Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Task Space: An Integrative Framework for Team Research

Xinlan Emily Hu, Mark E. Whiting, Linnea Gandhi, Duncan J. Watts, Abdullah Almaatouq

To cite this article:
Xinlan Emily Hu, Mark E. Whiting, Linnea Gandhi, Duncan J. Watts, Abdullah Almaatouq (2026) The Task Space: An Integrative Framework for Team Research. Management Science

Published online in Articles in Advance 05 Mar 2026

. https://doi.org/10.1287/mnsc.2023.03544

https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–20
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

The Task Space: An Integrative Framework for Team Research

Xinlan Emily Hu,[a,b,*] Mark E. Whiting,[a,c,d] Linnea Gandhi,[a] Duncan J. Watts,[a,d,e] Abdullah Almaatouq[b,f,g,*]

[a] Operations, Information, and Decisions Department, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [b] Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [c] Pareto AI, Stanford, California 94305; [d] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [e] Annenberg School for Communication, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [f] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [g] Center for Computational Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

*Corresponding authors

Contact: xehu@mit.edu, https://orcid.org/0000-0001-9439-3498 (XEH); mark@pareto.ai, https://orcid.org/0000-0002-6395-7833 (MEW); lgandhi@upenn.edu, https://orcid.org/0000-0001-5166-9431 (LG); djwatts@seas.upenn.edu, https://orcid.org/0000-0001-5005-4961 (DJW); amaatouq@mit.edu, https://orcid.org/0000-0002-8467-9123 (AA)

Received: October 31, 2023
Revised: September 3, 2024; April 5, 2025; September 17, 2025
Accepted: October 9, 2025
Published Online in Articles in Advance: March 5, 2026

https://doi.org/10.1287/mnsc.2023.03544

Abstract. Research on teams spans many contexts, but integrating knowledge from heterogeneous sources is challenging because studies typically examine different tasks that cannot be directly compared. Most investigations involve teams working on just one or a handful of tasks, and researchers lack principled ways to quantify how similar or different these tasks are from one another. We address this challenge by introducing the "Task Space," a multidimensional space in which tasks—and the distances between them—can be represented formally, and use it to create a "Task Map" of 102 crowd-annotated tasks from the published experimental literature. We then demonstrate the Task Space's utility by performing an integrative experiment that addresses a fundamental question in team research: *when do interacting groups outperform individuals?* Our experiment samples 20 diverse tasks from the Task Map at three complexity levels and recruits 1,231 participants to work either individually or in groups of three or six (180 experimental conditions). We find striking heterogeneity in group advantage, with groups performing anywhere from three times worse to 60% better than the best individual working alone, depending on the task context. Critically, the Task Space makes this heterogeneity predictable: it significantly outperforms traditional typologies in predicting group advantage on unseen tasks. Our models also reveal theoretically meaningful interactions between task features; for example, group advantage on creative tasks depends on whether the answers are objectively verifiable. We conclude by arguing that the Task Space enables researchers to integrate findings across different experiments, thereby building cumulative knowledge about team performance.

History: Accepted by Sameer Srivastava, organizations.

Funding: The authors thank the Alfred P. Sloan Foundation [Grant #202-13924] and the MIT Wade Fund for their generous support of this research.

Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2023.03544.

Keywords: organizational studies • organizational studies: effectiveness-performance • group tasks • generalizability • statistics: design of experiments

1. Introduction

Teams are ubiquitous. They play key roles in areas as diverse as the military (Goodwin et al. 2018), healthcare (Valentine et al. 2015), and corporate governance (Peterson et al. 1998, Barrick et al. 2007). However, the variety in both teams and their related research has made it difficult to produce a coherent science of teamwork. Hollenbeck et al. (2012) astutely observed that "the diversity of this expanding research … creates certain challenges. Perhaps the greatest is the problem it causes in generating a cumulative knowledge base for meaningfully integrating and aggregating results across studies."

Consider, for example, the question of when groups outperform individuals—what researchers call *group advantage* or *team synergy* (Larson 2010). This question is at the heart of lively scholarly debates in the social and behavioral sciences (Hill 1982, Larson 2010, Almaatouq et al. 2021a) and underlies critical managerial and organizational decisions (Barrick et al. 2007). Although it is

generally accepted that interacting groups can have a performance advantage over independent work, the specific conditions under which this occurs remain unclear (Almaatouq et al. 2024a). As with many questions in team research, there is not so much a single answer as there is a collection of phrases starting with "it depends." In particular, researchers have identified the task being performed as a key determinant of group performance (Morris 1966, Steiner 1972, Herold 1978, McGrath 1984, Stewart and Barrick 2000, Whiting et al. 2019, Meluso and Hébert-Dufresne 2023). For example, Husband (1940)'s early study of puzzle-solving teams concluded that groups are better for tasks "requiring definite originality and insight," but not for those that are predictable and routine. More recent work by Almaatouq et al. (2021b), using a room assignment task, demonstrated that groups outperform individuals only when task complexity is sufficiently high. Still other studies show conflicting patterns: groups are advantageous at intellective tasks where members can verify correct answers—the "truth wins" phenomenon (Laughlin et al. 2006)—but generate fewer creative ideas than individuals working separately (Diehl and Stroebe 1987, Mullen et al. 1991).

While researchers have long recognized that tasks matter, the problem is that tasks are complex and multidimensional, and we lack a systematic way to quantify which specific attributes drive these different outcomes. And though there have been multiple proposed frameworks for classifying tasks (Shaw 1963, Steiner 1972, McGrath 1984, Laughlin and Ellis 1986), these classifications not only rely on overly broad categories but also lack consensus about which task dimensions are most important (Larson 2010). Categorizing tasks into broad, unidimensional "types"—such as "Intellective" and "Maximizing"—obscures sources of variation that are not captured by the category label. For example, while completing a crossword puzzle and solving an algebraic equation are both "Intellective" tasks that involve identifying a correct answer, they differ in other respects that could meaningfully impact group outcomes (such as the degree of mathematical reasoning required, the role of vocabulary knowledge, or the potential for collaborative problem-solving). Making matters worse, task frameworks can describe the same task in different, often incomparable, terms. According to McGrath (1984), both crossword puzzles and algebra questions are Intellective tasks; under Zigurs et al. (1999)'s framework, the algebra question is a "Simple Task" because it has only one unambiguously correct answer, while the crossword puzzle is a "Problem Task" because it involves uncertainty about which words might fit a given clue. Thus, two tasks that appear equivalent in one framework can be placed in entirely different categories in another.

The lack of a common set of dimensions (*incommensurability*) makes it difficult to create meaningful comparisons across studies; when a finding from one setting fails to generalize to another (Chang et al. 2021, Yarkoni 2022, Almaatouq et al. 2024a) it is unclear whether to attribute the discrepancy to replication failure, methodological error, or sufficiently dissimilar tasks. Enabling commensurability across tasks—and allowing knowledge to be better integrated into a cohesive whole—is therefore critical to the research of teams.

To fill this gap, we introduce a 24-dimensional design space of team tasks, which we call the "Task Space," that synthesizes task taxonomies[1] from the literature to date. Each dimension in the space represents a specific construct motivated by prior research (e.g., Shaw 1963, Steiner 1972, McGrath 1984, Laughlin and Ellis 1986). By positioning tasks as points in this multidimensional space rather than forcing them into broad categories, we enable quantitative comparisons between any two tasks. We additionally label a repository of 102 tasks sourced from the empirical literature on team performance, creating what we call the "Task Map." The Task Map serves as the basis of a large-scale study that demonstrates our approach, as well as a standalone contribution for use in future research.

The Task Space solves the incommensurability problem: researchers can now quantify how similar Husband's puzzle-solving tasks are to Almaatouq's room assignment task, and determine whether conflicting findings reflect genuine boundary conditions or incomparable contexts. Beyond reconciling past findings, the Task Space enables systematic research design (Almaatouq et al. 2024a), allowing researchers to select tasks with specific features or sample tasks that are maximally different. Finally, the Task Space can serve as a living resource, in which new tasks and dimensions will be added as our understanding evolves. This unified framework allows future discoveries to be seamlessly integrated into the cumulative knowledge base rather than be viewed as isolated findings.

We empirically demonstrate the utility of the Task Space through a large-scale integrative experiment examining whether groups outperform individuals. After systematically sampling 20 diverse tasks from the Task Map, we build each task at three levels of complexity and randomly assign 1,231 participants to work either individually, in groups of three, or in groups of six. This design resulted in 5,972 observations across 180 unique experimental conditions. Our experiment reveals striking heterogeneity in group advantage depending on the task context. Critically, the heterogeneity is predictable—by using dimensions from the Task Space in predictive models, we achieve significantly higher out-of-sample performance than models relying on traditional categorical approaches. These

findings illustrate how the Task Space can transform isolated, task-specific findings into cumulative knowledge about when and why it is advantageous to hire a team.

2. Background

The Task Space extends a long history of prior work that has examined the key categories or features of tasks and that has produced numerous competing taxonomies (e.g., Shaw 1963, Hackman 1968, McCormick et al. 1972, Tushman 1979, Wood 1986, Driskell et al. 1987, Cohen and Bailey 1997, Peterson et al. 2001, Wildman et al. 2012). In our review of the literature, we identified 15 such taxonomies (Online Appendix A) and found significant heterogeneity across them. More problematically, though each taxonomy, when viewed in isolation, offers a plausible account of how tasks differ from each other, when viewed in aggregate, the accounts are not collectively coherent—partly because they focus on different attributes of tasks and partly because they adopt different definitions of a "task" to begin with. For example, some frameworks emphasize the task's stimulus material (e.g., Hackman 1968, McGrath 1984), while others focus on the primary goal (e.g., Laughlin and Ellis 1986), the work process (e.g., Steiner 1972), the required skills (e.g., Roby and Lanzetta 1958, Fleishman 1975), and even the level of participants' interest in the task (Shaw 1963).

These divergent perspectives reflect three related sources of incommensurability: (1) the conceptual paradigm; (2) the data source or setting; and (3) the level of detail used to describe the task. The first source, one originally noted by Hackman (1969), is that researchers use different conceptual paradigms when thinking about tasks. Some researchers use a *task qua task* paradigm, which defines a task purely from its stimulus material. Others expand the focus to participants' downstream interactions: *task as behavior requirement* emphasizes the behaviors that successful participants should achieve, while *task as behavior description* focuses on the typical behaviors that participants display. Still others adopt a *task as ability requirement* approach, focusing on personal characteristics that influence how participants approach the task. These different paradigms naturally require different types of information: a *task qua task framework* needs only details about the stimulus itself, while other approaches might require participant demographic information or empirical data on typical participant reactions.

A second source of incommensurability is that taxonomies draw on data from different settings, and consequently make different assumptions about the nature of the task and the individuals completing it. Some frameworks are oriented toward the laboratory (e.g., Shaw 1963, Hackman 1968, Steiner 1972, McGrath 1984) and therefore assume that tasks are short, self-contained activities with clear goals. Others are oriented toward real work environments (e.g., Hackman and Oldham 1975, Driskell et al. 1987, Peterson et al. 2001, Wildman et al. 2012) where tasks tend to be more open-ended and place greater demands on employees' abilities and skills. Still other frameworks synthesize features from both laboratory and field contexts (e.g., Driskell et al. 1987). Categorization of tasks can therefore be highly specific to the setting being studied, such as "Basic Research" versus "Applied Research" (Tushman 1979) or "Managing Others," "Advising Others," and "Human Service" (Wildman et al. 2012).

Finally, task frameworks specify their features using varying levels of detail. While some frameworks require information about the exact mechanics of performing a task—down to the precise number of actions involved (Wood 1986) and the duration of each stimulus (Fleishman 1975)—others place tasks into broad categories, such as "Production," "Discussion," and "Problem-Solving" in Hackman (1968) and the eight categories in McGrath (1984). Although a general description ("this task is about solving math problems") would be enough to categorize a task as "Problem-Solving" in Hackman's framework, Wood would require additional implementation details to determine the problem's complexity.

