The Task Space: An Integrative Framework for Team Research

Appendices and Supplementary Materials

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

[a] Operations, Information, and Decisions Department, The Wharton School, University of Pennsylvania

[b] Institute for Data, Systems, and Society, Massachusetts Institute of Technology

[c] Pareto

[d] Department of Computer and Information Science, University of Pennsylvania

[e] Annenberg School for Communication, University of Pennsylvania

[f] Sloan School of Management, Massachusetts Institute of Technology

[g] Center for Computational Engineering, Massachusetts Institute of Technology

* Correspondence to: xehu@mit.edu; amaatouq@mit.edu

List of Appendices

Appendix A: 15 Task Frameworks Proposed in The Behavioral and Organizational Sciences

Table 1: Review of 15 Task Frameworks

No.	*Citation*	*Task Categories or Dimensions*
1	Roby and Lanzetta (1958)	(1) Orientation: the ability to take stock of all the variables in the task environment, and the group's standing in relation to those variables (2) Mapping: anticipating or learning the consequences of various action alternatives (3) Jurisdiction: choosing and implementing decisions
2	Shaw (1963)	(1) Amount of effort required (2) Number of specialized operations, skills, and types of knowledge required (3) Clarity of the stated task (4) Degree to which there is more than one correct solution (5) Degree to which there is more than one way to reach a correct solution (6) Degree to which there is immediate feedback about whether the solution is correct (7) Degree to which group members' actions are integrated (8) Ratio of mental to motor requirement (9) Degree to which task is familiar to members of society (10) Degree to which task is interesting
3	Hackman (1968a)	(1) Production (producing ideas, images, or arrangements) (2) Discussion (discussing values or issues) (3) Problem-Solving (requiring that a solution to a problem is worked out)
4	Steiner (1972)	(1) Subtask structure (Divisible v. Unitary): (a) Divisible: easily-separated parts that require different skills or abilities (b) Unitary: cannot be meaningfully subdivided into separate parts (2) Nature of the goal (Maximizing v. Optimizing): (a) Maximizing: goal is to do something as much or as quickly as possible (b) Optimizing: production of a specific, desirable product (3) Permitted Group Processes (Additive, Conjunctive, Disjunctive, Discretionary): (a) Additive: product involves summing together people's efforts (b) Conjunctive: group members contribute identically to the product (c) Disjunctive: one member's contribution determines the group's product (d) Discretionary: no constraints on how member contributions are combined
5	Fleishman (1975)	(1) Ability Requirements: 11 Perceptual-Motor Abilities: (a) Control precision (b) Multilimb coordination (c) Response orientation (d) Reaction time (e) Speed of arm movement (f) Rate control (timing) (g) Manual dexterity (h) Finger dexterity (i) Arm-hand steadiness (j) Wrist-finger speed (k) Aiming 9 Physical Proficiency Abilities: (a) Extent flexibility (b) Dynamic flexibility (c) Static strength (d) Dynamic strength (e) Explosive strength (f) Trunk strength (g) Gross body coordination (h) Equilibrium (i) Stamina (2) Task Requirements: (a) Stimulus duration (b) Number of output units (c) Duration for which output unit is maintained

