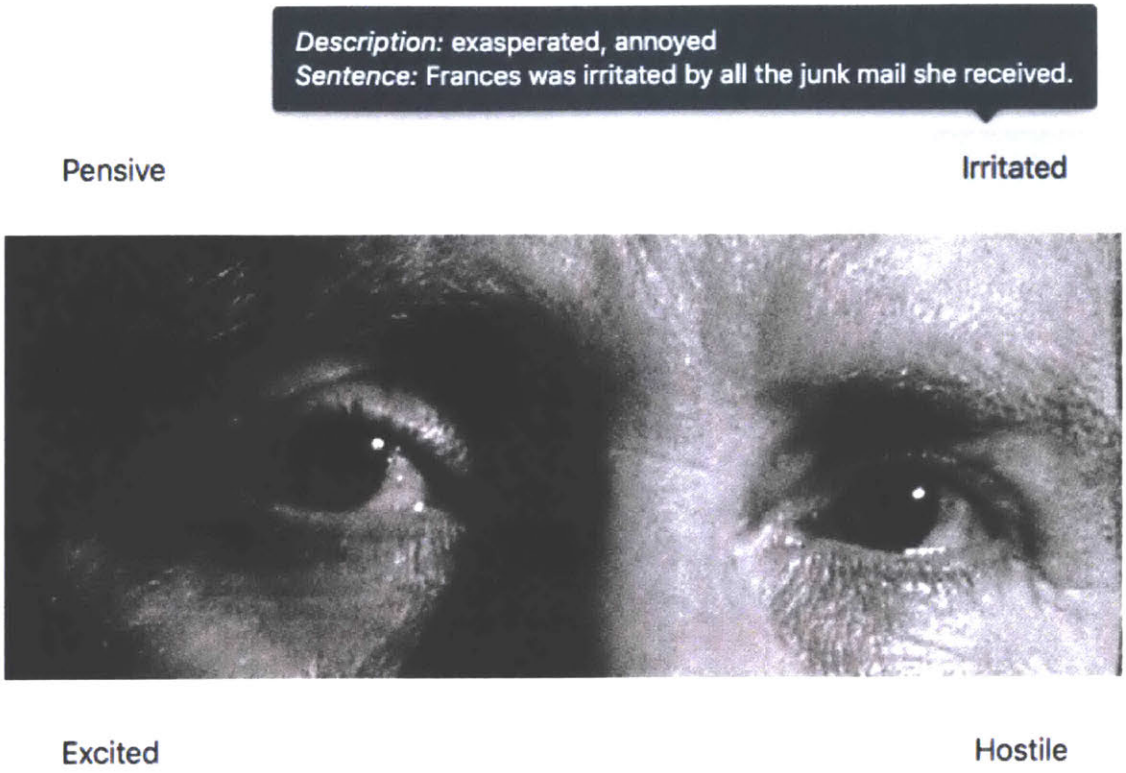


Figure A-4: An illustration of the "Reading the Mind in the Eye" test used in phase one of the experiment. The participant is shown a pair of eyes and asked to choose the emotion that best describes what the individual in the picture is feeling or thinking of.

A.3 Screenshots of the instructions and comprehension check

Game Overview

In this game, you will be **asked to solve a sequence of resource allocation tasks**. In each task, you are going to **assign a group of students into dorm rooms**. You are asked to find the room assignment plan that maximizes overall satisfaction for the group while respecting certain constraints (e.g., some students can not live together in one room).

You have at most **10 minutes** to work on each task. Completing the entire game may take you as long as 60 minutes. **If you do not have at least 60 minutes available to work on this HIT please return it now.**

You will play this game simultaneously with 2 other participants in real-time. As we will explain in more detail later, in each task, you and your teammates will submit a single room assignment plan.

At the end of the game, you will have the opportunity to earn a bonus payment and the amount is dependent on your accumulated score in all tasks. **Note that "free riding" is not permitted**. *If we detect that you are inactive during a task, you will not receive a bonus for that task*

The game **must be played on a desktop or laptop**. There is NO mobile support

For the best experience, please maximize the window containing this task or make it as large as possible.

« Previous Next »

Room Assignment Tasks

In each task (or round), you will be asked to **assign students to dorm rooms**. Students express their degree of satisfaction for living in a room as a number between 0 and 100 (the higher the rating, the more satisfied the student is).

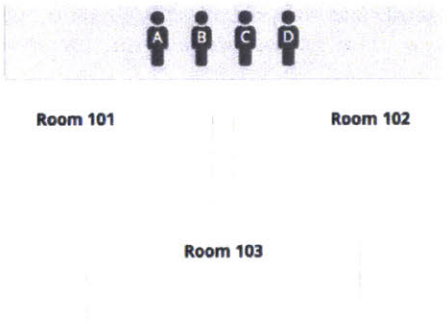
You are provided with a handy **drag and drop** tool to solve the problem. To assign a student into a room, drag the icon of that student and drop it into the room. Try this example:

Payoff

Rooms	101	102	103
Student A	20	80	65
Student B	67	90	76
Student C	85	82	79
Student D	20	75	78

Score

N/A



NOTE: ALL the students HAVE to be assigned to a room in order for your score to count.

« Previous Next »

Respecting the Constraints

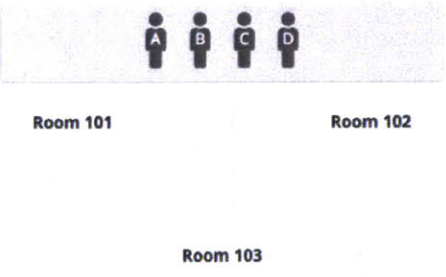
You need to **consider some constraints when assigning students to rooms**. Some students can't live together in the same room and some students must be neighbors.

These constraints vary from task to task, and there are no additional constraints you need to respect other than the ones stated (e.g., feel free to leave one room empty if no constraint requires you to assign at least one student in each room).

Try this example again and see what will happen if a constraint is violated:

Constraints

- A and B can't live in the same room or be neighbors.
- B and C must live in the same room.



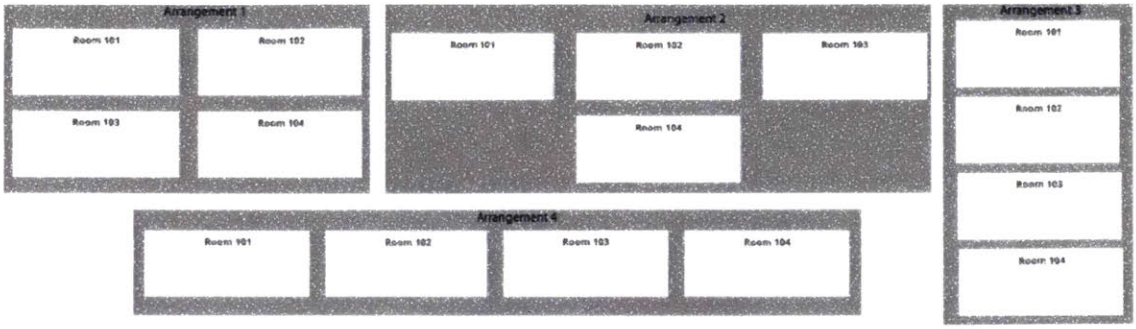
NOTE: Every violated constraint will result in deducting 100 points from your score.

NOTE: It is OK to leave some rooms empty, but you have to assign all the students.

« Previous Next »

Task Room Arrangements

Depending on the number of rooms, number of students, and your screen/browser size and resolution, the arrangement of the rooms might "look" different on your screen.



In all cases and for any arrangement that appears for you, you only need to consider the numbers on those rooms when addressing constraints in a task. In particular, **"neighbor" is defined as rooms with consecutive numbers**. For example, regardless of the arrangement you have on the screen, Room 102 is next door to both Room 101 and Room 103. On the other hand, Room 101 is only next door to Room 102.

« Previous Next »

You will be part of a team

In this game, you will **play together with 2 other participants (your teammates)**. They are other MTurk workers who are undertaking the same study simultaneously. Throughout all the tasks, the team will submit only one answer, and therefore, **all members of the team will receive the same score**. To help you identify yourself and differentiate each other in the team, we will assign a color to you when the game starts (as shown in the following example).

Note that the game allows for simultaneous and real-time actions. That means that you will be able to drag students to assign them to rooms while your teammates are doing the same. However, when any member in the team starts dragging a student, that student will be locked (i.e., no one else can move it) until it is assigned to a room. **The student that is being moved will have the color of the participant**.

« Previous Next »

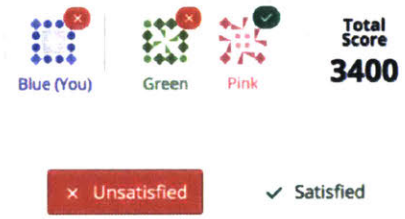
Event Logs and In-Game Chat

We will log every action taken by you or any of your teammates, and this log will be shown to you to help you keep track of all the actions that have taken place so far.

Also, you may communicate with your teammates through the in-game chat. This chat room is public so whatever you write will appear to the other 2 teammates. You can use this in anyway you want.

Remember, you and your teammates have 10 minutes in each task to find a room assignment plan. You will automatically **progress to the next task when the time is up**.

However, you can always indicate whether you are satisfied with the answer before the timer is up (indicated by the check mark on the avatar). Click on the "Satisfied" button in the following example and see what happens!



If all team members are satisfied with the answer before the timer is up, the answer will be submitted and your team will proceed to the next task. If the "Satisfied" button is unclickable (i.e., inactive) for you for more than 10 seconds, try to refresh the page..

« Previous Next »

Scores and Bonuses

In each task, we use "score" to evaluate the quality of the room assignment plan that your team came up with. **Your score starts counting only when you have a complete assignment** (that is, each student has been assigned to a room).

The score of your assignment is calculated as:

S = The sum of students' ratings of their assigned rooms - 100 * the number of violated constraints

That means, **for each constraint you violate, you get 100 points deducted.**

As a team, **you will submit ONE answer per task** and therefore **all team members will have the same score on each task**.

There are two parts of the bonus that you will have opportunity to earn in each task:

1. **"performance-based bonus":** When your score is positive, no matter whether your answer is the BEST possible assignment or not. The exchange rate is **1000 game points = $1 bonus**.

2. **"optimal assignment bonus"** : When your answer is the BEST possible assignment, you get **an additional bonus of $0.7 in that task**.

Therefore,**big part of the bonus is for finding the BEST possible assignment** (i.e., "optimal assignment bonus", which can be up to $3.5 total). Also,**you can earn more game points (i.e., more performance-based bonuses) from the difficult tasks** compared to the easier ones (more students/rooms means more possible bonus).

Together with your teammates, you should try to find a complete room assignment with a score that is as high as possible to earn more bonus in each task!

Remember, free riding is not permitted. If we detect that you are inactive during a task, you will not receive a bonus for that task.

« Previous Next »

Quiz

How many participants will play at the same time, including yourself?

Select the true statement about the score:

- I will score points only based on the assignments that I make
- We will submit only one answer as a team and therefore we will all get the same score.

is it ok to have some rooms empty? (the answer is 'Yes')

- Yes!
- No!

If your team ended up NOT assigning all students to room (i.e., at least one student remained in the deck) then your score in that task will be:

For each unsatisfied (i.e., violated) constraint, how many points will be deducted from you?

Which of the following rooms is a neighbor of Room 101? Please select all that apply.

- Room 101
- Room 102
- Room 103
- Room 104
- Room 105

Which of the following rooms is a neighbor of Room 103? Please select all that apply.

- Room 101
- Room 102
- Room 103
- Room 104
- Room 105

« Back to instructions Submit ↵

A.4 Validity of participant's individual skill measure

In our experiment, we defined an individual participant's skill score as the sum of her scores on the two hard tasks in phase one experiment, and we further labeled the participant as "high" or "low" on skill by examining whether her skill score was larger or smaller than the median score obtained among all participants. To illustrate the validity of this measurement of skill level, Figure A-5 contrasts the normalized scores (i.e., actual score obtained in a task instance / the maximum possible score for that task instance) obtained by "high skill" participants with those obtained by "low skill" participants, on each of the six tasks in phase one, including one practice task (hard) and five actual tasks (3 easy and 2 hard). Clearly, on all task instances, participants that are determined as "high skill" outperformed those participants that are determined as "low skill." In other words, participants' scores on the two hard task instances are highly correlated with their scores on any single task instance, regardless of whether it is easy or hard, which suggests that it is valid to use participants' scores on the two hard tasks to measure skill levels.