Thus, differences in *paradigms*, *settings*, and *levels of detail* often result in task frameworks that cannot "speak to one another." Given the same details regarding two tasks, one framework might place the tasks into the same category; another might place them into opposing categories; and a third framework might not be able to categorize them at all, because it requires a different set of information or makes a different assumption about the environment.

Our observation of incommensurability in task frameworks has parallels with other studies of the features underlying team-related constructs. For example, the critique by Hollenbeck et al (2012) of using taxonomies to determine "types" of teams also applies to our case ("types" of tasks). In fact, substituting the word "team" for "task," their passage almost exactly replicates our point:

> The literature on ~~teams~~ [*tasks*] proposes a dizzying array of different ~~team~~ [*task*] types, even though the number of actual underlying dimensions used as building blocks to construct ~~team~~ [*task*] types is limited. This state of affairs impedes the meaningful accumulation of results across studies and, in general, makes it very difficult for researchers or consumers of research to answer the question, "What kind of ~~team~~ [*task*] is this?"

In Section 3, we seek to resolve task frameworks' incommensurability by building a flexible design space of tasks. First, we describe our multidimensional

approach, in which each task is allowed to have a nonzero association with any task dimension, rather than being assigned to a single "category" or "type." Second, we define two related but distinct notions of a task (the *task class* and *task instance*), allowing us to clarify issues around the conceptual paradigm and level of detail. Finally, we identify a set of 24 dimensions that are relevant at the level of the task class, and we describe our methods for evaluating a task along these dimensions. By positioning tasks within our 24-dimensional design space, we produce a unified answer for how similar or different two tasks are.

3. Introducing the Task Space: Design and Construction

3.1. Representing Tasks as Vectors in Multidimensional Space

In the Task Space, each task is represented as a vector with M elements, where each element corresponds to the task's value on one of M dimensions. A collection of N tasks can thus be represented as an $N \times M$ matrix, where each row is a task and each column is an attribute of the task (Figure 1, left panel). This representation enables researchers to compute distances or similarities between tasks using standard metrics (e.g., Euclidean distance, cosine similarity), and to identify clusters of related tasks, systematically sample tasks for experiments, and quantify semantic relationships (Figure 1, right panel).

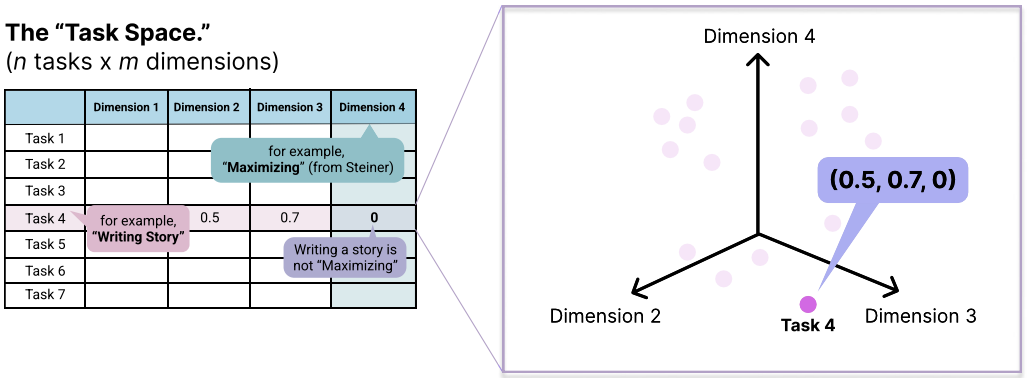
The vector representation integrates a variety of proposed task constructs while making few additional assumptions. For example, unlike in McGrath (1984), we do not assume mutually exclusive or collectively exhaustive categories, nor do we assume any particular topological structure (e.g., a circumplex). Reading the matrix in Figure 1, we can see how different frameworks would describe each task. For instance, "Writing Story" would not be considered a Maximizing task according to Steiner (1972), and hence the row corresponding to this task has a value of 0 in the column corresponding to the Maximizing dimension. In contrast, the Shopping Plan is (almost by definition) a Planning task according to McGrath (1984); thus, the row corresponding to this task has a value of 1 in the Planning dimension.

The matrix representation makes the Task Space inherently flexible and extensible. Researchers can easily add new tasks or dimensions, update existing values, or focus on subsets relevant to their research questions. Because these modifications involve only simple matrix operations, different research groups can maintain compatible versions while adapting the space to their specific needs.

More concretely, we can think of downstream applications in machine learning terms: theory-informed task dimensions serve as features to predict outcomes of interest, and the results of these predictions, in turn, inform the selection and development of task dimensions. This iterative process progressively refines the Task Space as more data becomes available. Features that do not predict relevant outcomes may be removed, and puzzling observations—where tasks occupying the same location have different outcomes—may suggest the need for new dimensions. New tasks may also inspire the addition of previously-unexplored dimensions. Figure 2 illustrates this process, in which two scientists use the multidimensional representation to unify different subsets of task dimensions, apply feature engineering techniques, and refine the Task Space through prediction.

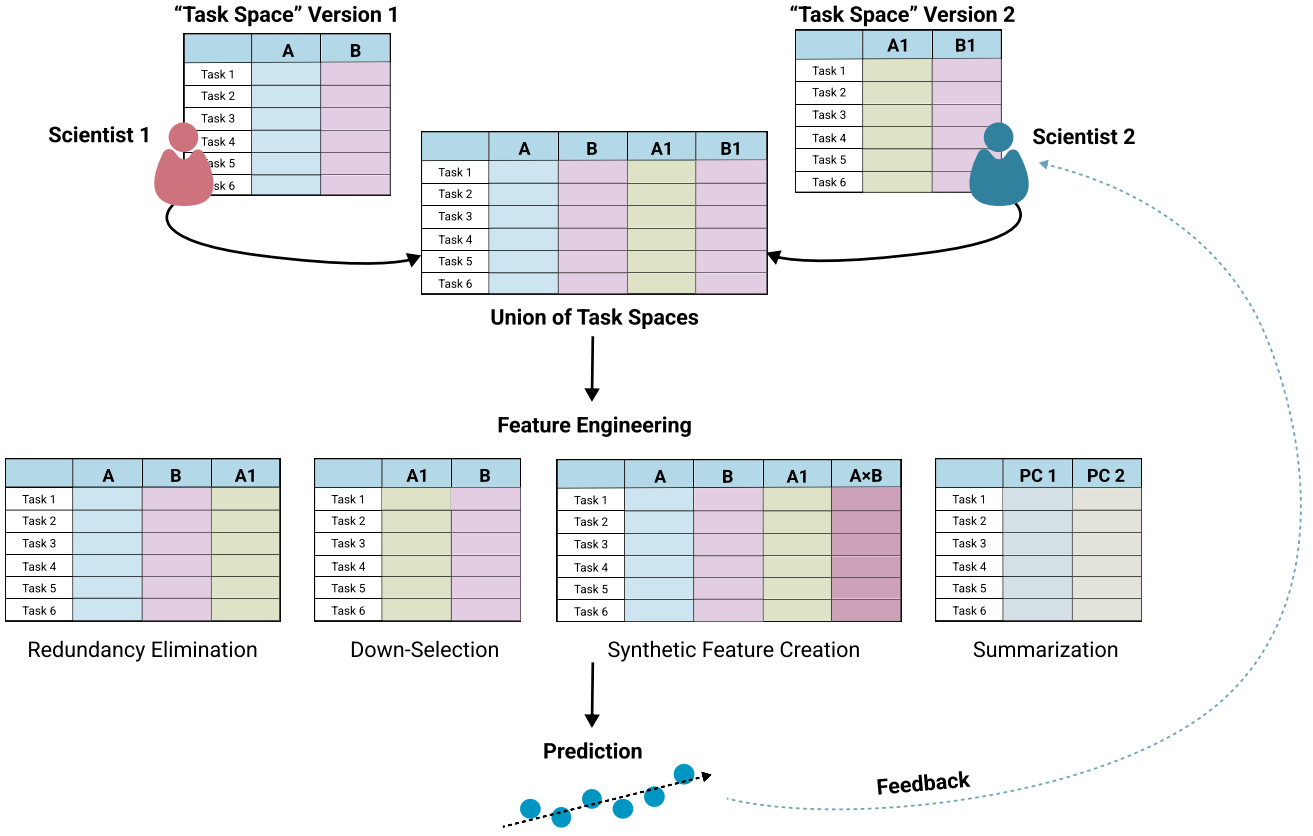
When constructing the Task Space, we deliberately refrained from applying factor analysis or combining conceptually similar dimensions *ex-ante*. Because there are many plausible methods of engineering features,

Figure 1. (Color online) An Illustration of the Task Space



Notes. Left: Our labeled set of tasks can be thought of as a 102 tasks × 24 dimensional matrix in which each task is represented as a row vector for which each element (column) is a dimension of the task. Right: Each row vector in this matrix can also be mapped to a point in 24-dimensional space. This representation makes the Task Space easily amenable to linear algebra-based analysis (e.g., finding similarity between vectors, clustering, and sampling).

Figure 2. (Color online) An Illustration of the Iterative Theory Construction Process



Notes. Scientists 1 and 2 each begin with different "versions" of the Task Space; Scientist 1's version is a two-dimensional space with features A and B, while Scientist 2's version is a two-dimensional space with features A1 and B1. To make their versions commensurate, the scientists can simply take the union of all features, creating a four-dimensional space. The unified space can then be transformed in a variety of ways ("feature engineering"), which may include eliminating redundant features, down-selecting to a subset of features, creating synthetic features, or summarizing key factors. The engineered features are then used to predict outcomes of interest. How well task features perform in predicting outcomes will inform additional changes to the task features. For example, task features that do not correlate with any known outcomes can be regarded as irrelevant and removed from the Task Space; task features that consistently improve model prediction on known outcomes can be retained.

our view is that refinements of the Space should be guided by applying it to practical questions, rather than by statistical relationships in an arbitrary set of existing tasks. This emphasis on empirical validation mirrors successful dimensional refinement processes in other domains. In early personality psychology, for example, researchers proposed thousands of personality dimensions (Allport and Odbert 1936) before eventually converging on what is now the "Big Five" (Goldberg 1993). The dimension reduction process was driven by a combination of empirical research and vigorous theoretical debate. For example, Hough (1992) argued that Affiliation (socialization preference) and Potency (energy level) should not have been combined into the single dimension of Extroversion, because they predicted different work-related outcomes (Potency predicted work success, but Affiliation did not).

In contrast, the team performance field has yet to begin a similar convergence, because researchers lack an effective means to make comparisons across frameworks. In a review of meta-analyses on team performance in the last 40 years (Online Appendix H), we found that no publication classified tasks using more than one taxonomy or typology; some only coded tasks according to a single dimension (e.g., task interdependence or task complexity); and some included no information about the task at all. To eventually converge to a potential "Big Five" of task dimensions, the first step is to test the predictive power of task dimensions across multiple task taxonomies. This process requires unifying taxonomies into a shared representation; in other words, to construct a *Task Map* that helps researchers connect the spatial position of a task to some specific outcome of interest.

3.2. Construction of the Task Map

The Task Map is a repository of 102 tasks that are annotated according to the 24 dimensions of the Task Space—it is a "worked example" of applying our proposed matrix representation to an initial set of commonly used tasks and frameworks. This section

considers the construction of the map in terms of its dimensions (columns), tasks (rows), and ratings (cells).

3.2.1. Dimensions (Columns). To establish explicit criteria for the dimensions we include in the Task Map, we introduce the notion of a *task class* and *task instance*. A *task class*, following the task definitions by Larson (2010) and Hackman (1968), represents the general blueprint of a task, as characterized by its stimuli and goals. For example, tasks in the Room Assignment class involve a set of students, rooms, and constraints (the stimuli), in which the students must be optimally allocated (the goal). Importantly, however, there may be many variations within the class—a "low complexity" version might involve assigning students to rooms under only one constraint, while a "high complexity" version might involve dozens of students, rooms, and constraints. We call these variations *task instances*.