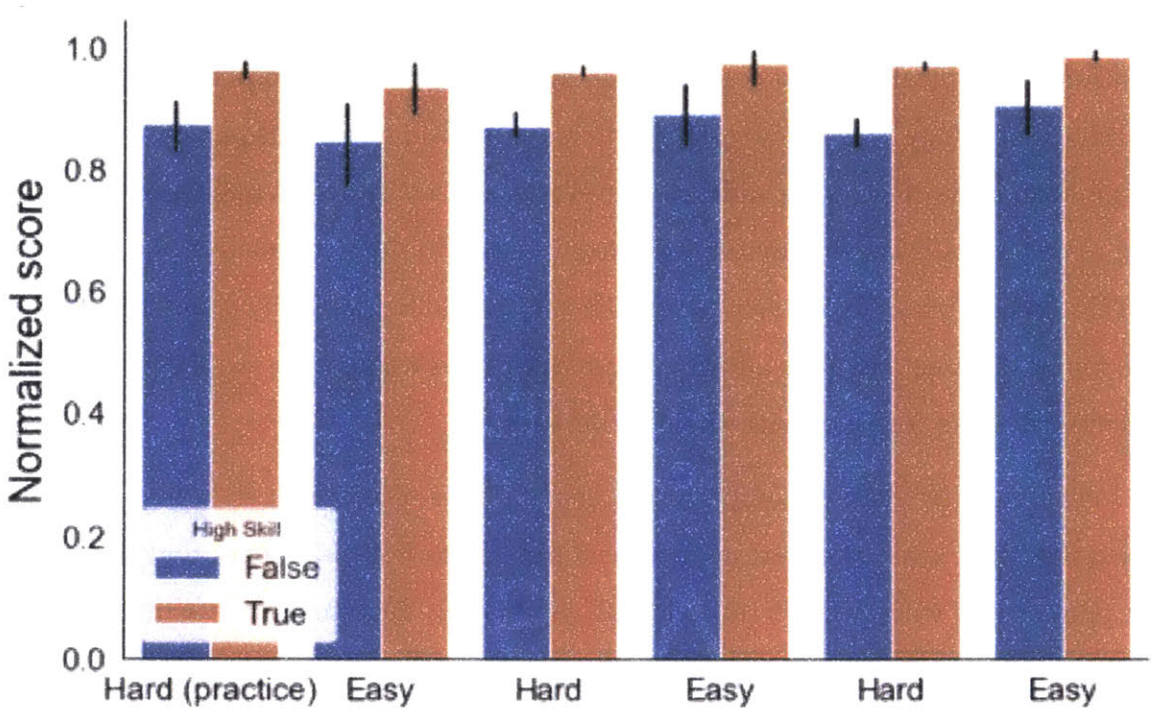


Figure A-5: Participants who obtained a higher score on the two hard tasks in the phase one experiment (i.e., "high skill") outperformed participants who obtained a lower score on those two hard tasks (i.e., "low skill") on each single task instance. Error bars represent 95% confidence intervals.

A.5 Comparing participants in phase one and two

One natural concern regarding the two-phase experimental design is whether different participants' experience in the phase one experiment will lead to a varying tendency to participating in the phase two experiment, implying potential self-selection that may result in biased experimental results. To examine whether self-selection bias would be a substantial concern, we first conducted a pilot study, in which 42 participants (these participants were not allowed to participate in the actual study) were recruited from Amazon Mechanical Turk to complete the first version of our two-phase experiment. In this pilot study, we asked each participant to complete a sequence of 5 room assignment tasks of varying difficulty levels as well as 36 RME questions in phase one. Two hours later, we invited all participants who had completed phase one to join the second-phase experiment, in which they would be randomly grouped together into teams of three members and they were asked to solve another sequence of 5 room assignment tasks together with their teammates.

Figure A-6 (left panel) compares the distributions of participants who completed phase one (i.e., gray bars and curves) and phase two (i.e., blue bars and curves) of the pilot study, with respect to their skill levels (i.e., the cumulative score a worker got in the 5 room assignment tasks of the phase one experiment; top row) and their social perceptiveness levels (i.e., the number of RME questions a worker answered correctly in the phase one experiment; bottom row). Visually, it is clear that during the pilot study, participants who decided to take the phase two experiment had both higher skill levels and higher social perceptiveness levels, compared to the entire pool of participants who had completed the phase one experiment. In other words, the experimental design and procedure that we adopted during our pilot study indeed led to a degree of self-selection bias. To decrease the level of self-selection bias, we made three changes during our main experiment. First, we altered the mix of tasks that we included in the phase one experiment to 3 "easy" tasks and 2 "hard" tasks. We hypothesized that with a higher fraction of easy tasks in phase one, participants would have a higher perceived self-efficiency in the room assignment tasks, and thus more

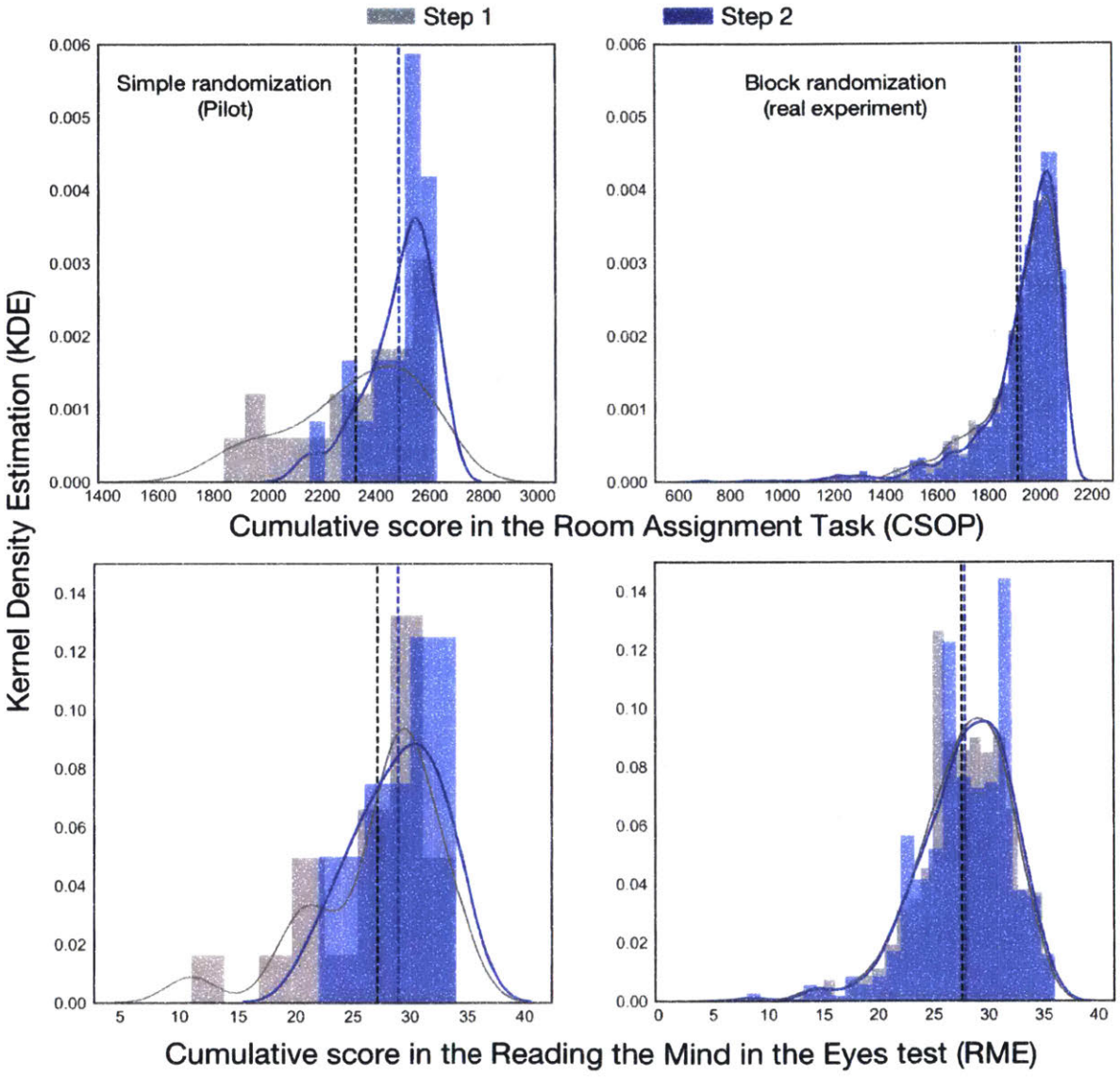
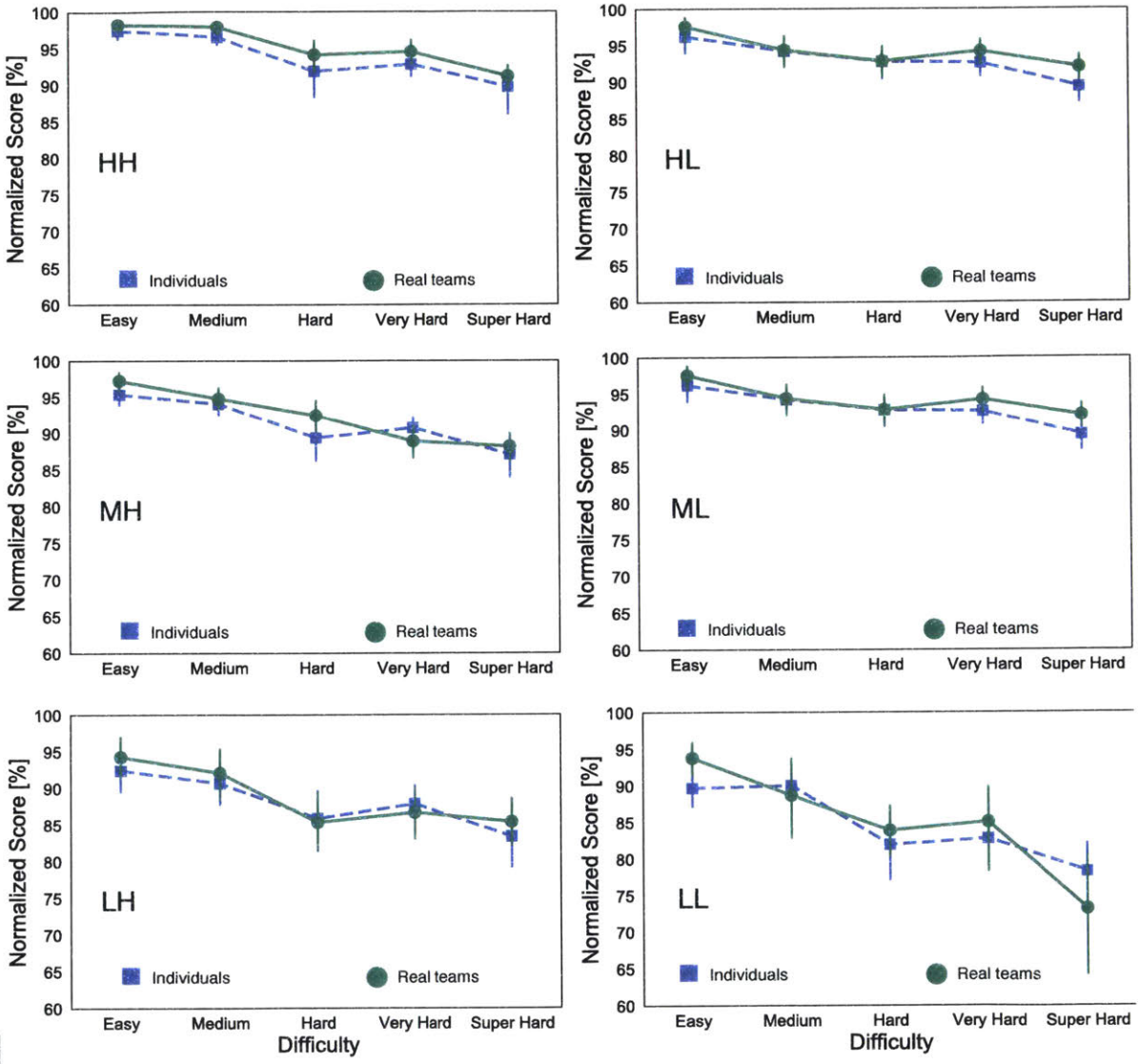


Figure A-6: Comparing the distributions of phase one participants and phase two participants with respect to their skill (i.e., scores obtained in room assignment tasks) and social perceptiveness levels (i.e., scores obtained in RME tests). Left: comparison results for the pilot study; Right: comparison results for the main experiment. Gaussian kernels are used for kernel density estimation.

likely to come back during phase two to complete more such tasks. Second, we adopted a block randomization scheme rather than a simple randomization scheme during our real experiment. Each block corresponded to a particular mixture of participants with high/mixed/low skill and high/low social perceptiveness (see Section 2.2.3 for more details), and we set the target number of workers to recruit at the block level. Doing so allowed us to effectively oversample the subgroups of participants who were potentially underrepresented in phase two, compared to the pool of participants in phase one (e.g., participants who had a lower skill and social perceptiveness levels)[1]. Finally, we extended the gap between the two phases of our experiment from two hours to six days, conjecturing that a longer gap would refresh participants' memory and potentially lead more of them to find it enjoyable to take similar types of tasks again in our phase two experiment. Figure A-6 (right panel) shows the distribution comparisons between participants who completed phase one and phase two of the real experiment. Here, we find there is no clear difference between the two groups of participants in terms of either their skill or their social perceptiveness. In other words, with the three changes that we made, we managed to minimize the self-selection biases between the two phases in our real experiment.

A.6 Performance as a function of the environment complexity

[1] As we mentioned in Section 2.2.3, another benefit brought up by the block randomization scheme is that we effectively oversampled less frequent combinations of workers (i.e., teams) even if self-selection bias was not present, such as teams with all three members being high on skill and social perceptiveness.



]\[h]

Figure A-7: Varying the room assignment task difficulty vs normalized score. The five task difficulty levels were characterized by the different number of students to be assigned, the number of dorm rooms available, and the number of constraints. Data is analyzed separately for individuals and teams from each of the six blocks. Increasing the task difficulty reduces the normalized score for both individuals and teams of all skill levels and social perceptiveness. Error bars indicate 95% confidence intervals.

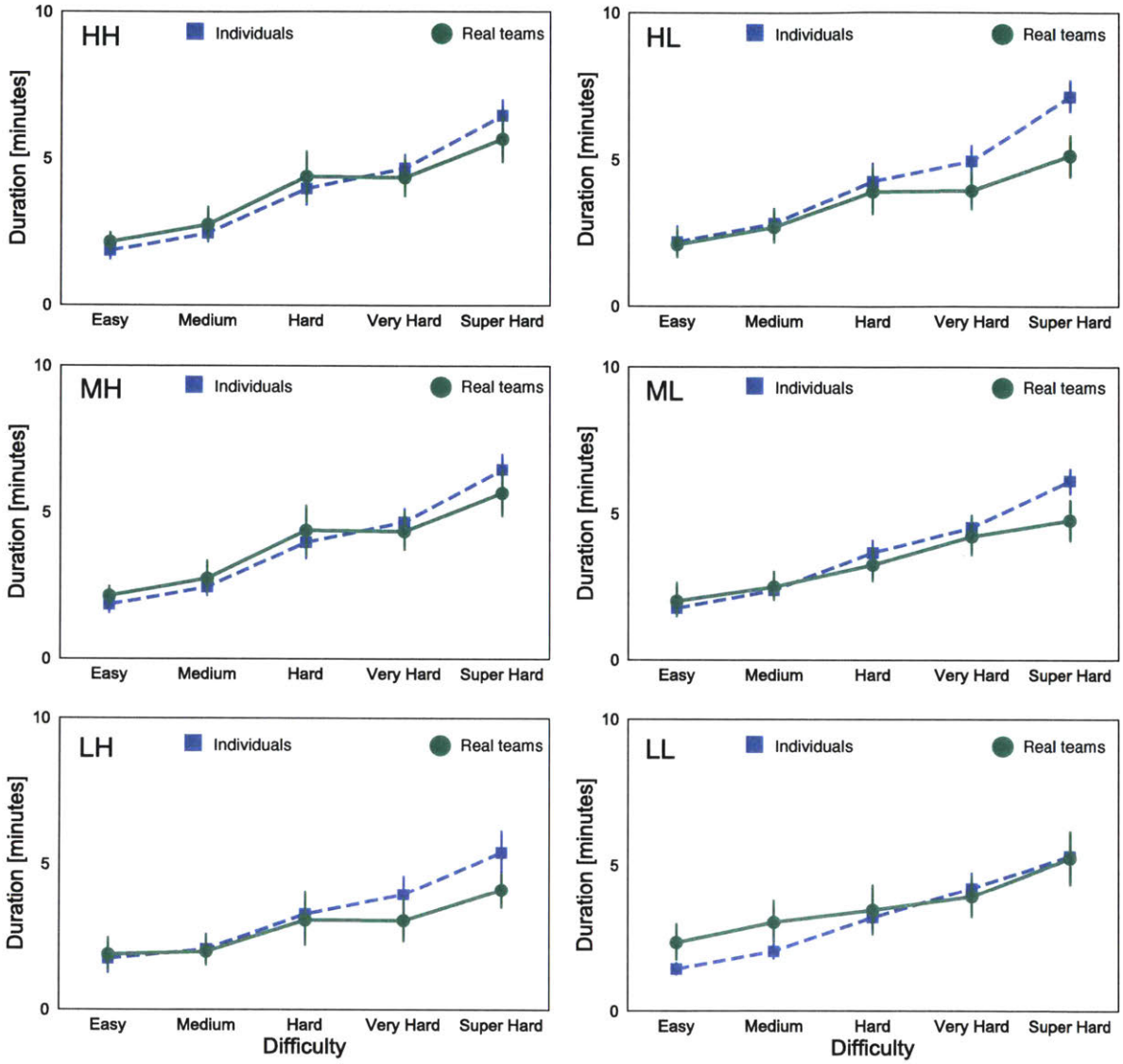


Figure A-8: Varying the room assignment task difficulty vs duration. The five task difficulty levels were characterized by the different number of students to be assigned, the number of dorm rooms available, and the number of constraints. Data is analyzed separately for individuals and teams from each of the six blocks. Increasing the task difficulty increases the time it takes participants to submit an assignment for both individuals and teams of all skill levels and social perceptiveness. Error bars indicate 95% confidence intervals.

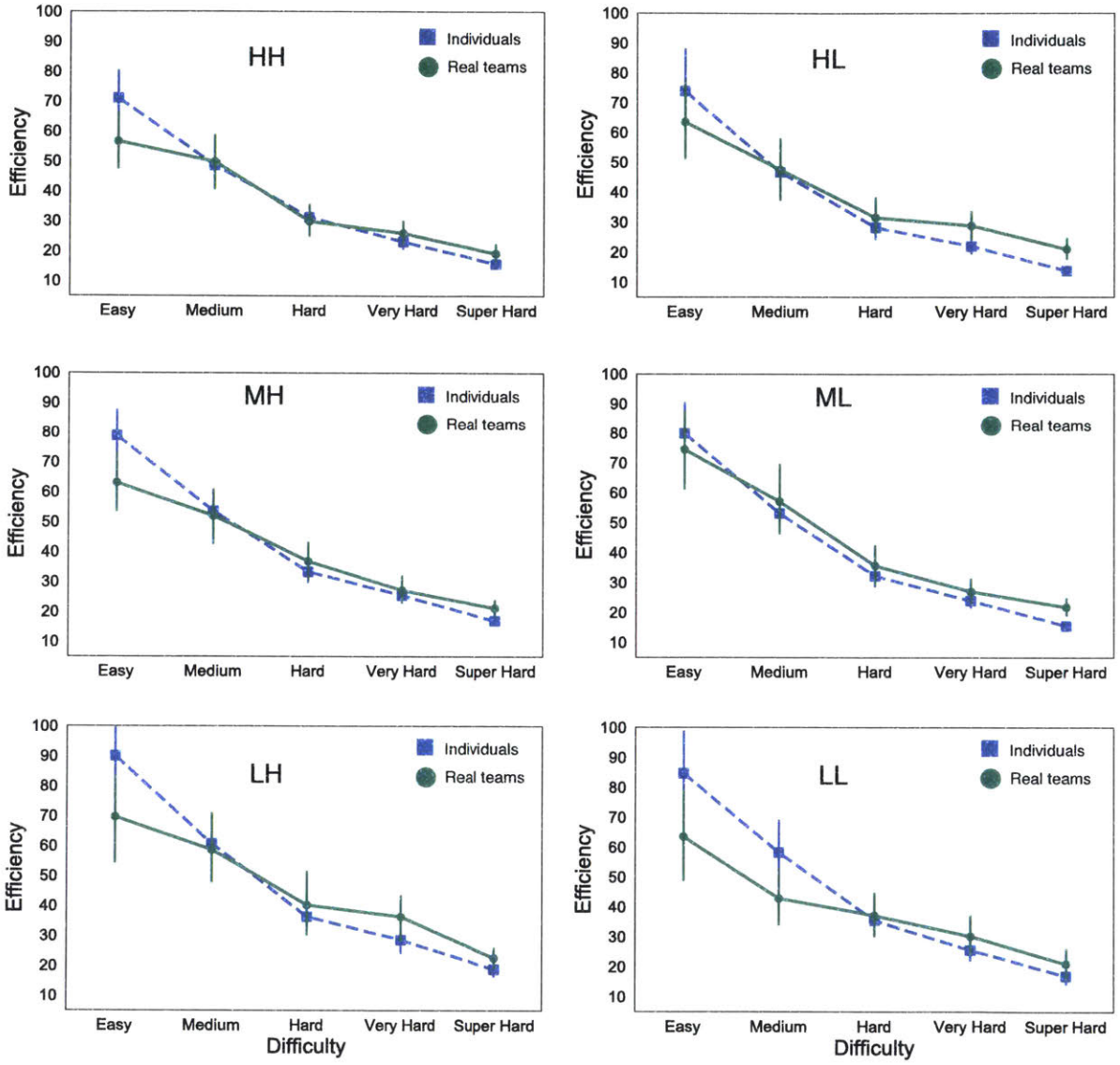


Figure A-9: Varying the room assignment task difficulty vs efficiency. The five task difficulty levels were characterized by the different number of students to be assigned, the number of dorm rooms available, and the number of constraints. Data is analyzed separately for individuals and teams from each of the six blocks. Increasing the task difficulty reduces the efficiency for both individuals and teams of all skill levels and social perceptiveness. Error bars indicate 95% confidence intervals.

A.7 Group composition; Supporting tables

Table A.1: The relation between the team's cognitive style diversity (in terms of whether all team members are fast/slow problem solvers or both types exist in the team) and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's cognitive style diversity. All models include random effects for teams as well as the team's skill level category as an intercept to account for dependence across tasks. Increasing a team's cognitive style diversity has no effect on the team's score, but reduces duration

	Score			Duration			Efficiency		
	β	CI (95%)	P	β	CI (95%)	P	β	CI (95%)	P
α	0.04	-0.38 - 0.47	0.843	-0.03	-0.14-0.08	0.578	0.05	-0.05 - 0.16	0.326
CogStyle. (Speed)	-0.02	-0.09 - 0.04	0.466	-0.15**	-0.26 - -0.04	0.007	0.14**	0.04 - 0.24	0.005
Random Effects									
σ^2	0.57			0.51			0.60		
τ_{00}	0.11 team_id			0.50 team_id			0.38 team_id		
	0.14 skill_type			0.00 skill_type			0.00 skill_type		
ICC	0.14 team_id			0.50 team_id			0.39 teame_id		
	0.17 skill_type			0.00 skill_type			0.00 skill_type		
N	980			980			980		
R^2	0.303			NA			0.400		

Table A.2: The relation between the team's cognitive style diversity (in terms of whether all team members have the same constraint violation tolerance or not) and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's cognitive style diversity. All models included random effects for teams as well as the team's skill level category as an intercept to account for dependence across tasks. Increasing a team's cognitive style diversity has no effect on the team's score, but reduces duration.

	Score			Duration			Efficiency		
	β	CI (95%)	P	β	CI (95%)	P	β	CI (95%)	P
α	0.04	-0.39 - 0.47	0.846	-0.03	-0.15 - 0.09	0.604	0.05	-0.08 - 0.18	0.421
CogStyle (tolerance)	0.01	-0.05 - 0.08	0.688	-0.12	-0.23 - -0.01	0.028	0.12	0.02 - 0.22	0.023
Random Effects									
σ^2	0.57			0.51			0.60		
τ_{00}	0.11 team_id			0.51 team_id			0.38 team_id		
	0.14 skill_type			0.00 skill_type			0.01 skill_type		
ICC	0.14 team_id			0.50 team_id			0.39 team_id		
	0.17 skill_type			0.00 skill_type			0.01 skill_type		
N	980			980			980		
R^2	0.305			0.508			0.401		

Table A.3: The relation between the team's cognitive style diversity (in terms of whether all team members are pragmatic/tenacious or both types exist in the team) and team performance. Data is combined across teams in all six blocks, and for all five tasks. Models relate performance measures (standardized within each task) with the team's cognitive style diversity. All models include random effects for teams as well as the team's skill level category as an intercept to account for dependence across tasks. Increasing a team's cognitive style diversity has no effect on the team's score and duration.