This definition clarifies a key issue with existing taxonomies, in which there is no explicit consensus about the boundaries of a task, leading various frameworks to implicitly mix class- and instance-level attributes. For example, three of the paradigms we reviewed in Section 2 (*task as behavior requirement*, *task as behavior description*, and *task as ability requirement*) implicitly operate at the instance level, as they require specific details about the participants and their reactions to the stimuli. Meanwhile, the *task qua task* paradigm, which focuses on the stimuli alone, can operate at either the class or instance levels, depending on whether the dimensions are general (e.g., does this task involve solving a problem?) or specific (e.g., how many operations does it take to complete the problem?).

Class- and instance-level features convey different amounts of information. Class-level features represent the coarsest level of detail; they are stable regardless of who completes the task, how they choose to complete the task, or the format in which the task is presented. Instance-level features comprise these fine-grained specifics. For example, every Room Assignment Task is an optimization problem, but the level of complexity, allowable group processes, user interface, and expected behaviors can vary. Following Larson, we treat class-level features as more "core" than instance-level features, and hence restrict the dimensions that we will use for the Task Map to those describing a task class. This restriction does not imply that we ignore instance-level attributes; rather, because there are too many variations to comprehensively describe, we propose that researchers incorporate instance-level attributes as additional dimensions (in Section 4, our empirical case study will account for task complexity in this manner).

Tasks in the laboratory and in the field are often different in nature, so we further restrict our attention in this paper to laboratory tasks. These tasks typically have very well-defined stimuli and goals due to their use in evaluating team performance in controlled settings. Thus, they are suitable for demonstrating our method in a setting that is complicated enough to highlight its benefits while keeping it simple enough to explain. We emphasize, however, that there is nothing about the method we outline that in principle excludes field tasks; hence, we expect that it will be possible to expand the Task Space to include them in the future (see Section 5.2 for more discussion).

Applying these two criteria resulted in 24 dimensions[2] sourced from five frameworks (Shaw 1963, Steiner 1972, McGrath 1984, Laughlin and Ellis 1986, Zigurs et al. 1999; see Table 1). Further implementation details can be found in Online Appendix C.4.[3]

3.2.2. Tasks (Rows). We sourced 102 tasks from published papers. Groups and teams are of interdisciplinary interest, with research spanning management (Marks et al. 2001, Ericksen and Dyer 2004), psychology (Mathieu et al. 2017, Salas et al. 2018), computer science (Harris et al. 2019), sociology (Gross 1954), economics (Weidmann and Deming 2021), complexity science (Almaatouq et al. 2021a), and other fields. Thus, our aim was not to create a systematic or fully representative sample of any one field, as each discipline has its own, often nonoverlapping, set of publication venues. Instead, we hand-curated a set of tasks from a diverse set of publications across different disciplines, focusing on common task paradigms and influential work. These include the Collective Intelligence Task Battery (Woolley et al. 2010), forecasting and prediction tasks (Silver et al. 2021), "classic" tasks used in group studies (Lorge and Solomon 1960), economic games (Camerer 1997), and others; see Online Appendix B for the complete list.[4] In our repository, each task is summarized in a standard format that describes the stimuli and goals (the two defining components of the task class). This repository is the basis of our empirical demonstration in Section 4, as well as a broader set of applications discussed in Section 5. It is also publicly available, alongside complete data and documentation,[5] enabling future researchers to replicate our task labeling for other domains of interest.

3.2.3. Ratings (Cells). To construct the Task Map, the 102 tasks were rated on 23 dimensions by a panel of 121 Amazon Mechanical Turk workers, who had passed our pretest with an average score of 84.56%.[6] We initially excluded a 24th dimension, Type 6 (Mixed-Motive), because we were focused on dimensions applicable to both individual and team tasks; however, this dimension was reintroduced at a later stage and separately annotated by the author team (see Online Appendix C.4). Over the 23 crowd-rated dimensions, each task received an average of 23.20 ratings, and ratings across workers were averaged at the question level

Table 1. List of Dimensions in the Task Space

No.	Dimension name	Question text	Source
1	Conceptual-Behavioral	Does this task primarily require physical effort, as opposed to primarily requiring mental effort?	McGrath
2	Intellectual-Manipulative*	What is the fraction of physical (as opposed to mental) effort required for the task?	Shaw
3	Type 1 (Planning)	Is this a "planning" task? In other words, is one of the main purpose(s) of this task to produce a sequence of concrete steps or actions that an individual can follow to achieve some goal?	McGrath
4	Type 2 (Generate)	Is this a "generation" or "brainstorming" task? In other words, is one of the main purpose(s) of this task to produce a number of ideas or examples, without any particular action associated with them?	McGrath
5	Creativity Input*	What fraction of the effort required for this task is creative thinking (as opposed to any other type of effort, whether physical, logical, etc.)?	New; Continuous version of Type 2 (Generate)
6	Type 3 and Type 4 (Objective Correctness)	Is there an objectively correct solution to this task that can be calculated or selected?	McGrath
7	Type 5 (Cognitive Conflict)	Is one of the main purpose(s) of this task to resolve people's differences in opinion, perspective, or viewpoint?	McGrath
8	Type 6 (Mixed-Motive)†	During this task, group members pursue different interests, objectives, or goals.	McGrath
9	Type 7 (Battle)	Can the outcome of this task be described in win/lose terms?	McGrath
10	Type 8 (Performance)	Does the task have an all-or-nothing outcome? In other words, is it sufficient to just meet a particular standard?	McGrath
11	Divisible-Unitary	Is it efficient and useful for members of the group to work on discrete parts (or subtasks) of this activity?	Steiner
12	Maximizing	Is the goal (or one of the goals) of this task to try to achieve doing something as much as possible, as many as possible, or as quickly as possible?	Steiner
13	Optimizing	Is the goal (or one of the goals) of this task to try to achieve a precise outcome?	Steiner
14	Outcome Multiplicity	Is there only one "best" solution (or possible solution) to this task?	Shaw, Zigurs et al.
15	Solution Scheme Multiplicity	Is there only one method for achieving the task, as opposed to many alternatives for task completion?	Shaw, Zigurs et al.
16	Decision Verifiability	Can acceptable solutions to this task be demonstrated or verified to be correct (e.g., by an expert or third-party)?	Laughlin and Ellis
17	Shared Knowledge	Can this task be written as a "formal model" that an algorithm could solve?	Laughlin and Ellis
18	Within-System Solution	Is there enough information in the problem to find a valid solution?	Laughlin and Ellis
19	Answer Recognizability	If someone who is able to solve the problem explains their answer, would others recognize it as correct without contest?	Laughlin and Ellis
20	Time Solvability	Is a participant able to come up with a provably correct solution, assuming sufficient ability, time, motivation, and resources?	Laughlin and Ellis
21	Intellective-Judgmental*	On a scale of 0 (entirely subjective, with no correct answer; a judgmental task) to 1 (entirely objective and demonstrable by pure logic; an intellective task), how would you classify the extent of demonstrable correctness of this task?	Laughlin and Ellis
22	Conflicting Tradeoffs	Does completing this task require participants to evaluate tradeoffs—conflicting possible solutions or conflicting pieces of information?	Zigurs et al.
23	Solution Scheme Outcome Uncertainty	When doing this task, will the participants have any uncertainty about whether their method or solution will lead to the desired outcome?	Zigurs et al.
24	Eureka Question	Is the solution to the question obvious as soon as it is proposed—for example, once people see the "trick," they know how to solve it?	Laughlin and Ellis

Notes. A summary of the 24 dimensions included in the Task Space. 20 of the dimensions were rated on a binary 0-1 scale; three dimensions, marked with an asterisk (*), were rated on a continuous scale from 0 to 1; one dimension, marked with a cross (†), was hand-coded by researchers (details in Online Appendix C.4). Note that some of these features capture similar or related concepts; we chose to ask separate questions for these concepts so that each question encodes a single idea and because similar concepts will be loaded onto the same factors (e.g., in factor analysis).

for each task and dimension.[7] A rating of 1 indicates that a rater perceived the feature as present, whereas a 0 indicates that a rater perceived the feature as absent; thus, the end result of the rating process is a 102 (tasks) × 24 (dimensional) matrix in which all cells contain a real value between 0 and 1, which we interpret as "the degree to which a task displays a particular feature." These values are then used to position tasks in an underlying vector space (for an illustration, please refer to Online Appendix E).

4. Empirical Case Study: A Large-Scale Integrative Experiment of Group Advantage

We next use the Task Map as part of a human-subject group experiment, in which we use task dimensions first to systematically select a diverse set of tasks, and second to predict group outcomes across the tasks. Our primary outcome is our running example of *group advantage*: the performance gain of working in a group, relative to an equivalent number of individuals working independently. Given heterogeneous findings across several decades of organizational research (e.g., Janis 1971, Steiner 1972, Diehl and Stroebe 1987, Stasson and Bradshaw 1995, Laughlin et al. 2006, Larson 2010), this question is an ideal test case for using the Task Space to integrate experimental results.

4.1. Study Design

To demonstrate the Task Space's potential, we followed an integrative experimental approach that allowed us to explicitly and quantitatively evaluate the effects of many task characteristics on group performance (Almaatouq et al. 2024a). Specifically, we selected 20 diverse tasks from the Task Map and created three versions of each task (with low, medium, and high complexity). We then randomly assigned participants to work on the tasks either alone or in a group of three or six members. This systematic variation yielded 180 unique experimental conditions (20 tasks × 3 complexity levels × 3 group sizes). To demonstrate the use of the Task Space for true out-of-sample generalization, we also used a sequential design of three experimental "waves" with a different collection of tasks in each. The first experimental wave contained 10 tasks, while the second and third waves each contained 5 tasks (details in Online Appendix F.2). In all analyses, we treat the first two waves as "training," since they represent a mixture of randomly-sampled and manually-selected tasks that maximize prior knowledge about group advantage; we then treat Wave 3 as the "held-out" set for evaluation.

Within each wave, the experiment proceeded identically. Participants were first assigned to work as part of a collaboration unit of either a single-player, three-player, or six-player group. The unit then completed a randomly-assigned block sequence of five tasks. Each task had bespoke specifications for parameters such as timing and the manipulation of complexity. When completing a given task, participants always experienced four versions: a trivial practice task followed by three graded versions at "low," "medium," and "high" complexity. For example, we increased the difficulty of Word Construction problems by requiring participants to generate words from lists that had fewer valid word combinations, and we increased the difficulty of the Room Assignment task by increasing the number of students, number of rooms, and the number/structure of constraints. This within-task manipulation allowed us to examine whether group advantage emerges only when the task instance is sufficiently complex (a phenomenon previously documented in Almaatouq et al. 2021b).

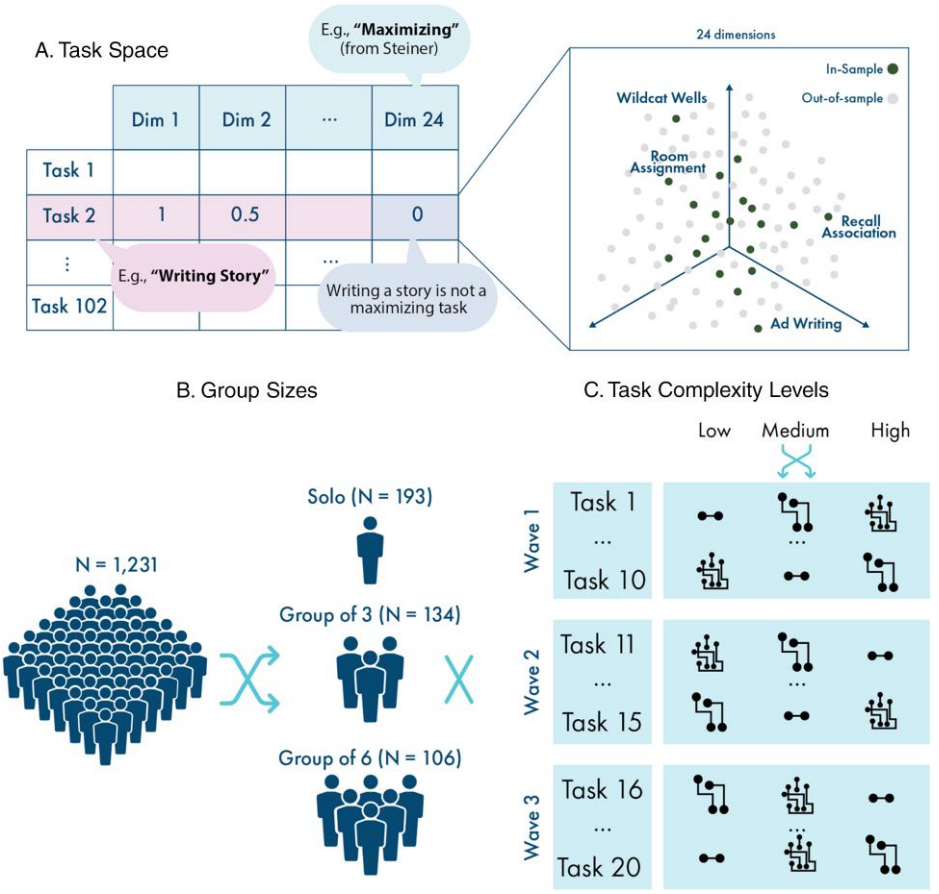
In summary, we collected data from 1,231 participants (193 individuals, 134 groups of three, and 106 groups of six), generating 5,972 observations across 180 conditions, where each observation represents one collaboration unit's performance on a given task at a given complexity level. The overall study design is shown in Figure 3 and detailed in Online Appendix F.