	Score			Duration			Efficiency		
	β	CI (95%)	P	β	CI (95%)	P	β	CI (95%)	P
α	0.04	-0.39 - 0.47	0.846	-0.03	-0.14 - 0.08	0.585	0.05	-0.07 - 0.17	0.379
CogStyle (conservative /progressive)	-0.02	-0.09 - 0.04	0.467	-0.04	-0.15 - 0.08	0.533	0.04	-0.06 - 0.14	0.473
Random Effects									
σ^2	0.57			0.51			0.60		
τ_{00}	0.11 game_id			0.52 game_id			0.40 game_id		
	0.14 skill_type			0.00 skill_type			0.00 skill_type		
ICC	0.13 game_id			0.51 game_id			0.40 game_id		
	0.17 skill_type			0.00 skill_type			0.00 skill_type		
N	980			980			980		
R^2	0.306			0.508			0.400		

A.8 Out of sample prediction accuracy

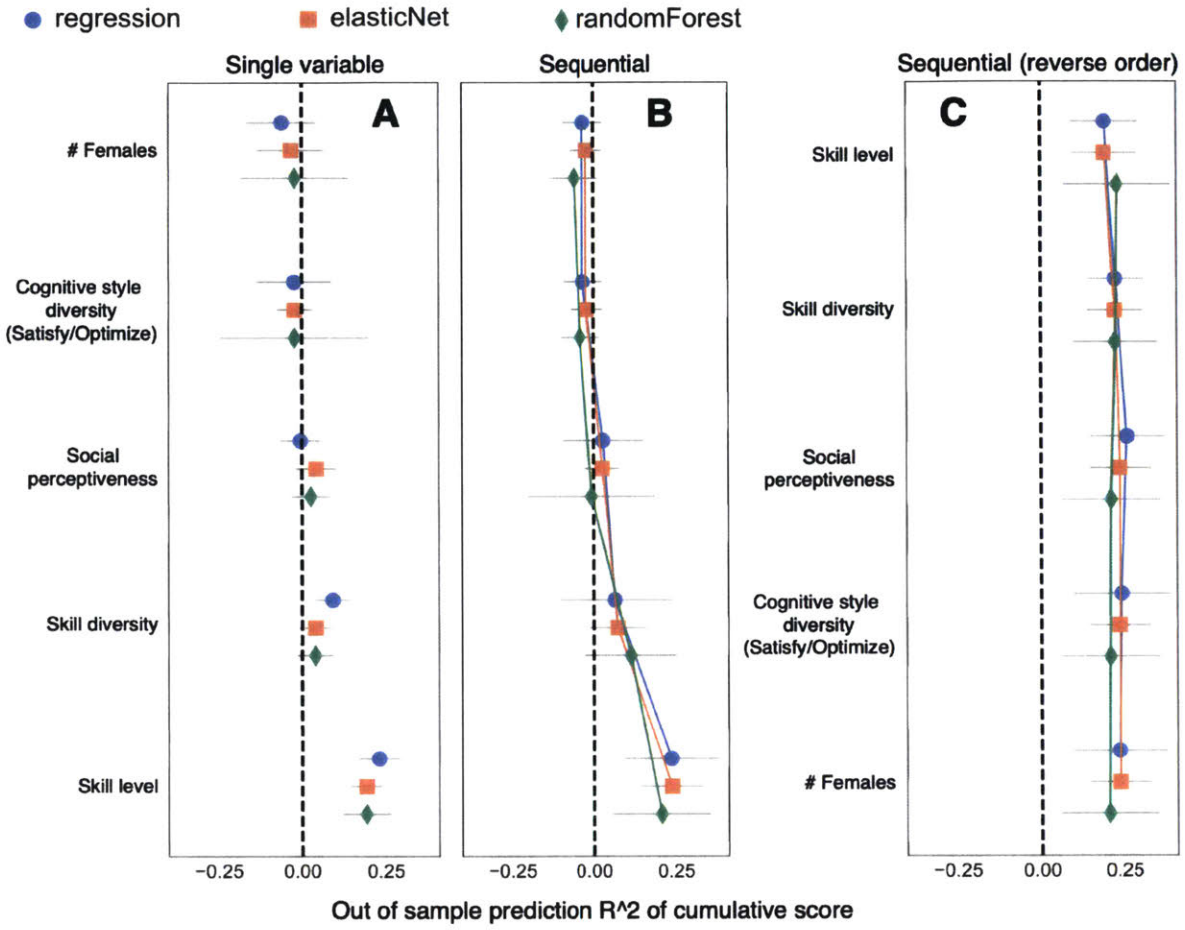


Figure A-10: Out of sample predictions on the team's cumulative score. Predict the team's normalized score with the team's skill level, skill diversity, social perceptiveness, cognitive style diversity, and the number of female team members. Three models (i.e., linear regression, elasticNet, and random forests) are used. Models are first learned on 70% of the teams and then tested on the rest 30% of the teams. This procedure is then repeated 5 times. Error bars indicate 95% confidence intervals. In all models, the majority of the explained variance in team's normalized score can be attributed to the team's skill level. .

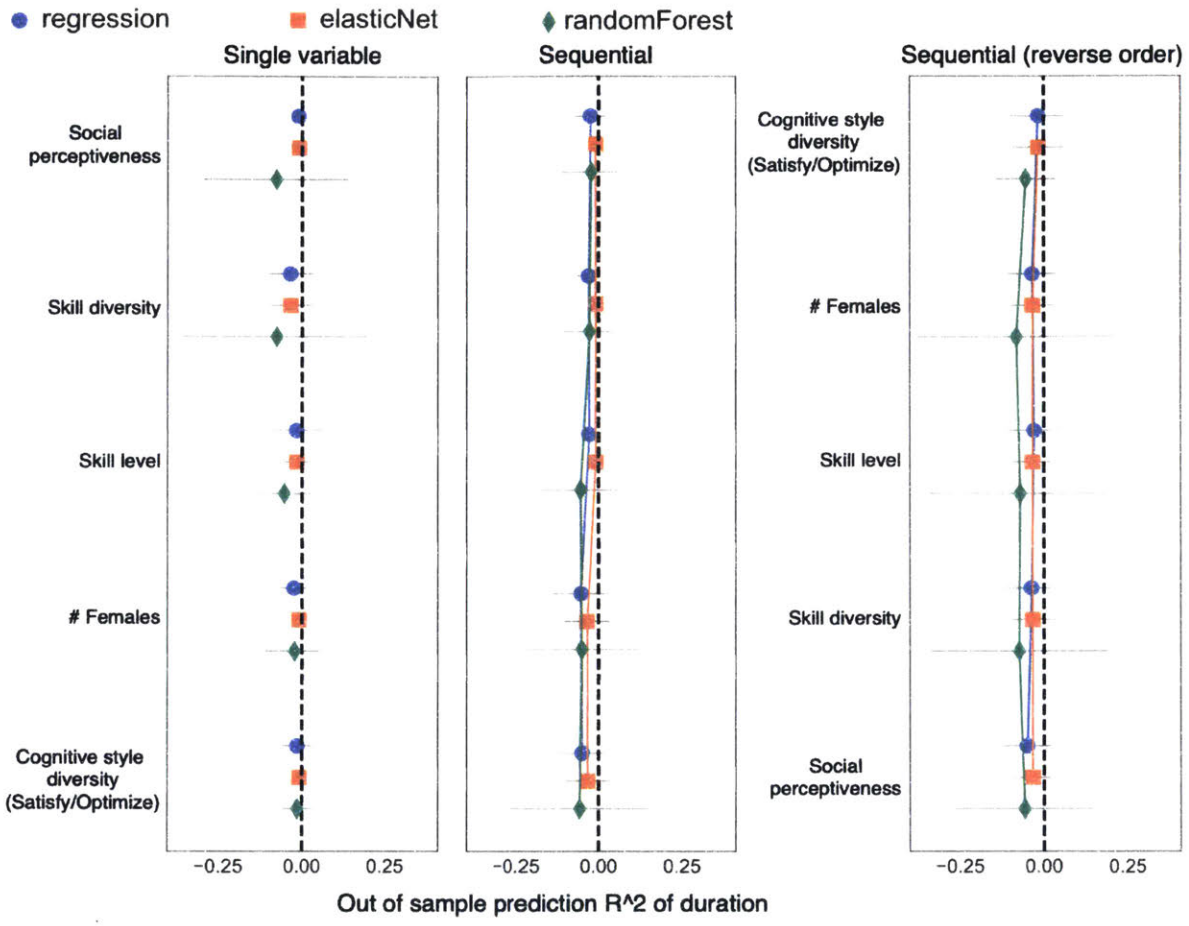


Figure A-11: Out of sample predictions on team's duration on tasks. Predict the team's duration on tasks with the team's skill level, skill diversity, social perceptiveness, cognitive style diversity, and the number of female team members. Three models (i.e., linear regression, elasticNet, and random forests) are used. Models are first learned on 70% of the teams and then tested on the rest 30% of the teams. This procedure is then repeated 5 times. Error bars indicate 95% confidence intervals. The set of independent variables can hardly be used to explain the variance in team's duration on tasks. .

Appendix B

Supplementary Information for Chapter 3

B.1 Guess the correlation game

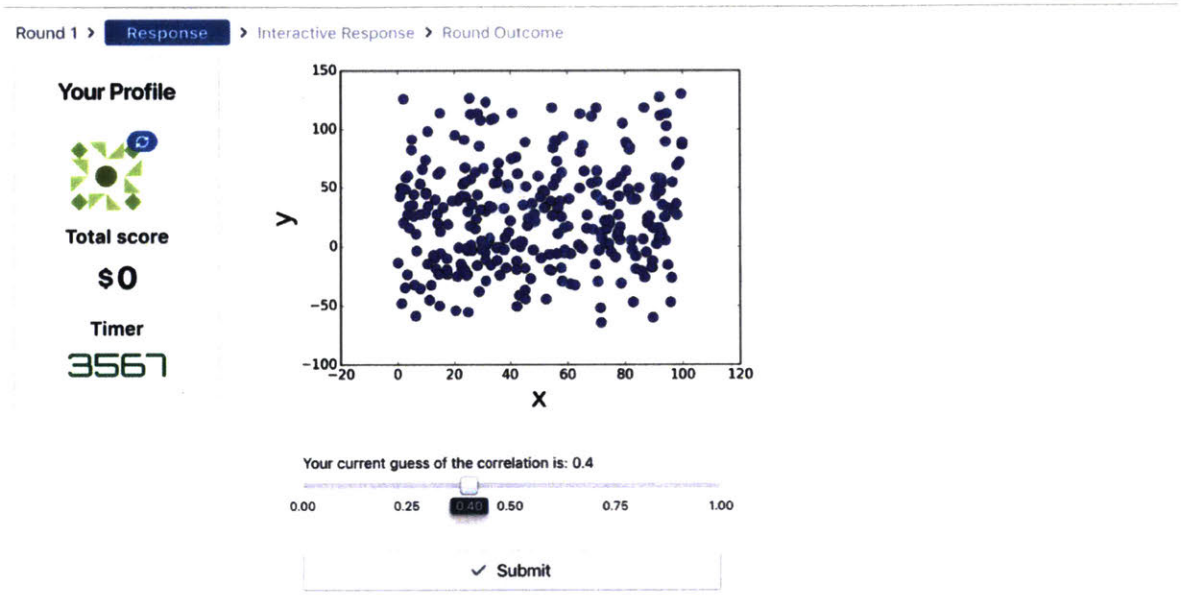


Figure B-1: Participants in all conditions make independent guesses about the correlation of two variables independently.

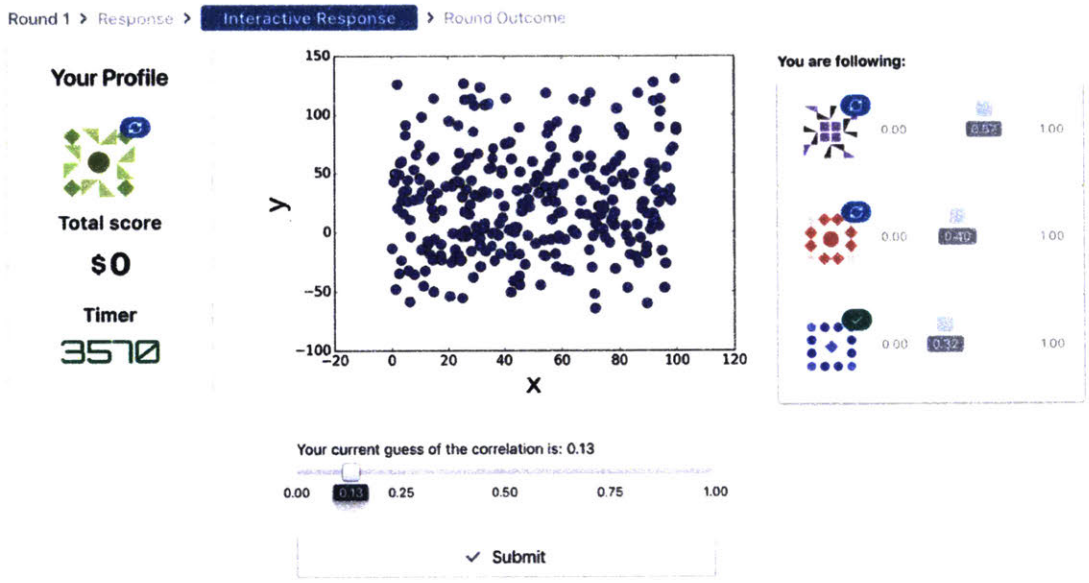


Figure B-2: Participants in the network condition engage in a an active social learning phase, where they are exposed to their ego-network's estimates in real time.