To contextualize the scale and scope of our demonstration, experimental studies of group performance routinely study just one task (e.g., Bavelas 1950; Krackhardt and Stern 1988; Laughlin et al. 2006; Bahrami et al. 2010; Mason and Watts 2011; Koriat 2012; Shore et al. 2015; Mao et al. 2016; Aggarwal and Wooley 2019; Almaatouq et al. 2020; 2021b; Straub et al. 2023; Almaatouq et al. 2024c), while one of the most ambitious previous studies of which we are aware (Woolley et al. 2010) comprised ten tasks with fixed complexity level and group size. Indeed, our study is comparable to many of the largest meta-analyses in the last 20 years (Mullen et al. 1991, Gully et al. 2002, De Dreu and Weingart 2003, Weber and Hertel 2007, Riedl et al. 2021), but offers several methodological advantages: all tasks were coded using a consistent framework (the Task Space); all experiments were conducted under identical conditions on the Empirica platform (Almaatouq et al. 2021b); and all participants were recruited from the same curated panel with detailed demographic information. Additionally, we avoided publication bias by treating all sampled conditions as informative regardless of statistical significance.

4.2. Computing Group Advantage

Our objective is to measure the extent to which interacting groups tend to outperform comparable individuals for a given task (*group advantage*). This requires us to compare group performance against a baseline of working independently. Here, following standard practice in the groups literature that distinguishes

Figure 3. (Color online) Illustration of the Experiment Design



Notes. Panel A shows how we use the Task Space to systematically sample 20 tasks for the experiment. Panel B shows the experiment design. Participants were randomized into one of three conditions: working independently, working in interacting groups of three, and working in interacting groups of six (Panel B). Participants experienced each task at three levels of complexity: "Low," "Medium," and "High" (Panel C). In our data, we consider each observation to be a single collaboration unit (an independent worker or an interacting group) performing a specific task (e.g., "Task 1") at a specific complexity level (e.g., "Low").

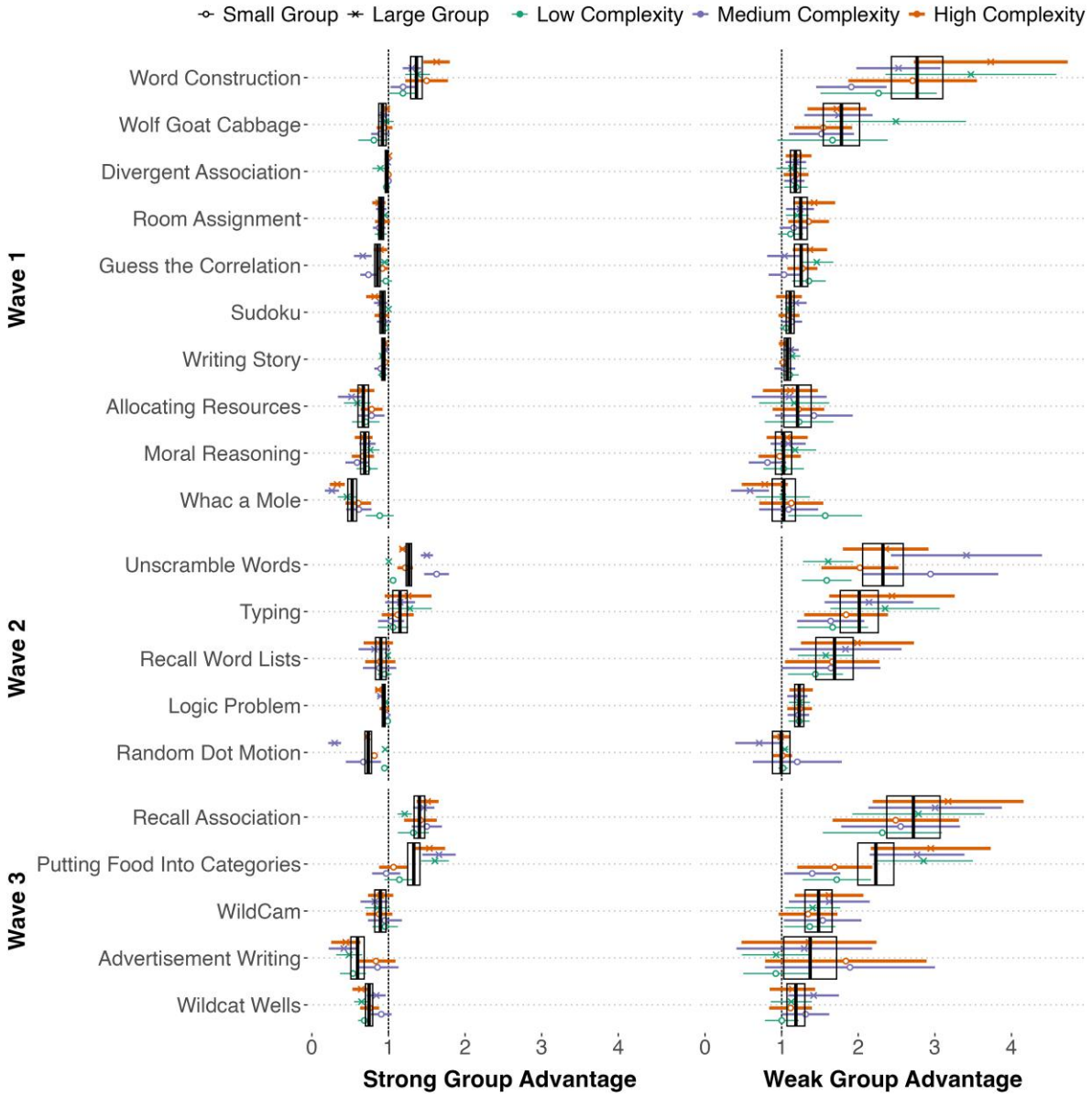
group advantage from mere economy of scale (Marquart 1955, Lorge and Solomon 1960, Larson 2010, Almaatouq et al. 2021a), we generated "nominal groups" by comparing each interacting group with a randomly permuted aggregation of individual participants who completed the same task instance. In other words, rather than directly comparing performance across the group and individual conditions, we treat the individuals as part of a "pool" from which we sample "nominal groups" that are equivalent in every respect except for the fact that they did not interact. We then ask, *what advantage does the ability to interact provide, compared with what we'd expect from an identical number of people who worked independently?*

Following prior work (Larson 2010, Hueffmeier and Hertel 2011, Meslec et al. 2014, Almaatouq et al. 2021a), we examine two variants of group advantage: *strong group advantage* is the ability of a group to perform better than even the best-performing individual in a nominal group, and *weak group advantage* is the ability of a group to perform better than a randomly-selected (i.e., an average) member in a nominal group. Together, these two measures capture different aspects of collaboration benefits (whether it is worthwhile to hire a group over an average individual, versus the most competent individual). Details for how we compute these two quantities are provided in Online Appendix G.1.

4.3. Results

4.3.1. Group Advantage is Heterogeneous. In the analyses that follow, we focus on the 120 group conditions (3- and 6-person groups), as the 60 individual conditions serve as the baseline for constructing nominal groups and are already incorporated into the group advantage calculation. Figure 4 shows the distribution of group advantage across each experimental condition. As expected, the phenomenon of group advantage is strikingly heterogeneous: across 20 tasks and 120 group conditions, interacting groups outperformed a

Figure 4. (Color online) Heterogeneity in Group Advantage Across Conditions



Notes. Each point represents the observed group advantage at the level of an experimental condition, which is a tuple of *task* (*y*-axis) × *level of complexity* (color/line thickness; low complexity is represented by a thin green line, medium complexity by a mid-thickness purple line, and high complexity by a thick orange line) × *group size* (point shape; small, three-person groups are represented by circles and large, six-person groups are represented by crosses). Tasks are grouped by the experimental wave in which they appeared (top three facets). Error bars represent the analytical 95% confidence intervals (1.96 * 1 standard error) for group advantage in a given condition. Boxes are centered around the mean and show one standard error.

random member of the nominal group (weak advantage) in 19 out of 20 tasks and nearly all experimental conditions (93%), but only outperformed the best member (strong advantage) in 5 tasks and a minority (26%) of conditions.

This heterogeneity aligns with longstanding observations that group advantage varies with factors such as the task, complexity, and group size (Diehl and Stroebe 1987, Mullen et al. 1991, Bahrami et al. 2010, Larson 2010, Thompson and Wilson 2015, Almaatouq et al. 2021a). However, Figure 4 also makes the stronger and less obvious point that this heterogeneity does not fall neatly along traditional lines of task "types." For example, among the types defined by the McGrath Circumplex, one of the most influential task typologies, there is almost as much within-type variation as there is overall. A Levene test of equality in variances shows that the McGrath task types do not significantly decrease the variance in strong and weak group advantage compared with the overall data (Generate, Strong: $W(1, 148) = 1.01$, $p = 0.32$; Generate, Weak: $W(1, 148) = 2.14$; $p = 0.15$; Intellective, Strong: $W(1, 184) = 3.38$, $p = 0.07$;

Intellective, Weak: $W(1, 184) = 0.43$; $p = 0.51$; Performance, Strong: $W(1, 130) = 3.61$; $p = 0.06$; Performance, Weak: $W(1, 130) = 0.07$, $p = 0.79$); refer to Online Appendix G.3.2 for details. For example, while Advertisement Writing and Word Construction are both classified as generation tasks (i.e., "Type 2" in McGrath's Group Task Circumplex), we observe strong group advantage for Word Construction but not for Advertisement Writing. Similarly, while other judgment tasks (such as Recall Word Lists and Random Dot Motion) show no strong advantage, the Recall Association task presents a notable exception, with the second-highest strong group advantage in our corpus.

This heterogeneity also has important implications for the way in which research on group performance is interpreted and generalized (Almaatouq et al. 2024a, b). To illustrate the potential problem, a researcher focusing solely on the Word Construction or Recall Association tasks (Figure 4; top of Wave 1 and Wave 3) would reasonably conclude that groups show both strong and weak group advantage, consistent with prior findings that communication between group members improves on members' individual decision making (Bahrami et al. 2010, Koriat 2012). However, if the researcher had instead chosen to focus on the Whac-a-Mole or Random Dot Motion task (Figure 4; bottom of Wave 1 and Wave 2), they would conclude that groups have no advantage over individuals at all, consistent with prior work that emphasizes the prominence of process losses (Steiner 1972). Finally, if they were to examine the Typing and Room Assignment tasks (Figure 4; middle of Wave 1 and Wave 2), they would find weak, but not strong, group advantage, leading them to conclude that, while groups are superior to a randomly-selected individual, it is preferable to rely on a small number of experts.