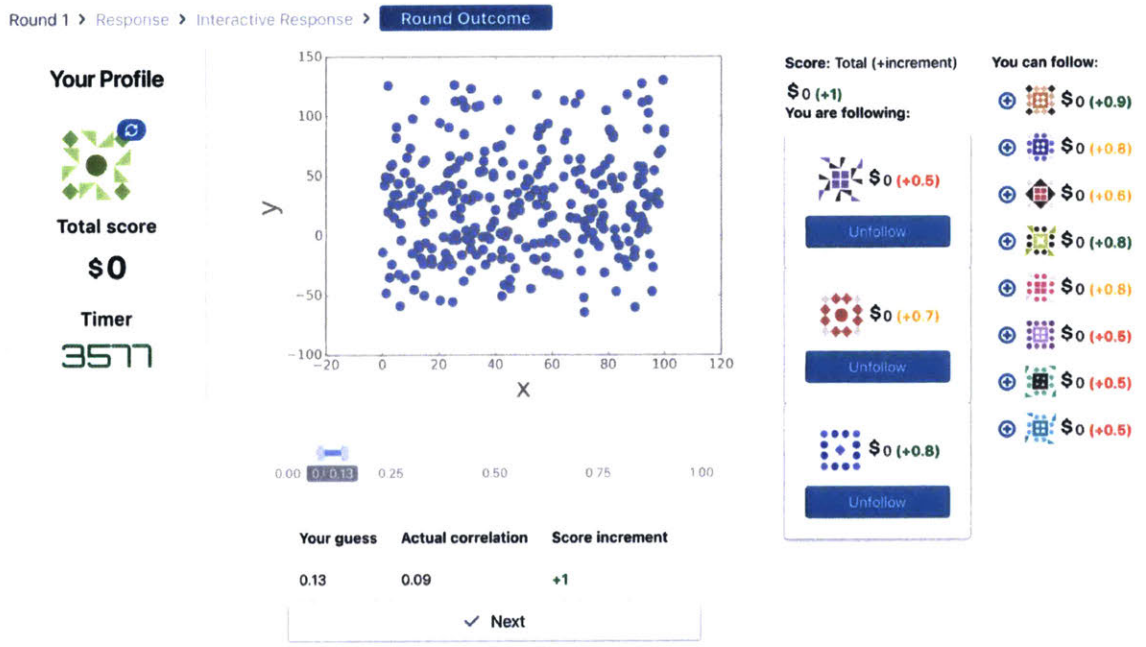


Figure B-3: After each task round, participants in the feedback conditions see the appropriate level of feedback for the conditions. This figure illustrates the dynamic network condition with full feedback (i.e., as opposed to no-feedback or only self-feedback). In all of our experiments, the maximum number of outgoing connections is three.

B.2 S1: Individual and collective error

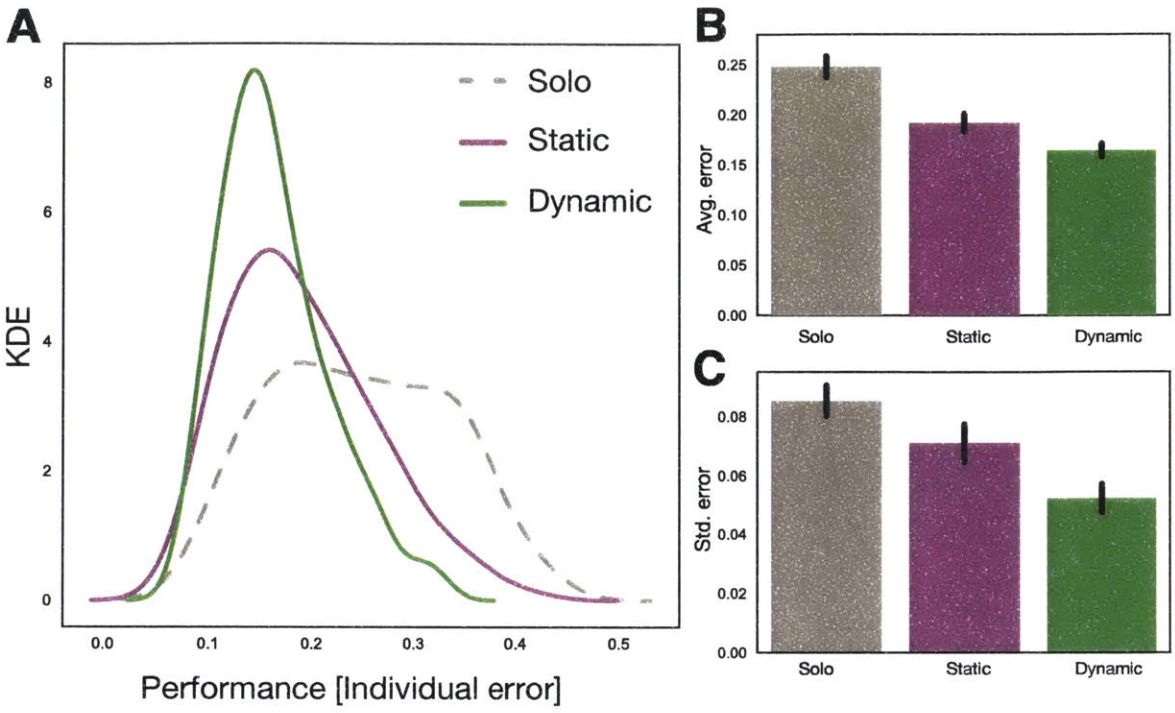


Figure B-4: Dynamic social influence benefits the performance of individuals in the crowd. (A) Kernel Density Estimate (KDE) of participants' individual performance (i.e., average error across all rounds) for the three experimental conditions. We find that participants in groups connected by dynamic influence networks (Dynamic condition) achieved 38% reduction in average error compared to participants in unconnected groups (Solo condition), and 12% reduction in average error compared to participants in groups connected by static influence networks (Static condition). Panel (B) compares the average performance of individuals across conditions. Two-sample t-tests show a significant difference between the average individual error of participants in the Solo and Static conditions ($P < 0.0001$), as well as between participants in the Static and Dynamic conditions ($P < 0.001$). Panel (C) compares the standard deviation of participant's individual performance across conditions, and shows that individual performance in groups connected by dynamic influence networks was, not only better on average, but also substantially more equal on its distribution among group members.

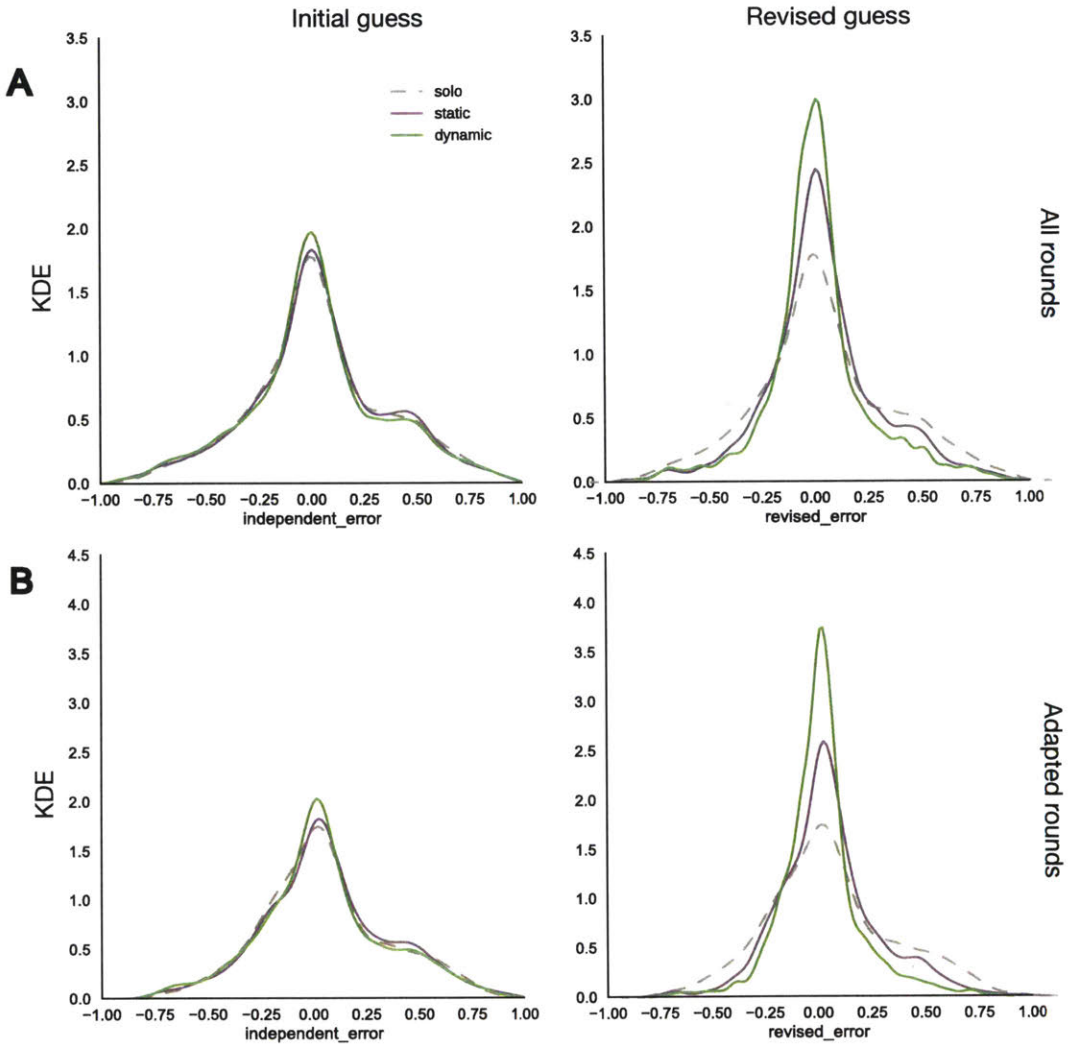


Figure B-5: Panel (A) shows individual errors in the full game and Panel (B) shows the error in the adapted period (i.e., periods [6-10] and [16-20]). The error for the initial guess in both panels is the same across conditions, however, the dynamic network condition incurs much lower errors in the adapted periods (as in Panel B).

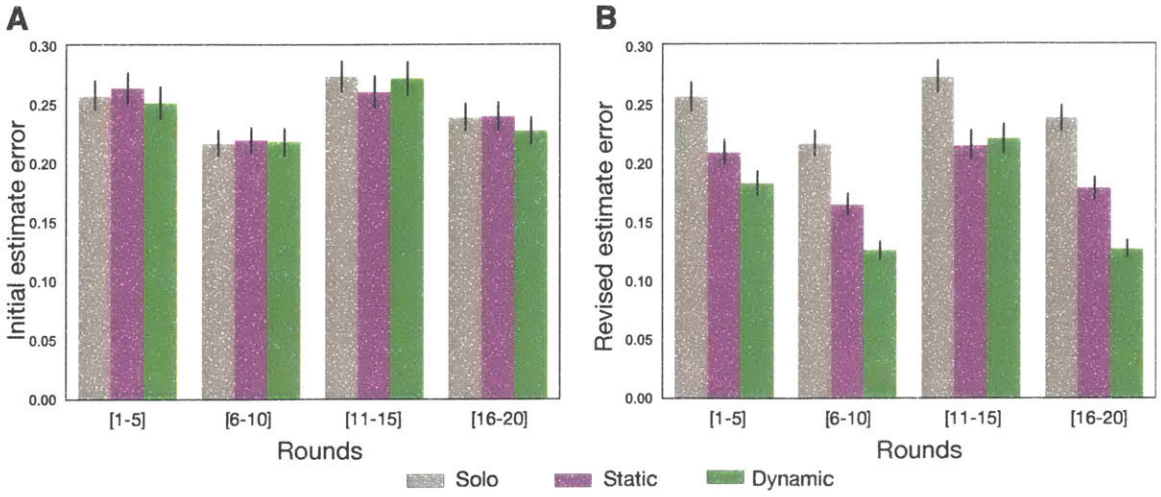


Figure B-6: Panel (A) shows the errors before the interactive estimation phase (i.e, pre-social learning). Panel (B) shows the errors after the participants revised their estimates in the static and dynamic network conditions (i.e., post-social learning).

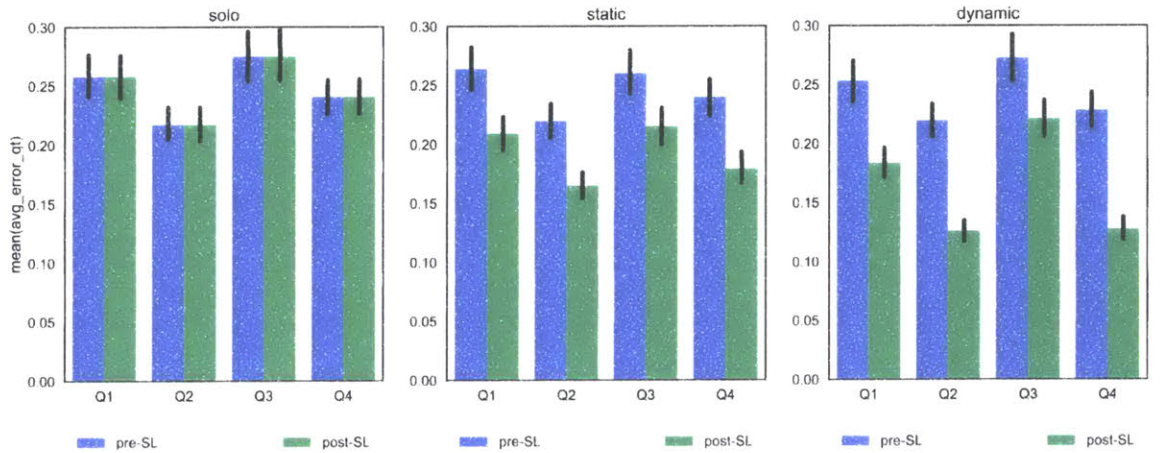


Figure B-7: The distribution of pre-social learning and post-social learning for the three conditions.

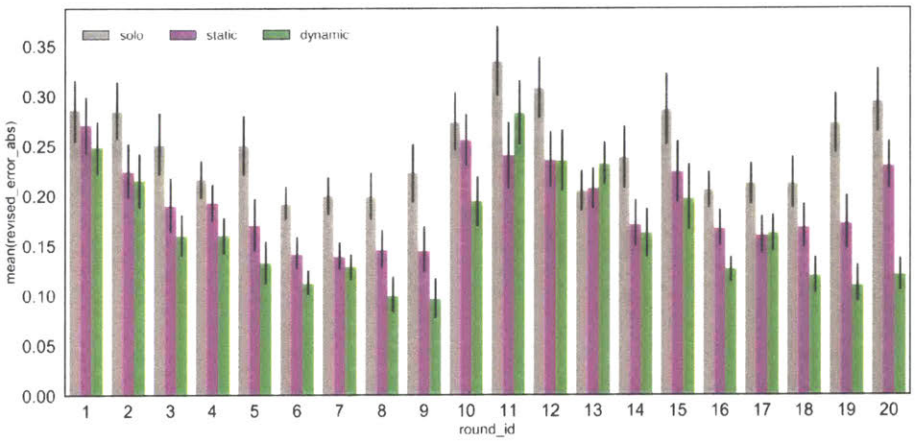


Figure B-8: Dynamic social influence effect in individual rounds: Adaptive with time and reduces individual error. All error rates are post-social learning errors.

Bibliography

[1] Daron Acemoglu, Angelia Nedic, and Asuman Ozdaglar. Convergence of rule-of-thumb learning rules in social networks. In *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, pages 1714–1720. IEEE, 2008.

[2] Ishani Aggarwal and Anita Williams Woolley. Do you see what i see? the effect of members' cognitive styles on team processes and errors in task execution. *Organizational Behavior and Human Decision Processes*, 122(1):92–99, 2013.

[3] Ishani Aggarwal and Anita Williams Woolley. Team creativity, cognition, and cognitive style diversity. *Management Science*, 2018.

[4] Faisal Aleissa, Riyadh Alnasser, Abdullah Almaatouq, Kamran Jamshaid, Fahad Alhasoun, Marta C González, and Anas Alfaris. Wired to connect: analyzing human communication and information sharing behavior during extreme events. In *KDD Workshop on Learning about Emergencies from Social Information*, 2014.

[5] Fahad Alhasoun, Abdullah Almaatouq, Kael Greco, Riccardo Campari, Anas Alfaris, and Carlo Ratti. The city browser: Utilizing massive call data to infer city mobility dynamics. In *3rd International Workshop on Urban Computing (UrbComp 2014). UrbComp: New York, NY*, 2014.

[6] Abdullah Almaatouq. Complex systems and a computational social science perspective on the labor market. Master's thesis, Massachusetts Institute of Technology, 2016.

[7] Abdullah Almaatouq, Ahmad Alabdulkareem, Mariam Nouh, Mansour Alsaleh, Abdulrahman Alarifi, Abel Sanchez, Anas Alfaris, and John Williams. A malicious activity detection system utilizing predictive modeling in complex environments. In *2014 IEEE 11th Consumer Communications and Networking Conference (CCNC)*, pages 371–379. IEEE, 2014.

[8] Abdullah Almaatouq, Ahmad Alabdulkareem, Mariam Nouh, Erez Shmueli, Mansour Alsaleh, Vivek K Singh, Abdulrahman Alarifi, Anas Alfaris, and Alex Sandy Pentland. Twitter: who gets caught? observed trends in social micro-blogging spam. In *Proceedings of the 2014 ACM conference on Web science*, pages 33–41. ACM, 2014.

[9] Abdullah Almaatouq, Peter Krafft, Yarrow Dunham, David G. Rand, and Alex Pentland. Turkers of the world unite: Multilevel in-group bias among crowdworkers on amazon mechanical turk. *Social Psychological and Personality Science*, 0(0):1948550619837002, 2019.

[10] Abdullah Almaatouq, Francisco Prieto-Castrillo, and Alex Pentland. Mobile communication signatures of unemployment. In *International conference on social informatics*, pages 407–418. Springer, 2016.

[11] Abdullah Almaatouq, Laura Radaelli, Alex Pentland, and Erez Shmueli. Are you your friends' friend? poor perception of friendship ties limits the ability to promote behavioral change. *PloS one*, 11(3):e0151588, 2016.

[12] Abdullah Almaatouq, Erez Shmueli, Mariam Nouh, Ahmad Alabdulkareem, Vivek K Singh, Mansour Alsaleh, Abdulrahman Alarifi, Anas Alfaris, et al. If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts. *International Journal of Information Security*, 15(5):475–491, 2016.

[13] Bedoor K AlShebli, Talal Rahwan, and Wei Lee Woon. The preeminence of ethnic diversity in scientific collaboration. *Nature communications*, 9(1):5163, 2018.

[14] Claes Andersson and Dwight Read. Group size and cultural complexity. *Nature*, 511(7507):E1, 2014.

[15] Sinan Aral. Commentary—identifying social influence: A comment on opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30(2):217–223, 2011.

[16] Sinan Aral and Marshall Van Alstyne. The diversity-bandwidth trade-off. *American Journal of Sociology*, 117(1):90–171, 2011.

[17] Antonio A. Arechar, Simon Gächter, and Lucas Molleman. Conducting interactive experiments online. *Experimental Economics*, pages 1–33, 2017.

[18] Bahador Bahrami, Karsten Olsen, Peter E Latham, Andreas Roepstorff, Geraint Rees, and Chris D Frith. Optimally interacting minds. *Science*, 329(5995):1081–1085, 2010.

[19] Stefano Balietti. nodegame: Real-time, synchronous, online experiments in the browser. *Behavior research methods*, 49(5):1696–1715, 2017.

[20] Albert Bandura. Social learning theory of aggression. *Journal of communication*, 28(3):12–29, 1978.

[21] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[22] Albert-László Barabási et al. *Network science*. Cambridge university press, 2016.

[23] Beth Baribault, Chris Donkin, Daniel R Little, Jennifer S Trueblood, Zita Oravecz, Don Van Ravenzwaaij, Corey N White, Paul De Boeck, and Joachim Vandekerckhove. Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, 115(11):2607–2612, 2018.

[24] Daniel Barkoczi and Mirta Galesic. Social learning strategies modify the effect of network structure on group performance. *Nature communications*, 7:13109, 2016.

[25] Simon Baron-Cohen, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. The âĂIJreading the mind in the eyesâĂİ 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):241–251, 2001.

[26] Joshua Becker, Devon Brackbill, and Damon Centola. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences*, 114(26):E5070–E5076, 2017.

[27] Suzanne T Bell. Deep-level composition variables as predictors of team performance: a meta-analysis. *Journal of applied psychology*, 92(3):595, 2007.

[28] Adam J Berinsky, Gregory A Huber, and Gabriel S Lenz. Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. *Political Analysis*, 20(3):351–368, 2012.

[29] Silvia Bonaccio and Reeshad S Dalal. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, 101(2):127–151, 2006.

[30] Léon Bottou. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

[31] Robert Boyd and Jeffrey P Lorberbaum. No pure strategy is evolutionarily stable in the repeated prisoner's dilemma game. *Nature*, 327(6117):58, 1987.

[32] Robert Boyd, Peter J Richerson, and Joseph Henrich. The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, 108(Supplement 2):10918–10925, 2011.

[33] Daren C Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, 14(1):75–90, 2008.

[34] Egon Brunswik. Representative design and probabilistic theory in a functional psychology. *Psychological review*, 62(3):193, 1955.

[35] Egon Brunswik. *Perception and the representative design of psychological experiments.* Univ of California Press, 1956.

[36] David V Budescu and Eva Chen. Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2):267–280, 2014.

[37] Marc W Cadotte and Caroline M Tucker. Should environmental filtering be abandoned? *Trends in ecology & evolution*, 32(6):429–437, 2017.

[38] Alessandra Cassar. Coordination and cooperation in local, random and small world networks: Experimental evidence. *Games and Economic Behavior*, 58(2):209–230, 2007.

[39] Simone Cenci and Serguei Saavedra. Structural stability of nonlinear population dynamics. *Physical Review E*, 97(1):012401, 2018.

[40] Simone Cenci, Chuliang Song, and Serguei Saavedra. Rethinking the importance of the structure of ecological networks under an environment-dependent framework. *Ecology and evolution*, 8(14):6852–6859, 2018.

[41] Scott A Chamberlain, Judith L Bronstein, and Jennifer A Rudgers. How context dependent are species interactions? *Ecology letters*, 17(7):881–890, 2014.

[42] Jesse Chandler, Gabriele Paolacci, Eyal Peer, Pam Mueller, and Kate A Ratliff. Using nonnaive participants can reduce effect sizes. *Psychological science*, 26(7):1131–1139, 2015.

[43] Arun G Chandrasekhar, Horacio Larreguy, and Juan Pablo Xandri. Testing models of social learning on networks: Evidence from a framed field experiment. *Work. Pap., Mass. Inst. Technol., Cambridge, MA*, 2012.

[44] Daniel L Chen, Martin Schonger, and Chris Wickens. otreeâĂŤan open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016.

[45] Thomas Chesney, Swee-Hoon Chuah, and Robert Hoffmann. Virtual world experimentation: An exploratory study. *Journal of Economic Behavior & Organization*, 72(1):618–635, 2009.

[46] Wing S Chow and Lai Sheung Chan. Social network, social trust and shared goals in organizational knowledge sharing. *Information & management*, 45(7):458–465, 2008.

[47] Kevin A Clarke and David M Primo. Overcoming 'physics envy'. *New York Times*, 30, 2012.

[48] Marcus Credé and Garett Howardson. The structure of group task performanceâĂŤa second look at "collective intelligence": Comment on woolley et al.(2010). 2017.

[49] Clintin P Davis-Stober, David V Budescu, Jason Dana, and Stephen B Broomell. When is a crowd wise? *Decision*, 1(2):79, 2014.

[50] Robyn M Dawes. Behavioral decision making and judgment. 1998.

[51] Joshua R De Leeuw. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47(1):1–12, 2015.

[52] Yves-Alexandre De Montjoye, Laura Radaelli, Vivek Kumar Singh, et al. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015.

[53] Stephanie de Oliveira and Richard E Nisbett. Demographically diverse crowds are typically not much wiser than homogeneous crowds. *Proceedings of the National Academy of Sciences*, 115(9):2066–2071, 2018.

[54] Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[55] Peter M DeMarzo, Jeffrey Zwiebel, and Dimitri Vayanos. Persuasion bias, social influence, and uni-dimensional opinions. *Social Influence, and Uni-Dimensional Opinions (November 2001). MIT Sloan Working Paper*, (4339-01), 2001.