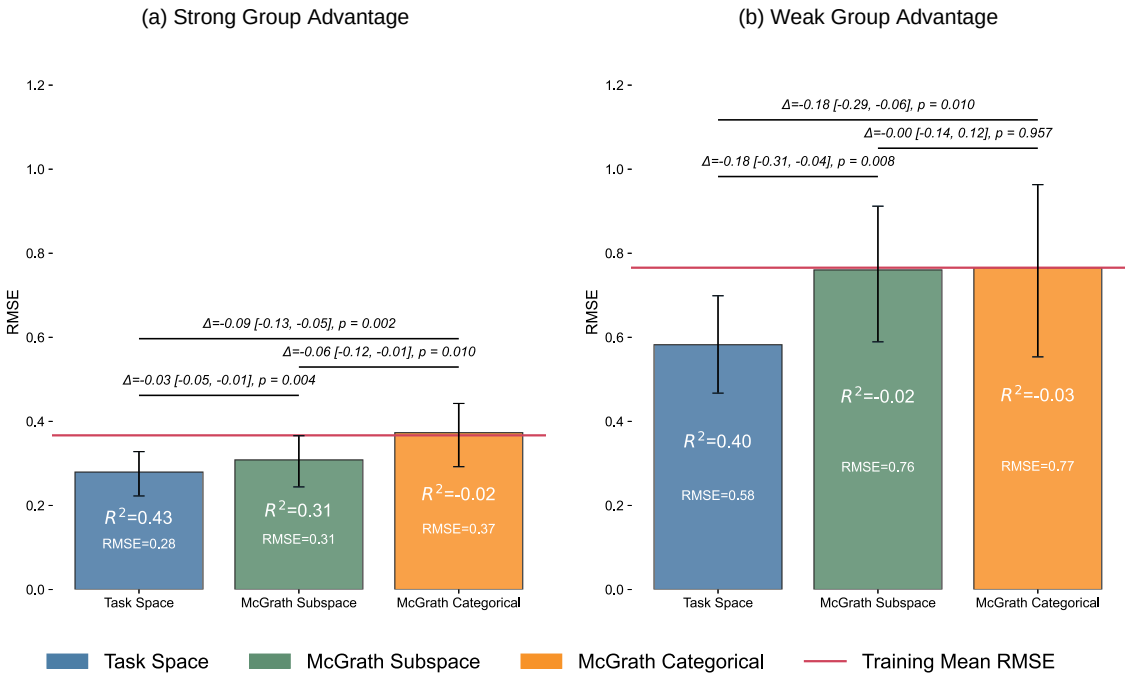
A similar point can be made about interaction effects. For example, Almaatouq et al. (2021b) found that weak group advantage for small groups increases with the complexity of the Room Assignment Task. We can replicate this effect using a linear mixed-effects model on our observational-level data, with random effects for the group identifier ($\beta = 0.114$, $p < 0.001$). However, this pattern does not hold across all tasks. In fact, we observe the opposite pattern in the Whac-a-Mole task: group advantage *decreases* as task complexity increases ($\beta = -0.107$ for strong; $\beta = -0.176$ for weak; $p < 0.01$). Online Appendix G.4 further explores interaction effects for complexity and group size, and we show that interaction patterns are heterogeneous across the 20 tasks in our data.

As all these examples illustrate, such substantial between-task heterogeneity implies that generalizations about group performance based on individual tasks are likely brittle and may not hold even within seemingly similar task types. Unfortunately, insisting that theoretical claims about group performance can only be assumed to hold for the exact task or tasks under consideration would effectively prevent any generalization to new settings, thereby severely undermining a fundamental benefit of theory (Yarkoni 2022). And without a way to reconcile potentially inconsistent results across tasks, the sheer number of existing results and possible tasks would lead to an unmanageable proliferation of theories, creating confusion about what is and is not known (Watts 2017, Levinthal and Rosenkopf 2020). Taking a middle ground, the Task Space allows for an *integrative* approach (Almaatouq et al. 2024a)—identifying the "coordinates" (i.e., the features) of the task at which groups are likely to outperform individuals. To quantify the informational content of the task features that we have defined, we frame this problem as an out-of-sample prediction exercise, where results acquired from one set of tasks can be used to predict group advantage in entirely new tasks. The associated models can, in turn, drive theory development by illuminating important relationships between task attributes, complexity levels, and group sizes.

4.3.2. Task Features Predict Group Advantage. To assess the predictive value of the Task Space, we use task features as covariates to predict group advantage at the condition level. We then train hyperparameter-tuned ElasticNets with two-way interactions on the 15 training tasks and evaluate them on the 5 held-out tasks. Our primary model uses all 24 Task Space dimensions; we compare this model against two baseline models inspired by the McGrath typology and a naïve baseline with no task information. The two McGrath baselines help to assess the Task Space's gains relative to a representative traditional typology: "McGrath Categorical" uses McGrath's task types as dummy variables to evaluate the gains attributable to using continuous rather than categorical dimensions, while "McGrath Subspace" uses the continuous Task Space dimensions inspired by McGrath's taxonomy to evaluate the gains attributable to expanding the dimensionality beyond a single taxonomy. Finally, the naïve baseline assesses the predictive performance of all three models against predicting the training data mean without accounting for any task features. All models include the exogenous factors of group size and task complexity and are evaluated using root mean squared error (RMSE). See Online Appendix G.2 for McGrath Categorical details, Online Appendix G.5 for model details, and Online Appendix G.6 for robustness checks.

Figure 5 shows that the Task Space models substantially outperform both the McGrath models and the naïve baseline. Strikingly, the McGrath Categorical model performs no better than ignoring task information and simply predicting the training mean. For

Figure 5. (Color online) Bar Plots of the Root Mean Squared Errors (RMSE) of Models Predicting Condition-Level Group Advantage; Lower RMSE is Better



Notes. All models are hyperparameter-tuned ElasticNets, trained on 15 tasks (Waves 1–2) and evaluated on 5 held-out tasks (Wave 3). The red line is the naïve baseline from predicting the training mean (i.e., ignoring task information). Error bars show bootstrapped 95% intervals for each model's performance on Wave 3 (resampling 1,000 Wave 3 datapoints with replacement and taking the 2.5th and 97.5th percentiles). Statistical significance is assessed using *paired* bootstrap differences computed on the same resamples; for each comparison, we report the mean difference in RMSE, the 95% CI and a two-sided p-value derived directly from the empirical distribution of the differences. Finally, we report an out-of-sample R^2 metric, which quantifies the proportion of out-of-sample variance (squared error) explained by our model relative to the naïve baseline.

strong group advantage, the Task Space achieves the lowest RMSE (0.28), significantly outperforming both McGrath Categorical (RMSE = 0.37, mean ΔRMSE = −0.093, 95% $CI_{\Delta RMSE}$[−0.131, −0.054], p = 0.002) and McGrath Subspace (RMSE = 0.31, mean ΔRMSE = −0.028, 95% $CI_{\Delta RMSE}$[−0.048, −0.010], p = 0.004). Additionally, McGrath Subspace significantly outperforms McGrath Categorical (mean ΔRMSE = −0.065, 95% $CI_{\Delta RMSE}$[−0.115, −0.012], p = 0.010). For weak group advantage, the patterns are similar: the Task Space again achieves the lowest RMSE (0.58), significantly outperforming both McGrath Categorical (RMSE = 0.77, mean ΔRMSE = −0.181, 95% $CI_{\Delta RMSE}$[−0.291, −0.056], p = 0.010) and McGrath Subspace (RMSE = 0.76, mean ΔRMSE = −0.178, 95% $CI_{\Delta RMSE}$[−0.310, −0.036], p = 0.008). However, the difference between McGrath Subspace and McGrath Categorical is not statistically significant (mean ΔRMSE = −0.004, 95% $CI_{\Delta RMSE}$[−0.137, 0.123], p = 0.957). We base our assessment of statistical significance on a resampling of 1,000 datapoints with replacement from the Wave 3 data, computing pairwise RMSE differences for each resample. We then report 95% confidence intervals and two-sided p-values derived directly from the empirical distribution of these differences (see Online Appendix G.5.2 for additional details).

Our results also suggest that strong group advantage is inherently more predictable than weak group advantage (baseline RMSE of ~0.4 versus ~0.8), likely due to lower overall variation in the dependent variable (Figure 4). Moreover, the same covariates can vary in predictive power across outcomes. For example, dimensions drawn from McGrath's taxonomy effectively predict strong group advantage but fail to do so for weak group advantage. In contrast, the multidimensional Task Space yields effective predictions for both, underscoring the value of capturing a wider range of task features—beyond the shift from categorical to continuous dimensions. Reinforcing this point, we conceptually replicate our results using the subset of dimensions inspired by the Steiner (1972) and Laughlin and Ellis (1986) taxonomies; in both cases, a more restricted set of features from a single taxonomy underperforms the complete Task Space (see Online Appendix G.6.2).

4.3.3. Key Task Space Dimensions for Predicting Group Advantage. To understand which specific task dimensions drive our prediction results, we next use

two complementary approaches to evaluate feature importance in our primary model. The first method, permutation analysis (Altmann et al. 2010), measures the increase in RMSE after a feature is randomly shuffled. Features that lead to worse RMSE when permuted are considered more important. The second method, Shapley Additive Explanations (SHAP; Lundberg and Lee 2017), quantifies the relative influence of each feature on the model's individual predictions. While permutation feature importance enables us to assess the magnitude of change on predictive accuracy, SHAP is useful for inspecting the directional impact of each feature on the model's predictions. Applying both methods in parallel allows us to identify both the features that are most important for determining group advantage, as well as whether these features help or hurt groups' expected performance relative to that of individuals.

Figure 6 presents our feature importance analysis, with permutation feature importance shown in Panels a and b and SHAP plots shown in Panels c and d. Features are ordered by their permutation importance rank. For both outcomes, the most predictive features is the extent to which a task has a demonstrably correct answer (Demonstrable Correctness), followed by the extent to which a task involves generating creative ideas (Creative). Additional top features include the extent to which a task has discrete subtasks that can be divided-and-conquered (Divisible-Unitary) and the extent to which there is uncertainty about solution correctness (Solution Scheme Outcome Uncertainty). The top feature (Demonstrable Correctness) accounts for ~20% and ~15% of the model's RMSE for strong and weak group advantage, respectively; the second-most important feature, Creative, accounts for just over 10% of the RMSE for both measures of group advantage. Both of these features also positively predict group advantage, as evidenced by the warm-colored/dark-shaded points (indicating high feature values) clustering to the right of the SHAP plot (indicating positive impact on model output).

Beyond these top predictors, importance drops considerably: the third-most important feature (Divisible-Unitary for strong group advantage, and Solution Scheme Outcome Uncertainty for weak group advantage) each accounts for close to 5% of the model's RMSE. Having clearly divisible subtasks positively correlates with group advantage, as groups can efficiently "divide and conquer" the activities. In contrast, uncertainty about solution correctness negatively correlates with group advantage.

Additionally, while some features show consistent patterns across both dependent variables—whether a task has an all-or-nothing outcome ("Performance" Tasks) ranks among the least predictive features for both strong and weak group advantage—other features vary in importance depending on the outcome. For example, whether a task has multiple possible outcomes (Outcome Multiplicity) ranks as the fifth-most important dimension for predicting strong group advantage, but is the second-least dimension for predicting weak group advantage.

We next turn our attention to the top two predictors, Demonstrable Correctness and Creative. The observation that groups outperform individuals on problems with a correct answer originates from Laughlin's early work on social decision schemes and intellective tasks (Laughlin et al. 1975, Laughlin and Ellis 1986), which posits that, once one member of the group identifies the correct answer, "truth wins." The same notion also underlies contemporary work on estimation and forecasting (Almaatouq et al. 2020, Becker et al. 2021, Almaatouq et al. 2022). Put another way, if the task has a correct answer that can be easily verified, it is easy for group members to efficiently search for the answer and to know when one of them has it right.