[56] Maxime Derex, Marie-Pauline Beugin, Bernard Godelle, and Michel Raymond. Experimental evidence for the influence of group size on cultural complexity. *Nature*, 503(7476):389, 2013.

[57] Maxime Derex and Robert Boyd. The foundations of the human cultural niche. *Nature communications*, 6:8398, 2015.

[58] Maxime Derex and Robert Boyd. Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences*, 113(11):2982–2987, 2016.

[59] Mukund Deshpande and George Karypis. Selective markov models for predicting web page accesses. *ACM transactions on internet technology (TOIT)*, 4(2):163–184, 2004.

[60] Dennis J Devine and Jennifer L Philips. Do smarter teams do better: A meta-analysis of cognitive ability and team performance. *Small Group Research*, 32(5):507–532, 2001.

[61] Mandeep K Dhami, Ralph Hertwig, and Ulrich Hoffrage. The role of representative design in an ecological approach to cognition. *Psychological bulletin*, 130(6):959, 2004.

[62] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh acm international conference on web search and data mining*, pages 135–143. ACM, 2018.

[63] Lex Donaldson. *The contingency theory of organizations.* Sage, 2001.

[64] RI Dunbar. The social brain hypothesis. *brain*, 9(10):178–190, 1998.

[65] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36):15274–15278, 2009.

[66] Alice H Eagly. When passionate advocates meet research on diversity, does the honest broker stand a chance? *Journal of Social Issues*, 72(1):199–222, 2016.

[67] Bryan D Edwards, Eric Anthony Day, Winfred Arthur Jr, and Suzanne T Bell. Relationships among team ability composition, team mental models, and team performance. *Journal of Applied Psychology*, 91(3):727, 2006.

[68] Naomi Ellemers and Floor Rink. Diversity in work groups. *Current Opinion in Psychology*, 11:49–53, 2016.

[69] David Engel, Anita Williams Woolley, Lisa X Jing, Christopher F Chabris, and Thomas W Malone. 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, 2014.

[70] Michael A Erskine and Dawn G Gregg. Utilizing volunteered geographic information to develop a real-time disaster mapping tool: A prototype and research framework. In *CONF-IRM*, page 27, 2012.

[71] Douglas H Erwin and Eric H Davidson. The evolution of hierarchical gene regulatory networks. *Nature Reviews Genetics*, 10(2):141–148, 2009.

[72] Ann E Feyerherm and Cheryl L Rice. Emotional intelligence and team performance: The good, the bad and the ugly. *The International Journal of Organizational Analysis*, 10(4):343–362, 2002.

[73] Urs Fischbacher. z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178, 2007.

[74] James H Fowler and Nicholas A Christakis. Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, 107(12):5334–5338, 2010.

[75] Bruno S Frey and Stephan Meier. Social comparisons and pro-social behavior: Testing" conditional cooperation" in a field experiment. *The American Economic Review*, 94(5):1717–1722, 2004.

[76] Daniel Friedman and Ryan Oprea. A continuous dilemma. *American Economic Review*, 102(1):337–63, 2012.

[77] Edoardo Gallo and Chang Yan. The effects of reputational and social knowledge on cooperation. *Proceedings of the National Academy of Sciences*, page 201415883, 2015.

[78] Francis Galton. Vox populi (the wisdom of crowds). *Nature*, 75:450–451, 1907.

[79] Khaled Ghédira and Bernard Dubuisson. Constraint satisfaction and optimization problems. *Constraint Satisfaction Problems*, pages 165–180, 2013.

[80] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[81] Daniel G Goldstein and Gerd Gigerenzer. Models of ecological rationality: the recognition heuristic. *Psychological review*, 109(1):75, 2002.

[82] Benjamin Golub and Matthew O Jackson. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010.

[83] Diego A Gutnisky and Valentin Dragoi. Adaptive coding of visual information in neural populations. *Nature*, 452(7184):220–224, 2008.

[84] Kenneth R Hammond. Probabilistic functionalism: Egon brunswik's integration of the history, theory, and method of psychology. *The psychology of Egon Brunswik*, 15:80, 1966.

[85] Kenneth R Hammond. Representative design. *Retrieved October*, 9:1998, 1998.

[86] Morten T Hansen. The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative science quarterly*, 44(1):82–111, 1999.

[87] Andrew Hargadon and Robert I Sutton. Technology brokering and innovation in a product development firm. *Administrative science quarterly*, pages 716–749, 1997.

[88] Robert XD Hawkins. Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, 47(4):966–976, 2015.

[89] Joseph Henrich. *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press, 2015.

[90] Joseph Henrich, Maciej Chudek, and Robert Boyd. The big man mechanism: how prestige fosters cooperation and creates prosocial leaders. *Phil. Trans. R. Soc. B*, 370(1683):20150013, 2015.

[91] Joseph Henrich, Steven J Heine, and Ara Norenzayan. Beyond weird: Towards a broad-based behavioral science. *Behavioral and Brain Sciences*, 33(2-3):111–135, 2010.

[92] Joseph Henrich, Steven J Heine, and Ara Norenzayan. Most people are not weird. *Nature*, 466(7302):29, 2010.

[93] Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

[94] César A Hidalgo. Disconnected, fragmented, or united? a trans-disciplinary review of network science. *Applied Network Science*, 1(1):6, 2016.

[95] Yu-Chi Ho and David L Pepyne. Simple explanation of the no-free-lunch theorem and its implications. *Journal of optimization theory and applications*, 115(3):549–570, 2002.

[96] Jake M Hofman, Amit Sharma, and Duncan J Watts. Prediction and explanation in social systems. *Science*, 355(6324):486–488, 2017.

[97] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[98] Lu Hong and Scott E Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.

[99] John J Horton, David G Rand, and Richard J Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14(3):399–425, 2011.

[100] Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[101] Ute R Hülsheger, Neil Anderson, and Jesus F Salgado. Team-level predictors of innovation at work: a comprehensive meta-analysis spanning three decades of research. *Journal of Applied psychology*, 94(5):1128, 2009.

[102] Christian Igel and Marc Toussaint. A no-free-lunch theorem for non-uniform distributions of target functions. *Journal of Mathematical Modelling and Algorithms*, 3(4):313–322, 2005.

[103] Bertrand Jayles, Hye-rin Kim, Ramón Escobedo, Stéphane Cezera, Adrien Blanchet, Tatsuya Kameda, Clément Sire, and Guy Theraulaz. How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences*, 114(47):12620–12625, 2017.

[104] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the*

2nd ACM SIGKDD international workshop on Urban Computing, page 2. ACM, 2013.

[105] Márton Karsai, Mikko Kivelä, Raj Kumar Pan, Kimmo Kaski, János Kertész, A-L Barabási, and Jari Saramäki. Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2):025102, 2011.

[106] Michael Kearns, Stephen Judd, Jinsong Tan, and Jennifer Wortman. Behavioral experiments on biased voting in networks. *Proceedings of the National Academy of Sciences*, 106(5):1347–1352, 2009.

[107] Michael Kearns, Stephen Judd, and Yevgeniy Vorobeychik. Behavioral experiments on a network formation game. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 690–704. ACM, 2012.

[108] Michael Kearns, Siddharth Suri, and Nick Montfort. An experimental study of the coloring problem on human subject networks. *Science*, 313(5788):824–827, 2006.

[109] Norbert L Kerr and R Scott Tindale. Group performance and decision making. *Annu. Rev. Psychol.*, 55:623–655, 2004.

[110] Young Ji Kim, David Engel, Anita Williams Woolley, Jeffrey Yu-Ting Lin, Naomi McArthur, and Thomas W Malone. 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*, pages 2316–2329. ACM, 2017.

[111] Alexander Kindel, Vineet Bansal, Kristin Catena, Thomas Hartshorne, Kate Jaeger, Dawn Koffman, Sara McLanahan, Maya Phillips, Shiva Rouhani, Ryan Vinh, et al. Improving metadata infrastructure for complex surveys: Insights from the fragile families challenge. 2018.

[112] Richard Klein, Kate Ratliff, Michelangelo Vianello, Reginald Adams Jr, Stěpán Bahník, Michael Bernstein, Konrad Bocian, Mark Brandt, Beach Brooks, Claudia Brumbaugh, et al. Data from investigating variation in replicability: A âĂIJmany labsâĂİ replication project. *Journal of Open Psychology Data*, 2(1), 2014.

[113] Michio Kondoh. Foraging adaptation and the relationship between food-web complexity and stability. *Science*, 299(5611):1388–1391, 2003.

[114] Asher Koriat. When are two heads better than one and why? *Science*, 336(6079):360–362, 2012.

[115] Christina N Lacerenza, Shannon L Marlow, Scott I Tannenbaum, and Eduardo Salas. Team development interventions: Evidence-based approaches for improving teamwork. *American Psychologist*, 73(4):517, 2018.

[116] David Lazer and Allan Friedman. The network structure of exploration and exploitation. *Administrative Science Quarterly*, 52(4):667–694, 2007.

[117] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.

[118] John M Levine, Richard L Moreland, and Leslie RM Hausmann. Managing group composition: Inclusive and exclusive role transitions. In *Social psychology of inclusion and exclusion*, pages 155–178. Psychology Press, 2004.

[119] Aming Li, Sean P Cornelius, Y-Y Liu, Long Wang, and A-L Barabási. The fundamental advantages of temporal networks. *Science*, 358(6366):1042–1046, 2017.

[120] Michael P Lillis. Emotional intelligence, diversity, and group performance: The effect of team composition on executive education program outcomes. *Journal of Executive Education*, 6(1):4, 2013.

[121] David Liu and Matthew Salganik. Successes and struggles with computational reproducibility: Lessons from the fragile families challenge. 2019.

[122] Adeline Lo, Herman Chernoff, Tian Zheng, and Shaw-Hwa Lo. Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences*, 112(45):13892–13897, 2015.

[123] Andrew W Lo. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. 2004.

[124] Andrew W Lo. Reconciling efficient markets with behavioral finance: the adaptive markets hypothesis. *Journal of investment consulting*, 7(2):21–44, 2005.

[125] Andrew W Lo and Mark T Mueller. Warning: physics envy may be hazardous to your wealth! *Available at SSRN 1563882*, 2010.

[126] Dmitrii Logofet. *Matrices and Graphs Stability Problems in Mathematical Ecology: 0*. CRC press, 2018.

[127] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025, 2011.

[128] Ian Lundberg, Arvind Narayanan, Karen Levy, and Matthew J Salganik. Privacy, ethics, and data access: A case study of the fragile families challenge. *arXiv preprint arXiv:1809.00103*, 2018.

[129] Gabriel Madirolas and Gonzalo G de Polavieja. Improving collective estimations using resistance to social influence. *PLoS Comput Biol*, 11(11):e1004594, 2015.

[130] Albert E Mannes, Jack B Soll, and Richard P Larrick. The wisdom of select crowds. *Journal of personality and social psychology*, 107(2):276, 2014.

[131] Andrew Mao, Yiling Chen, Krzysztof Z Gajos, David C Parkes, Ariel D Procaccia, and Haoqi Zhang. Turkserver: Enabling synchronous and longitudinal online experiments. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[132] Andrew Mao, Lili Dworkin, Siddharth Suri, and Duncan J Watts. Resilient cooperators stabilize long-run cooperation in the finitely repeated prisonerâĂŹs dilemma. *Nature communications*, 8:13800, 2017.

[133] Andrew Mao, Winter Mason, Siddharth Suri, and Duncan J Watts. An experimental study of team size and performance on a complex task. *PloS one*, 11(4):e0153048, 2016.

[134] Qiushi Mao. *Experimental studies of human behavior in social computing systems.* PhD thesis, 2015.

[135] Travis Martin, Jake M Hofman, Amit Sharma, Ashton Anderson, and Duncan J Watts. Exploring limits to prediction in complex social systems. In *Proceedings of the 25th International Conference on World Wide Web*, pages 683–694. International World Wide Web Conferences Steering Committee, 2016.

[136] Winter Mason and Duncan J Watts. Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769, 2012.

[137] Carolyn McAndrews and Justine Marcus. The politics of collective public participation in transportation decision-making. *Transportation Research Part A: Policy and Practice*, 78:537–550, 2015.

[138] Joseph Edward McGrath. *Groups: Interaction and performance*, volume 14. Prentice-Hall Englewood Cliffs, NJ, 1984.

[139] Mark E McKnight and Nicholas A Christakis. Breadboard: Software for online social experiments, 2016.

[140] Edgar C Merkle, Mark Steyvers, Barbara Mellers, and Philip E Tetlock. A neglected dimension of good forecasting judgment: The questions we choose also matter. *International Journal of Forecasting*, 33(4):817–832, 2017.

[141] Robert King Merton and Robert C Merton. *Social theory and social structure.* Simon and Schuster, 1968.

[142] Mehdi Moussaïd, Alejandro Noriega Campero, and Abdullah Almaatouq. Dynamical networks of influence in small group discussions. *PloS one*, 13(1):e0190541, 2018.

[143] Mehdi Moussaïd, Stefan M Herzog, Juliane E Kämmer, and Ralph Hertwig. Reach and speed of judgment propagation in the laboratory. *Proceedings of the National Academy of Sciences*, page 201611998, 2017.

[144] Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

[145] Satyam Mukherjee, Yun Huang, Julia Neidhardt, Brian Uzzi, and Noshir Contractor. Prior shared success predicts victory in team competitions. *Nature Human Behaviour*, 3(1):74, 2019.

[146] Kevin Patrick Murphy. *Dynamic bayesian networks*. PhD thesis, University of California, Berkeley, 2002.

[147] Michael Muthukrishna and Joseph Henrich. Innovation in the collective brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1690):20150192, 2016.

[148] Joaquin Navajas, Tamara Niella, Gerry Garbulsky, Bahador Bahrami, and Mariano Sigman. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, page 1, 2018.

[149] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2):026118, 2001.

[150] Bernard A Nijstad, Michael Diehl, and Wolfgang Stroebe. Cognitive stimulation and interference in idea generating groups. *Group creativity: Innovation through collaboration*, pages 137–159, 2003.

[151] Alejandro Noriega-Campero, Abdullah Almaatouq, Peter Krafft, Abdulrahman Alotaibi, Mehdi Moussaid, and Alex Pentland. The wisdom of the network: How adaptive networks promote collective intelligence. *arXiv preprint arXiv:1805.04766*, 2018.

[152] Mariam Nouh, Abdullah Almaatouq, Ahmad Alabdulkareem, Vivek K Singh, Erez Shmueli, Mansour Alsaleh, Abdulrahman Alarifi, Anas Alfaris, et al. Social information leakage: Effects of awareness and peer pressure on user behavior. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 352–360. Springer, 2014.

[153] Martin A Nowak. Five rules for the evolution of cooperation. *science*, 314(5805):1560–1563, 2006.

[154] Lynn R Offermann, James R Bailey, Nicholas L Vasilopoulos, Craig Seal, and Mary Sass. The relative contribution of emotional competence and cognitive ability to individual and team performance. *Human performance*, 17(2):219–243, 2004.

[155] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences*, 104(18):7332–7336, 2007.

[156] Nicolas Paton and Abdullah Almaatouq. Empirica: Open-Source, Real-Time, Synchronous, Virtual Lab Framework, November 2018.

[157] Pablo Piedrahita, Javier Borge-Holthoefer, Yamir Moreno, and Sandra González-Bailón. The contagion effects of repeated activation in social networks. *Social Networks*, 54:326–335, 2018.

[158] Flávio L Pinheiro and Dominik Hartmann. Intermediate levels of network heterogeneity provide the best evolutionary outcomes. *Scientific reports*, 7(1):15242, 2017.

[159] Amy R Poteete and Elinor Ostrom. Heterogeneity, group size and collective action: The role of institutions in forest management. *Development and change*, 35(3):435–461, 2004.

[160] Dražen Prelec, H Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532, 2017.

[161] David G Rand, Samuel Arbesman, and Nicholas A Christakis. Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*, 108(48):19193–19198, 2011.

[162] David G Rand, Martin A Nowak, James H Fowler, and Nicholas A Christakis. Static network structure can stabilize human cooperation. *Proceedings of the National Academy of Sciences*, 111(48):17093–17098, 2014.

[163] David G Rand, Alexander Peysakhovich, Gordon T Kraft-Todd, George E Newman, Owen Wurzbacher, Martin A Nowak, and Joshua D Greene. Social heuristics shape intuitive cooperation. *Nature communications*, 5:3677, 2014.

[164] Ray Reagans and Bill McEvily. Network structure and knowledge transfer: The effects of cohesion and range. *Administrative science quarterly*, 48(2):240–267, 2003.

[165] Ray Reagans and Ezra W Zuckerman. Networks, diversity, and productivity: The social capital of corporate r&d teams. *Organization science*, 12(4):502–517, 2001.

[166] Christoph Riedl and Anita Williams Woolley. Teams vs. crowds: A field test of the relative contribution of incentives, member ability, and emergent collaboration to crowd-based problem solving performance. *Academy of Management Discoveries*, 3(4):382–403, 2017.

[167] Daniel E Rigobon, Eaman Jahani, Yoshihiko Suhara, Khaled AlGhoneim, Abdulaziz Alghunaim, Abdullah Almaatouq, et al. Winning models for gpa, grit, and layoff in the fragile families challenge. *arXiv preprint arXiv:1805.11557*, 2018.

[168] Simon Rodan and Charles Galunic. More than network structure: How knowledge heterogeneity influences managerial performance and innovativeness. *Strategic management journal*, 25(6):541–562, 2004.

[169] Rudolf P Rohr, Serguei Saavedra, Guadalupe Peralta, Carol M Frost, Louis-Félix Bersier, Jordi Bascompte, and Jason M Tylianakis. Persist or produce: a community trade-off tuned by species evenness. *The American Naturalist*, 188(4):411–422, 2016.

[170] Ariel Rubinstein. Finite automata play the repeated prisoner's dilemma. *Journal of economic theory*, 39(1):83–96, 1986.

[171] Serguei Saavedra, Rudolf P Rohr, Jordi Bascompte, Oscar Godoy, Nathan JB Kraft, and Jonathan M Levine. A structural approach for understanding multispecies coexistence. *Ecological Monographs*, 87(3):470–486, 2017.

[172] Serguei Saavedra, Rudolf P Rohr, Vasilis Dakos, and Jordi Bascompte. Estimating the tolerance of species to the effects of global environmental change. *Nature communications*, 4:2350, 2013.

[173] Serguei Saavedra, Rudolf P Rohr, Luis J Gilarranz, and Jordi Bascompte. How structurally stable are global socioeconomic systems? *Journal of the Royal Society Interface*, 11(100):20140693, 2014.

[174] Matthew J Salganik. *Bit by bit: social research in the digital age.* Princeton University Press, 2017.

[175] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

[176] Matthew J Salganik and Karen EC Levy. Wiki surveys: Open and quantifiable social data collection. *PloS one*, 10(5):e0123483, 2015.

[177] Takao Sasaki, Boris Granovskiy, Richard P Mann, David JT Sumpter, and Stephen C Pratt. Ant colonies outperform individuals when a sensory discrimination task is difficult but not when it is easy. *Proceedings of the National Academy of Sciences*, 110(34):13769–13773, 2013.

[178] Jonathan Schulz, Duman Bahrami-Rad, Jonathan Beauchamp, and Joseph Henrich. The origins of weird psychology. 2018.

[179] Patricia Shih, Mark Shen, Birgit Öttl, Brandon Keehn, Michael S Gaffrey, and Ralph-Axel Müller. Atypical network connectivity for imitation in autism spectrum disorder. *Neuropsychologia*, 48(10):2931–2939, 2010.

[180] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

[181] Herbert A Simon. Bounded rationality and organizational learning. *Organization science*, 2(1):125–134, 1991.

[182] Leo K Simon and Maxwell B Stinchcombe. Extensive form games in continuous time: Pure strategies. *Econometrica: Journal of the Econometric Society*, pages 1171–1214, 1989.

[183] Vernon L Smith. Constructivist and ecological rationality in economics. *American economic review*, 93(3):465–508, 2003.

[184] Stanislav Sobolevsky, Izabela Sitko, Remi Tachet Des Combes, Bartosz Hawelka, Juan Murillo Arias, and Carlo Ratti. Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in spain. In *2014 IEEE international congress on big data*, pages 136–143. IEEE, 2014.

[185] Chuliang Song, Rudolf P Rohr, and Serguei Saavedra. Why are some plant–pollinator networks more nested than others? *Journal of Animal Ecology*, 86(6):1417–1424, 2017.

[186] Ivan D Steiner. Group process and productivity (social psychological monograph). 2007.

[187] Greg L Stewart. A meta-analytic review of relationships between team design features and team performance. *Journal of management*, 32(1):29–55, 2006.

[188] Gijsbert Stoet. Psytoolkit: A software package for programming psychological experiments using linux. *Behavior Research Methods*, 42(4):1096–1104, 2010.

[189] Gijsbert Stoet. Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1):24–31, 2017.

[190] Wolfgang Stroebe and Michael Diehl. Why groups are less effective than their members: on productivity losses in idea-generating groups. *European review of social psychology*, 5(1):271–303, 1994.

[191] Nathalie Stroeymeyt, Anna V Grasse, Alessandro Crespi, Danielle P Mersch, Sylvia Cremer, and Laurent Keller. Social network plasticity decreases disease transmission in a eusocial insect. *Science*, 362(6417):941–945, 2018.

[192] Monic Sun. How does the variance of product ratings matter? *Management Science*, 58(4):696–707, 2012.

[193] Siddharth Suri and Duncan J Watts. Cooperation and contagion in web-based, networked public goods experiments. *PloS one*, 6(3):e16836, 2011.

[194] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[195] Jeannette Sutton, C Ben Gibson, Nolan Edward Phillips, Emma S Spiro, Cedar League, Britta Johnson, Sean M Fitzhugh, and Carter T Butts. A cross-hazard analysis of terse message retransmission on twitter. *Proceedings of the National Academy of Sciences*, 112(48):14793–14798, 2015.

[196] Nassim Nicholas Taleb. *Skin in the game: Hidden asymmetries in daily life*. Random House, 2018.

[197] Leigh L Thompson and Elizabeth Ruth Wilson. Creativity in teams. *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource*, pages 1–14, 2015.

[198] Peter M Todd and Gerd Gigerenzer. *Ecological rationality: Intelligence in the world*. OUP USA, 2012.

[199] Wataru Toyokawa, Andrew Whalen, and Kevin N Laland. Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour*, page 1, 2019.

[200] Edward Tsang. *Foundations of constraint satisfaction: the classic text*. BoD–Books on Demand, 2014.

[201] Yakov Zalmanovich Tsypkin and Zivorad Jezdimir Nikolic. *Adaptation and learning in automatic systems*, volume 73. Academic Press New York, 1971.

[202] Brian Uzzi. The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American sociological review*, pages 674–698, 1996.

[203] Brian Uzzi. Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative science quarterly*, pages 35–67, 1997.

[204] Brian Uzzi and Jarrett Spiro. Collaboration and creativity: The small world problem. *American journal of sociology*, 111(2):447–504, 2005.

[205] Sebastián Valenzuela, Namsu Park, and Kerk F Kee. Is there social capital in a social network site?: Facebook use and college students' life satisfaction, trust, and participation. *Journal of Computer-Mediated Communication*, 14(4):875–901, 2009.

[206] Hua Wang and Barry Wellman. Social connectivity in america: Changes in adult friendship network size from 2002 to 2007. *American Behavioral Scientist*, 53(8):1148–1169, 2010.

[207] Jing Wang, Siddharth Suri, and Duncan J Watts. Cooperation and assortativity with dynamic partner updating. *Proceedings of the National Academy of Sciences*, 109(36):14363–14368, 2012.

[208] Duncan J Watts. Networks, dynamics, and the small-world phenomenon. *American Journal of sociology*, 105(2):493–527, 1999.

[209] Duncan J Watts. The âĂIJnewâĂİ science of networks. *Annu. Rev. Sociol.*, 30:243–270, 2004.

[210] Duncan J Watts. Should social science be more solution-oriented? *Nature Human Behaviour*, 1(1):0015, 2017.

[211] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *nature*, 393(6684):440, 1998.

[212] Geoffrey B West. *Scale: the universal laws of growth, innovation, sustainability, and the pace of life in organisms, cities, economies, and companies.* Penguin, 2017.

[213] Norbert Wiener. *Cybernetics or Control and Communication in the Animal and the Machine*, volume 25. MIT press, 1961.

[214] Thomas N Wisdom, Xianfeng Song, and Robert L Goldstone. Social learning strategies in networked groups. *Cognitive science*, 37(8):1383–1425, 2013.

[215] Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of economic perspectives*, 18(2):107–126, 2004.

[216] David H Wolpert, William G Macready, et al. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

[217] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688, 2010.

[218] Lingfei Wu, Dashun Wang, and James A Evans. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378, 2019.

[219] Stefan Wuchty, Benjamin F Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

[220] Tal Yarkoni and Jacob Westfall. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122, 2017.

[221] Morris Zelditch Jr. Can you really study an army in the laboratory.

[222] Martín G Zimmermann, Víctor M Eguíluz, and Maxi San Miguel. Coevolution of dynamical states and interactions in dynamic networks. *Physical Review E*, 69(6):065102, 2004.