Curiously, however, we find that groups also outperform individuals at creative tasks, such as Word Construction and Putting Food Into Categories (which involves generating ideas for grouping food items). Here, our finding ostensibly contradicts the established result that interacting groups generate fewer ideas than those working independently (Diehl and Stroebe 1987, Mullen et al. 1991, Larson 2010). However, this contradiction can be resolved by considering the details of operationalizing group advantage; recall that, in typical idea generation tasks, researchers compute the nominal group's performance by combining the deduplicated raw ideas of all individuals, whereas we compare the group's final score to the final score of the best ("strong") and a randomly-selected ("weak") individual member of the nominal group. Group advantage is then represented as a ratio between the interacting group's score and the nominal group's score. In our data, groups in the Putting Food Into Categories task generate on average 1.6 times as many ideas as the "best" individual from a nominal team; however, this falls short of the theoretical maximum in which each group member contributes at their individual best level—effectively requiring a ratio of n (the number of group members) to demonstrate true superiority over simply combining independent work. In these cases, one might say that the group has an advantage over a single contributor, but it may not be "worthwhile" to facilitate real-time interaction. However, questions about the threshold of group worthiness can be highly context dependent. In a highly competitive market, even a tiny advantage might be extremely valuable and hence worth assigning to a group; and in different settings, the cost of hiring a group may differ substantially. Thus, interpreting our results will require careful consideration of these factors.

Figure 6. (Color online) Feature Importance Among Task Dimensions for Predicting Strong (Panels A and C) and Weak (Panels B and D) Group Advantage

Notes. Analyses are for the ElasticNet models, trained on the first 15 tasks (Waves 1–2) and evaluated on the final 5 tasks (Wave 3). Permutation feature importance is presented in Panels A and B, and SHAP values are presented in Panels C and D, ordered by their rank in permutation feature importance. In Permutation plots, positive bars indicate that a feature improves prediction, as error increases when the feature is shuffled; negative bars indicate that a feature merely adds noise, as shuffling the feature reduces error in out-of-sample prediction. In SHAP plots, color and shading indicate whether the feature was high (warm/dark color) or low (cool/light color) for a given observation, while location relative to the center line indicates whether they are positively (right of center) or negatively (left of center) influential in making a prediction for the observation. Note that, in order to avoid contaminating the permutation analysis with copies of highly correlated features, we create three sets of combined dimensions; "Creative" combines two dimensions related to the extent of creative thinking required; "Demonstrably Correct Answer" combines seven dimensions related to the extent to which the task has a correct answer that can be verified; and "Physical" combines two dimensions related to the extent to which a task requires physical effort.

In contrast to Demonstrably Correct and Creative tasks, one of the strongest negative predictors of group advantage is having uncertainty over whether a solution is correct (Solution Scheme Outcome Uncertainty). In some sense, the dimension of Solution Scheme Outcome Uncertainty is the inverse of Demonstrable Correctness; rather than having provably correct answers, tasks high in Solution Scheme Outcome Uncertainty have many plausible solutions, with uncertainty about whether each one might lead to the desired outcome. Indeed, these two dimensions are often connected in the literature; one review of 21 collective

intelligence studies finds that groups exhibit collective intelligence only for "well-structured" tasks, defined as tasks with "(1) one or more verifiably correct solutions (i.e., solution demonstrability) and/or the presence of several decision options and (2) a guaranteed procedure available to reach that solution" (Graf-Drasch et al. 2022, p. 742). Our findings align with Graf-Drasch et al.'s claim. In our data, groups display the strongest advantage for tasks that would be considered "well-structured." For example, one of the tasks with the highest group advantage is Unscramble Words, an anagram-solving task (which is well-structured in the sense of having both clearly-defined answers and a path to reaching them). In contrast, Moral Reasoning, a task that involves high uncertainty over the correct answer to a moral quandary, ranks as one of the tasks with the lowest group advantage. One possible explanation is that groups can be inefficient when navigating uncertainty, a pattern that aligns with groups' known biases in discussing limited information (Stasser and Titus 1985).

An important note is that, although we directly assign participants to different tasks (e.g., Word Construction versus Advertisement Writing), we do not independently manipulate the underlying features that characterize these tasks—these are the ratings taken directly from the Task Space. Thus, tasks will naturally exhibit different combinations of features, and some of this natural variation is useful for explaining puzzling observations in our data. For example, we observe strong group advantage for Word Construction, but not for Advertisement Writing, even though both tasks are similar on the Creative dimension. The puzzle is resolved by recognizing that Word Construction is also high in Demonstrable Correctness, since it is trivial to verify when one has generated a valid English word, whereas Advertisement Writing is low in Demonstrable Correctness and high in Solution Scheme Outcome Uncertainty, as evidenced by the difficulty of predicting click-through rates in headlines (Batista and Ross 2024). Consequently, groups working on Word Construction can much more efficiently combine their efforts (with each member generating as many words as they can), while groups working on Advertisement Writing are caught up in discussions about whether an advertisement is appealing or not. These findings further show that categorizing tasks into "types" fails to tell the whole story. Our multidimensional approach integrates disparate ideas from across social science (e.g., "truth wins," productivity losses) into a cohesive whole.

5. Conclusion

Writing in 1966, Charles Morris argued for "a systematic map of the effects of different task characteristics on various aspects of the group-interaction process." But in the decades since his writing, such a map was never created, despite many calls to better understand task features and their impact on social science theories (Hackman 1968, Fleishman 1975, Wood 1986). Over the years, continued challenges in building cumulative knowledge across conflicting taxonomies led to renewed calls to shift from categorical representations to multidimensional methods (Larson 2010, Hollenbeck et al. 2012, Lee et al. 2015, Almaatouq et al. 2024a). The Task Space answers these calls for the domain of team tasks. And while we present one application via our integrative experiment and prediction exercise, we next discuss our work within a broader conversation on developing research methods for generalizable, contextually-bounded theories across social science disciplines.

5.1. Discussion

5.1.1. Systematically Selecting Study Stimuli. In psychology, theories often deal with broad phenomena: that people seek to conform with groups (Asch 1956), that knowledge of the outcome biases human judgment (Baron and Hershey 1988), and that previous numerical information anchors new responses (Epley and Gilovich 2006). Each theory may have a plethora of moderating variables—for example, are people more likely to anchor to larger or smaller numbers? Do people commit more outcome bias when the stakes are higher?—and the theory is sometimes silent about the moderator. But just because the theory is silent does not mean the moderator is irrelevant. To address this concern, a substantial stream of literature advocates for using systematic variations in a study's stimuli to determine the limits of an effect's generalizability. For example, rather than study anchoring by giving all participants the same anchor number, the experimenter might select a range of numbers that represent different categories: very large numbers (10 billion), very small numbers (0.0001), and other values in between.

An early proponent of this idea, Egon Brunswik, advocated for *representative design*, or finding a "representative sampling of [the] situations" in which one intends to generalize their results (Brunswik 1956, Dhami et al. 2004). Similarly, Wells and Windschitl (1999) argued that researchers should engage in *stimulus sampling*, or using a variety of stimuli within each category of interest. If one is studying the effects of gender on some phenomenon, for example, using a stimulus of one man and one woman is "functionally equivalent to conducting an experiment with a sample size of $n = 1$." Because only one task represents each category, "what might be portrayed as a category effect could in fact be due to the unique characteristics of the stimulus selected to represent that category."

By making it possible to quantify task dimensions, the Task Space extends this line of work and enables researchers to systematically choose tasks along many dimensions of interest. Rather than studying just one task, researchers can use the Task Space to "stimulus sample" across different task contexts, and to do so in a reproducible manner. Possibilities include random exploration (Baribault et al. 2018), selecting tasks that are maximally far apart, or even adaptive sampling—from using the classic Thompson sampling technique (Thompson 1933) to leveraging more recent developments in adaptive experimental design (Eyke et al. 2020, Balietti et al. 2021, Simonsohn et al. 2024).

5.1.2. Constructing and Testing "Middle Range" Theories. Our contribution also speaks to longstanding concerns in sociology about the domain specificity of theories. As Robert Merton once observed, the path to understanding general insights about the social world is not to seek a single, all-encompassing theory, but to progressively unify "theories of the middle range"—which are limited in context and scope, yet empirically testable (Merton 1949). Similar themes of understanding the influence of contextual variables recur across social science: Yarkoni (2022) discusses the need to measure "stimuli, task, [and] instructions"; Bareinboim and Pearl (2013) call such variables the "difference-generating factors." Manzi (2012) describes the world as being "causally dense," with many potential contextual factors influencing a given phenomenon. Most directly, our research builds on recent work by Almaatouq et al. (2022) and Levinthal and Rosenkopf (2020), who propose quantifying a study's defining attributes in its latent design space. Similarly, Simons et al. (2017) advocate for publishing statements of Constraints on Generality (COG) that explicitly identify the context for which findings are expected to apply. Listing one's design decisions and the implications they may have on generalizability can distinguish a bounded "middle range" theory from a replication failure.

Thus, in our domain, if the benefit of hiring an interacting group rather than a competent individual depends on whether the task involves a correct answer, the Task Map can coherently encode this idea alongside other possible moderators. This integrative potential makes the Task Map useful for enhancing meta-analysis and systematic review, which have been criticized for aggregating studies that are not "remotely comparable" (Eysenck 1994).

Moreover, examining a study's underlying dimensions may inspire more context-specific, or "solution-oriented," research (Watts 2017). Managers working in different domains may justifiably wonder whether a case study from another industry would apply to their own, or whether best practices from a traditional colocated workplace would translate into one adopting distributed, technology-mediated work. The general challenge of understanding how and why theories translate across contexts is especially salient in a world in which the context of work is constantly changing. Thus, describing prior studies in terms of their underlying dimensions—and isolating the ones that are the key "drivers" of an effect—would help practitioners identify which research studies and insights are most relevant to them. We envision the Task Space as a tool that connects theory to practice, creating theory not for theory's sake, but for producing actionable, context-sensitive insights.

5.2. Limitations and Future Directions

Although we have already shown that the Task Space, as presented, is a useful device for reconciling findings in the group and teams literature, we emphasize that this first step is associated with several limitations. For example, we chose a human annotation-based method of quantifying task dimensions, on the grounds that it captures the impressions of task attributes from the perspective of a general population; however, our framework is agnostic to the method of measurement. Future researchers can introduce new methods, including expert annotations, Elo-style head-to-head comparisons, and large language model-based annotation (see Gilardi et al. 2023). Ultimately, many of the constructs in the Task Space are subjective: for example, what qualifies as a "conflicting tradeoff" may be up to interpretation. We therefore expect that methods for measuring task dimensions, as well as their relative locations within the Task Space, will be refined over time. Indeed, just as the semantic meanings of words have shifted alongside the evolution of language and culture, so the meanings of tasks may also shift alongside the landscape of work and automation. The Task Space is designed to embrace this change.

Beyond pure measurement issues, we also acknowledge that tasks are still far more complex than we have allowed in our 24-dimensional representation. As discussed in Section 3, our choice of focusing on the task class means that the current Task Space considers only general attributes of a task's stimuli and goals. Dimensions outside of this scope include problem-solving strategies (e.g., Steiner (1972)'s *conjunctive*, *disjunctive*, and *discretionary* tasks), typical participant behaviors (Roby and Lanzetta 1958), participants' perceptions of the task (Shaw 1963, Tushman 1979), and instance-level specifics (complexity, technology affordances, incentive structures, and environmental factors). Each of these dimensions constitute design spaces in their own right. For example, participant behaviors (the "group process") are very well-studied (Barrick et al. 1998, Li et al. 2018, Oldeweme et al. 2021), with extensive reviews of subtopics, such as leadership (Carton 2022), as well as taxonomies on the subject (e.g., Marks et al. 2001). Similarly, group composition is another extensively

researched field (Bell 2007, Mathieu et al. 2014). Incorporating these richer conceptions of tasks (i.e., beyond the stimulus and goals) will entail considerable additional work. Nonetheless, we see our approach—of quantifying the design space of an empirical question—as a proof of concept that can be extended to other aspects of tasks, and even to other domains. That is, the procedure we have outlined of measuring the underlying variables in a domain is naturally extensible to far more general spaces; analogous to the Task Space that we have described here, one could also create a Process Space, Composition Space, and so on.

Another line of future work will be extending the Task Space to field research. Whereas laboratory tasks are short, self-contained, and can be completed by generalists with minimal training, these properties rarely hold for tasks performed in an organizational context. In the field, members of teams must respond to an ever-changing environment ("rugged fitness landscapes," Levinthal 1997) and perform specialized tasks embedded within long-term projects. Thus, the notion of a "task" will need to be redefined—a process that we caution will not be trivial. Among the most significant changes will be to shift the focus from class-level attributes to instance-level attributes, as well as to emphasize specialized skills. For example, O*NET (originally introduced by Peterson et al. 2001), describes a "Sales Manager" in terms of activities such as "Selling and Influencing Others" and "Establishing and Maintaining Interpersonal Relationships." The analysis by Driskell, et al. (1987) of military teams includes skill-oriented task types such as "Mechanical/Technical," "Manipulative/Persuasive," and "Logical/Precision." Wildman et al. (2012) include features such as "Managing Others," "Advising Others," and "Human Service." As with the laboratory tasks we included in this work, there exist a myriad of ways to describe skill- and role-related features, and we anticipate that building a multidimensional space of tasks for the field will involve developing a shared language for describing work activities, then integrating relevant features from across these frameworks.

A concern with this philosophy of continued expansion is that it will create many different "versions" of the Task Space, which could undermine the original goal of commensurability. While the need to narrow down to a relevant set of attributes is inevitable—it is impossible, in a "causally dense" world, to account for every conceivable variable—incompatibility is not. The Task Space, at its most fundamental level, is a way of thinking about tasks, presenting them not as discrete "types," but as points within a multidimensional space. Whereas categorical systems imply interdependencies between dimensions (e.g., "Creative" tasks are implicitly not "Intellective;" a "Maximizing" task is implicitly not "Optimizing"), multidimensional spaces make minimal assumptions. Consequently, it is easier to update a multidimensional space, because adding a new dimension does not require redefining relationships between the existing categories. Different ways of thinking about tasks can be combined by simply taking the union; and even when researchers select different subsets of task dimensions, we find in a robustness check that key relationships between tasks are generally preserved (Online Appendix I.2).

This flexibility in the Task Space allows it to serve its integrative purpose. Broadly, the integrative approach that inspires our work "emphasizes contingencies and enables practitioners to distinguish between the most general result and the result that is most useful in practice" (Almaatouq et al. 2024a). Rather than assume that a finding from a single case study extends to teams in any context, the Task Space adopts the paradigm that one should first delineate a study's fundamental features. By forming and testing theories about these features, teams working in organizations can create practices that survive rugged terrain.

Acknowledgments

The authors thank the members of the Penn Computational Social Science Lab's Advisory Committee: Eytan Bakshy, Eric Bradlow, Colin Camerer, Angela Duckworth, James Larson, Jim Manzi, Brian Uzzi, Ming Yin, and the late Daniel Kahneman. The authors are grateful for the work of Daria Paniukhina, Joyce An-Jie Wang, Sarika Subramanian, Karan Sampath, Dodge Hill, and Vikram Balasubramanian, who were research assistants instrumental in the literature review and task extraction process, and Vincent Cai, Kaily Liu, Eda Orakci, Alan Qiao, Jason Ren, Adil Shahid, Anna Vazhaeparambil, Bryan Yao, Huifang Ye, and Daniel Xue, who were research assistants involved in engineering the multi-task experiment through the Empirica platform. The authors thank Eric Shapiro for managing the software engineering components of this project, and are also grateful for research support from Wharton Computing and the Wharton Behavioral Lab. Finally, the authors thank Matthew Cronin, Mohammed Alsobay, Michael Cooper, James Houghton, Michelle Vaccaro, and Robin Na for their feedback on drafts of this work.

Endnotes

[1] In some cases, prior work refers to taxonomies, whereas in other cases it refers to typologies. Although there are subtle differences between taxonomies and typologies (see Bailey 1989, 1994), the terms are often used interchangeably. Reflecting this common practice, we will use the term "taxonomy" to refer to both concepts, noting also that the technical differences are not relevant to our framework.

[2] After completing a review of the literature on task taxonomies, we initially piloted a set of 71 questions. After a substantial iterative process to ensure reliability and applying exclusions based on the criteria discussed above (task-class focused and applicable to laboratory studies) our final set included 24 dimensions. As we have noted, we have designed the Task Space such that it is easy to omit or append additional task dimensions. Therefore, we do not make any claim that this is the final design space of all tasks, and we

welcome contributions by other researchers to further develop the design space (a topic that we will expand upon later in this paper).

[3] We derived the content of each dimension from the source papers, editing them for clarity and conciseness after pilot testing. In addition, we wrote a longer elaboration for each question that clarified common misconceptions we found during pilots, which we also displayed to raters (see Online Appendix C.2 for piloting procedure and Section C.4.3 for elaboration text).

[4] We note that the Task Map is not intended to represent tasks in general. "Representativeness" is determined by one's research question, rather than being a universal truth: for example, a researcher studying economic games may be interested in whether the Ultimatum Game is overrepresented in their field, while another studying conflict may be interested in whether distributive negotiations are more commonly studied than integrative ones. Answering such questions requires representatively sampling tasks in those respective fields. However, the method of evaluating tasks according to their dimensions, and positioning them in a "space" where one can evaluate and draw boundaries between them, remains unchanged.

[5] We have released the data, rater training materials, and data cleaning scripts associated with the Task Map on both OSF (https://osf.io/4pftv/) and GitHub (https://github.com/Watts-Lab/task-mapping). In addition, readers can interact with our data on our public website (taskmap.seas.upenn.edu).

[6] All workers were heavily vetted through a three-stage process, in which they were screened for reading comprehension abilities, provided with step-by-step interactive training, and required to pass a final qualification quiz (see Online Appendix C.4.1).

[7] According to the literature on the wisdom of crowds and prediction markets, aggregating independent opinions generates accurate estimates for the true value of a quantity, and the arithmetic mean is a robust pooling function (Chen et al. 2005, Arrow et al. 2008, Chen and Li 2012).

References

Aggarwal I, Wooley AW (2019) Team creativity, cognition, and cognitive style diversity. *Management Sci.* 65(4):1586–1599.

Allport GW, Odbert HS (1936) Trait-names: A psycho-lexical study. *Psych. Monographs* 47(1):i–171.

Almaatouq A, Alsobay M, Yin M, Watts DJ (2021a) Task complexity moderates group synergy. *Proc. Natl. Acad. Sci. USA* 118(36): e2101062118.

Almaatouq A, Alsobay M, Yin M, Watts DJ (2024c) The effects of group composition and dynamics on collective performance. *Topics Cognitive Sci.* 16(2):302–321.

Almaatouq A, Amin Rahimian M, Burton JW, Alhajri A (2022) The distribution of initial estimates moderates the effect of social influence on the wisdom of the crowd. *Sci. Rep.* 12(1):16546.

Almaatouq A, Becker J, Houghton JP, Paton N, Watts DJ, Whiting ME (2021b) Empirica: A virtual lab for high-throughput macro-level experiments. *Behav. Res. Methods* 53(5):2158–2171.

Almaatouq A, Griffiths TL, Suchow JW, Whiting ME, Evans J, Watts DJ (2024a) Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Brain Behav. Sci.* 47:1–55.

Almaatouq A, Griffiths TL, Suchow JW, Whiting ME, Evans J, Watts DJ (2024b) Replies to commentaries on beyond playing 20 questions with nature. *Brain Behav. Sci.* 47:e65.

Almaatouq A, Noriega-Campero A, Alotaibi A, Krafft PM, Moussaid M, Pentland A (2020) adaptive social networks promote the wisdom of crowds. *Proc. Natl. Acad. Sci. USA* 117(21): 11379–11386.

Altmann A, Toloşi L, Sander O, Lengauer T (2010) Permutation importance: A corrected feature importance measure. *Bioinformatics* 26(10):1340–1347.

Arrow KJ, Forsythe R, Gorham M, Hahn R, Hanson R, Ledyard JO, Levmore S, et al. (2008) The promise of prediction markets. *Science (1979)* 320(5878):877–878.

Asch S (1956) Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psych. Monographs General Appl.* 70(9):1–70.

Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD (2010) Optimally interacting minds. *Science (1979)* 329(5995):1081–1085.

Bailey KD (1989) Taxonomies and disaster: Prospects and problems. *Internat. J. Mass Emerg. Disasters* 7(3):419–431.

Bailey KD (1994) *Typologies and Taxonomies in Social Science: An Introduction to Classification Techniques*, Quantitative Applications in the Social Sciences, Series No. 07-102 (Sage, Thousand Oaks, CA).

Balietti S, Klein B, Riedl C (2021) Optimal design of experiments to identify latent behavioral types. *Exp. Econom.* 24(3):772–799.

Bareinboim E, Pearl J (2013) A general algorithm for deciding transportability of experimental results. *J. Causal Inference* 1(1):107–134.

Baribault B, Donkin C, Little DR, et al. (2018) Metastudies for robust tests of theory. *Proc. Natl. Acad. Sci. USA.* 115(11):2607–2612.

Baron J, Hershey JC (1988) Outcome bias in decision evaluation. *J. Personality Soc. Psych.* 54(4):569–579.

Barrick MR, Bradley BH, Kristof-Brown AL, Colbert AE (2007) The moderating role of top management team interdependence: Implications for real teams and working groups. *Acad. Management J.* 50(3):544–557.

Barrick MR, Stewart GL, Neubert MJ, Mount MK (1998) Relating member ability and personality to work-team processes and team effectiveness. *J. Appl. Psych.* 83(3):377–391.

Batista RM, Ross J (2024) Words that work: Using language to generate hypotheses. Preprint, submitted September 16, http://dx.doi.org/10.2139/ssrn.4926398.

Bavelas A (1950) Communication patterns in task-oriented groups. *J. Acoust. Soc. Amer.* 22(August):725–730.

Becker J, Almaatouq A, Horvát E-Á (2021) Network structures of collective intelligence: The contingent benefits of group discussion. Preprint, submitted March 8, http://arxiv.org/abs/2009.07202.

Bell ST (2007) Deep-level composition variables as predictors of team performance: A meta-analysis. *J. Appl. Psych.* 92(3): 595–615.

Brunswik E (1956) Representative design and probabilistic theory in a functional psychology. *Psych. Rev.* 62(3):193–217.

Camerer CF (1997) Progress in behavioral game theory. *J. Econom. Perspect.* 11(4):167–188.

Carton AM (2022) The science of leadership: A theoretical model and research agenda. *Annual Rev. Organ. Psych. Organ. Behav.* 9(1):61–93.

Chang EH, Kirgios EL, Smith RK (2021) Large-scale field experiment shows null effects of team demographic diversity on outsiders' willingness to support the team. *J. Experiment. Soc. Psych.* 94(May):104099.

Chen W, Li X (2012) Deciphering wisdom of crowds from their influenced binary decisions. *Proc. 2012 IEEE Internat. Conf. Intelligence Security Informatics* (IEEE, Washington, DC), 235–240.

Chen Y, Chu C-H, Mullen T, Pennock DM (2005) Information markets vs opinion pools: An empirical comparison. *Proc. ACM* Conf. Electronic Commerce (ACM, New York), 58–67.

Cohen SG, Bailey DE (1997) What makes teams work: Group effectiveness research from the shop floor to the executive suite. *J. Management* 23(3):239–290.

De Dreu CKW, Weingart LR (2003) Task versus relationship conflict, team performance, and team member satisfaction: A meta-analysis. *J. Appl. Psych.* 88(4):741–749.

Dhami MK, Hertwig R, Hoffrage U (2004) The role of representative design in an ecological approach to cognition. *Psych. Bull.* 130(6):959–988.

Diehl M, Stroebe W (1987) Productivity loss in brainstorming groups: Toward the solution of a riddle. *J. Personality Soc. Psych.* 53(3):497–509.

Driskell JE, Salas E, Hogan R (1987) *A Taxonomy for Composing Effective Naval Teams* (Naval Training Systems Center, Orlando, FL).

Epley N, Gilovich T (2006) The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psych. Sci.* 17(4):311–318.

Ericksen J, Dyer L (2004) Right from the start: Exploring the effects of early team events on subsequent project team development and performance. *Admin. Sci. Quart.* 49(3):438–471.

Eyke NS, Green WH, Jensen KF (2020) Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *Reaction Chemistry Engrg.* 5(10):1963–1972.

Eysenck HJ (1994) Systematic reviews: Meta-analysis and its problems. Education and debate. *BMJ* 309(6957):789–792.

Fleishman EA (1975) Toward a taxonomy of human performance. *Amer. Psych.* 30(12):1127–1149.

Gilardi F, Alizadeh M, Kubli M (2023) ChatGPT outperforms crowd-workers for text-annotation tasks. Preprint, submitted March 27, http://arxiv.org/abs/2303.15056.

Goldberg LR (1993) The structure of phenotypic personality traits. *Amer. Psych.* 48(1):26–34.

Goodwin GF, Blacksmith N, Coats MR (2018) The science of teams in the military: Contributions from over 60 years of research. *Amer. Psych.* 73(4):322–333.

Graf-Drasch V, Gimpel H, Barlow JB, Dennis AR (2022) Task structure as a boundary condition for collective intelligence. *Personnel Psych.* 75(3):739–761.

Gross E (1954) Primary functions of the small group. *Amer. J. Sociol.* 60(1):24–29.

Gully SM, Incalcaterra KA, Joshi A, Beaubien JM (2002) A meta-analysis of team-efficacy, potency, and performance: Interdependence and level of analysis as moderators of observed relationships. *J. Appl. Psych.* 87(5):819–832.

Hackman JR (1968) Effects of task characteristics on group products. *J. Experiment. Soc. Psych.* 4(2):162–187.

Hackman JR (1969) Towards understanding the role of tasks in behavioral research. *Acta Psychologica* 31(2):97–128.

Hackman JR, Oldham GR (1975) Development of the job diagnostic survey. *J. Appl. Psych.* 60(2):159–170.

Harris AM, Gómez-Zará D, DeChurch LA, Contractor NS (2019) Joining together online: The trajectory of CSCW scholarship on group formation. *Proc. ACM Human Comput. Interact.* 3(CSCW):1–27.

Herold DM (1978) Improving the performance effectiveness of groups through a task-contingent selection of intervention strategies. *Acad. Management Rev.* 3(2):315–325.

Hill GW (1982) Group versus individual performance: Are $N + 1$ heads better than one? *Psych. Bull.* 91(3):517–539.

Hollenbeck JR, Beersma B, Schouten ME (2012) Beyond team types and taxonomies: A dimensional scaling conceptualization for team description. *Acad. Management Rev.* 37(1):82–106.

Hough LM (1992) The 'big five' personality variables—Construct confusion: Description versus prediction. *Human Perform.* 5(1&2): 139–155.

Hueffmeier J, Hertel G (2011) When the whole is more than the sum of its parts: Group motivation gains in the wild. *J. Experiment. Soc. Psych.* 47(2):455–459.

Husband RW (1940) Cooperative versus solitary problem solution. *J. Soc. Psych.* 11(2):405–409.

Janis IL (1971) *Groupthink.* Psychology Today (Ziff-Davis Publishing: New York), 1–7.

Koriat A (2012) When are two heads better than one and why? *Science (1979)* 336(6079):360–362.

Krackhardt D, Stern RN (1988) Informal networks and organizational crises: An experimental simulation. *Social Psych. Quart.* 51(2):123–140.

Larson JR (2010) *In Search of Synergy in Small Group Performance* (Psychology Press, New York).

Laughlin PR, Ellis AL (1986) Demonstrability and social combination processes on mathematical intellective tasks. *J. Experiment. Social Psych.* 22(3):177–198.

Laughlin PR, Hatch EC, Silver JS, Boh L (2006) Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *J. Personality Soc. Psych.* 90(4):644–651.

Laughlin PR, Kerr NL, Davis JH, Halff HM, Marciniak KA (1975) Group size, member ability, and social decision schemes on an intellective task. *J. Personality Soc. Psych.* 31(3):522–535.

Lee SM, Koopman J, Hollenbeck JR, Wang LC, Lanaj K (2015) The Team Descriptive Index (TDI): A multidimensional scaling approach for team description. *Acad. Management Discoveries* 1(1):91–116.

Levinthal DA (1997) Adaptation on rugged landscapes. *Management Sci.* 43(7):934–950.

Levinthal DA, Rosenkopf L (2020) Commensurability and collective impact in strategic management research: When non-replicability is a feature, not a bug. Accessed November 13, 2022, https://mackinstitute.wharton.upenn.edu/2020/commensurability-and-collective-impact-in-strategic-management-research/.

Li G, Rubenstein AL, Lin W, Wang M, Chen X (2018) The curvilinear effect of benevolent leadership on team performance: The mediating role of team action processes and the moderating role of team commitment. *Personnel Psych.* 71(3):369–397.

Lorge I, Solomon H (1960) Group and individual performance in problem solving related to previous exposure to problem, level of aspiration, and group size. *Behav. Sci.* 5(1):28–38.

Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. *Proc. 31st Conf. Neural Inform. Processing Systems* (NIPS, Long Beach, CA), 4768–4777.

Manzi J (2012) *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society* (Basic Books, New York).

Mao A, Mason W, Suri S, Watts DJ (2016) An experimental study of team size and performance on a complex task. *PLoS One* 11(4):e0153048.

Marks MA, Mathieu JE, Zaccaro SI (2001) A temporally based framework and taxonomy of team processes. *Acad. Management Rev.* 26(3):356–376.

Marquart DI (1955) Group problem solving. *J. Social Psych.* 41(1): 103–113.

Mason W, Watts DJ (2011) Collaborative learning in networks. *Proc. Natl. Acad. Sci. USA* 109(3):764–769.

Mathieu JE, Hollenbeck JR, van Knippenberg D, Ilgen DR (2017) A century of work teams in the Journal of Applied Psychology. *J. Appl. Psych.* 102(3):452–467.

Mathieu JE, Tannenbaum SI, Donsbach JS, Alliger GM (2014) A review and integration of team composition models: Moving toward a dynamic and temporal framework. *J. Management* 40(1): 130–160.

McCormick EJ, Jeanneret PR, Mecham RC (1972) A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). *J. Appl. Psych.* 56(4):347–368.

McGrath JE (1984) *Groups: Interaction and Performance* (Prentice-Hall, Inc, Engelwood Cliffs, NJ).

Meluso J, Hébert-Dufresne L (2023) Multidisciplinary learning through collective performance favors decentralization. *Proc. Natl. Acad. Sci. USA* 120(34):e2303568120.

Merton RK (1949) On sociological theories of the middle range. Calhoun C, Gerteis J, Moody J, Pfaff S, and Indermohan V, eds. *Classical Sociological Theory*, 2nd ed. (Blackwell Pub, Malden, MA), 448–459.

Meslec N, Curseu PL, Meeus MTH, Iederan Fodor OC (2014) When none of us perform better than all of us together: The role of analogical decision rules in groups. *PLoS One* 9(1): e85232.

Morris CG (1966) Task effects on group interaction. *J. Personality Soc. Psych.* 4(5):545–554.

Mullen B, Johnson C, Salas E (1991) Productivity loss in brainstorming groups: A meta-analytic integration. *Basic Appl. Soc. Psych.* 12(1):3–23.

Oldeweme M, Konradt U, Brede M (2021) The rhythm of teamwork: Discovering a complex temporal pattern of team processes. *Group Dynam. Theory Res. Practice* 27(1):50–64.

Peterson RS, Owens PD, Tetlock PE, Fan ET (1998) Group dynamics in top management teams: Groupthink, vigilance, and alternative models of organizational failure and success. *Organ. Behav. Human Decision Processes* 73(2/3):272–305.

Peterson NG, Mumford MD, Borman WC, Richard Jeanneret P, Fleishman EA, Levin KY, Campion MA, et al. (2001) Understanding work using the Occupational Information Network (O*NET): Implications for practice and research. *Personnel Psych.* 54(2):451–492.

Riedl C, Kim YJ, Gupta P, Malone TW, Woolley AW (2021) Quantifying collective intelligence in human groups. *Proc. Natl. Acad. Sci. USA* 118(21):e2005737118.

Roby TB, Lanzetta JT (1958) Considerations in the analysis of group tasks. *Psych. Bull.* 55(2):88–101.

Salas E, Reyes DL, McDaniel SH (2018) The science of teamwork: Progress, reflections, and the road ahead. *Amer. Psych.* 73(4): 593–600.

Shaw ME (1963) *Scaling Group Tasks: A Method for Dimensional Analysis* (American Psychological Association).

Shore J, Bernstein E, Lazer D (2015) Facts and figuring: An experimental investigation of network structure and performance in information and solution spaces. *Organ. Sci.* 26(5):1432–1446.

Silver I, Mellers BA, Tetlock PE (2021) Wise teamwork: Collective confidence calibration predicts the effectiveness of group discussion. *J. Experiment. Soc. Psych.* 96(September):104157.

Simons DJ, Shoda Y, Lindsay DS (2017) Constraints on generality (COG): A proposed addition to all empirical papers. *Perspect. Psych. Sci.* 12(6):1123–1128.

Simonsohn U, Montealegre A, Evangelidis I (2024) Stimulus sampling reimagined: Designing experiments with mix-and-match, analyzing results with stimulus plots. *J. Personality Soc. Psych.* 129(1):71–90.

Stasser G, Titus W (1985) Pooling of unshared information in group decision making: Biased information sampling during discussion. *J. Personality Soc. Psych.* 48(5):1467–1478.

Stasson MF, Bradshaw SD (1995) Explanations of individual-group performance differences: What sort of 'bonus' can be gained through group interaction? *Small Group Res.* 26(2): 296–308.

Steiner ID (1972) *Group Process and Productivity* (Academic Press, New York).

Stewart GL, Barrick MR (2000) Team structure and performance: Assessing the mediating role of intrateam process and the moderating role of task type. *Acad. Management J.* 43(2):135–148.

Straub VJ, Tsvetkova M, Yasseri T (2023) The cost of coordination can exceed the benefit of collaboration in performing complex tasks. *Collective Intelligence* 2(2):263391372311569.

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3–4):285–294.

Thompson LL, Wilson ER (2015) Creativity in teams. Thompson LL, Choi HK, eds. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource* (Lawrence Erlbaum Associates, Inc, Mahwah, NJ).

Tushman ML (1979) Work characteristics and subunit communication structure: A contingency analysis. *Admin. Sci. Quart.* 24(1): 82–98.

Valentine MA, Nembhard IM, Edmondson AC (2015) Measuring teamwork in health care settings: A review of survey instruments. *Med. Care* 53(4):e16–e30.

Watts DJ (2017) Should social science be more solution-oriented? *Nat. Human Behav.* 1(1):0015.

Weber B, Hertel G (2007) Motivation gains of inferior group members: A meta-analytical review. *J. Personality Soc. Psych.* 93(6):973–993.

Weidmann B, Deming DJ (2021) Team players: How social skills improve team performance. *Econometrica* 89(6):2637–2657.

Wells GL, Windschitl PD (1999) Stimulus sampling and social psychological experimentation. *Personality Social Psych. Bull.* 25(9):1115–1125.

Whiting ME, Blaising A, Barreau C, Fiuza L, Marda N, Valentine M, Bernstein MS (2019) Did it have to end this way? Understanding the consistency of team fracture. *Proc. ACM Human Comput. Interact.* 3(CSCW):1–23.

Wildman JL, Thayer AL, Rosen MA, Salas E, Mathieu JE, Rayne SR (2012) Task types and team-level attributes: Synthesis of team classification literature. *Human Resource Development Rev.* 11(1): 97–129.

Wood RE (1986) Task complexity: Definition of the construct. *Organ. Behav. Human Decision Processes* 37(1):60–82.

Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW (2010) Evidence for a collective intelligence factor in the performance of human groups. *Science (1979)* 330(6004):686–688.

Yarkoni T (2022) The generalizability crisis. *Behav. Brain Sci.* 45:e1.

Zigurs I, Buckland BK, Connolly JR, Wilson EV (1999) A test of task-technology fit theory for group support systems. *ACM SIGMIS Database Adv. Inform. Systems* 30(3–4):34–50.