Table 1: Review of 15 Task Frameworks

No.	*Citation*	*Task Categories or Dimensions*
		(d) Simultaneity of responses (e) Number of procedural steps (f) Variability of stimulus location
6	McCormick et al. (1972)	(1) Overall: Decision/communication/social responsibilities, skilled activities, physical activities, equipment/vehicle cooperation, information processing activities (2) Visual Input: Visual input from devices/materials, perceptual interpretation, information from people, visual input from distal sources, evaluation of information from distal sources, environmental awareness, awareness of body movement/posture (3) Mediation Processes: decision making, information processing (4) Work Output: Machine/process control, Manual control/coordination activities, control/equipment operation, general body activity, handling/manipulating activities, use of finger-controlled devices versus physical work, skilled/technical activities (5) Interpersonal Activities: Communication of decisions/judgments, job-related information exchange, staff/related activities, supervisor-subordinate relationships, public/related contact (6) Work Situation and Job Context: unpleasant/hazardous physical environment, personally demanding situations (7) Miscellaneous Aspects: businesslike work situations, attentive/discriminating work demands, unstructured v. structured work, variable v. regular work schedule
7	Hackman and Oldham (1975)	(1) Skill variety: the degree to which a job requires a variety of different activities in carrying out the work, which involve many skills/talents (2) Task identity: the degree to which the job requires completion of a "whole" and identifiable piece of work (3) Task significance: the degree to which the job has a substantial impact on the lives or work of other people (4) Autonomy: the degree to which the job provides substantial freedom, independence, and discretion to the employee (5) Feedback from the job itself: the degree to which carrying out the work activities required by the job results in the employee obtaining direct and feedback on their performance (6) Feedback from agents: the degree to which the employee receives clear information about their performance from their supervisors or coworkers (7) Dealing with others: the degree to which the job requires the employee to work closely with other people
8	Tushman (1979)	(1) Basic research: work of a general nature intended to have broad applications (2) Applied research: work involving basic knowledge for the solution of a specific problem (3) Development: work involving the combination of existing or feasible concepts to create a distinctly new product (4) Technical service: work involve cost/performance improvement of existing produces, processes, or systems Also describes task features that motivated the categories: - Degree of predictability (routineness) versus uncertainty - Degree of complexity (information-processing requirements) - Degree of interdependence (coordination, joint decision-making, and problem-solving) - Time span of feedback - Specific versus general problem orientation - Generation of new knowledge versus using existing knowledge
9	McGrath (1984)	Quadrant I: (1) Planning: generating an action-oriented plan (2) Creativity: generating ideas Quadrant II: (3) Intellective: solving problems with a correct answer (4) Decision-Making: tasks in which the preferred or agreed-upon answer is the right one Quadrant III: (5) Cognitive Conflict: resolving conflicts of viewpoint (6) Mixed-Motive: resolving conflicts of motive or interest Quadrant IV: (7) Contests/Battles/Competitive: resolving conflicts of power (8) Performance/Psycho-Motor: tasks performed against objective or absolute standards of excellence
10	Wood (1986)	(1) Component complexity: a function of the number of distinct acts that need to be executed (2) Coordinative complexity: the nature of relationships between task inputs and products (e.g., timing, frequency, intensity) (3) Dynamic complexity: changes in the states of the world which have an effect on the relationships between task inputs and products

Table 1: Review of 15 Task Frameworks

No.	*Citation*	*Task Categories or Dimensions*
11	Laughlin and Ellis (1986)	(1) Group shares a common verbal or mathematical system (2) There is sufficient information for a solution within the system (3) Group members who are not themselves able to solve the problem must have sufficient knowledge of the system to recognize and accept a correct solution (4) A person with the correct answer can demonstrate its correctness to others
12	Driskell et al. (1987)	(1) Mechanical/Technical: construction, operation, or maintenance of things (2) Intellectual/Analytic: generation, exploration, or verification of knowledge (3) Imaginative/Aesthetic: invention, arrangement, or production of expressive products (4) Social: training, assisting, or serving others (5) Manipulative/Persuasive: organization, motivation, or persuasion of others (6) Logical/Precision: performance of explicit, routine tasks or tasks requiring attention to detail
13	Zigurs et al. (1999)	(1) Outcome Multiplicity: the task has more than one desired outcome (2) Solution Scheme Multiplicity: there is more than one course of action to attain the group's goal (3) Conflicting Interdependence: the adoption of one possible solution scheme conflicts with the adoption of another (4) Solution Scheme Outcome Uncertainty: the extent to which there is uncertainty about whether a given solution will lead to a desired outcome
14	N. G. Peterson et al. (2001)	(1) Worker Characteristics (Abilities, Occupational Values and Interests, Work Styles) (2) Occupation Characteristics (Labor Market Information, Occupational Outlook, Wages) (3) Occupation-Specific Requirements (Occupational Skills, Knowledge; Tasks, Duties; Machines, Tools, and Equipment) (4) Worker Requirements (Basic Skills, Cross-Functional Skills, Knowledge, Education) (5) Experience Requirements (Training, Experience, Licensure) (6) Occupational Requirements (Generalized Work Activities, Work Context, Organizational Context) Example Abilities: - Cognitive abilities: verbal, idea generation and reasoning, quantitative, memory, perceptual, spatial, attentiveness - Psychomotor abilities: fine manipulative, control movement, reaction time and speed, physical strength, endurance, flexibility/balance/coordination, sensory Example Skills: - Content skills: active listening, reading comprehension, writing, speaking, mathematics, science - Process skills: active learning, learning strategies, monitoring, critical thinking - Problem-solving skills: problem identification, information gathering, information organization, synthesis/reorganization, idea generation, idea evaluation, implementation planning, solution appraisal - Social skills: social perceptiveness, coordination, persuasion, negotiation, instructing, service orientation
15	Wildman et al. (2012)	(1) Managing Others: directing, supervising, or overseeing the work of others (2) Advising Others: providing professional support, such as expertise or advise (3) Human Service: social interaction where an individual is providing a good or service to another party (4) Negotiation: social interaction in which two or more parties in conflict seek to resolve differences and reach agreement (5) Psychomotor Action: technical and/or motor functioning requiring psychological processing to perform calculated or elaborate movements (e.g., of a product, object, or machine) (6) Defined Problem Solving: problem-solving tasks with predetermined or conclusive solutions or correct answers (7) Ill-Defined Problem Solving: problem-solving tasks lacking predetermined or conclusive solutions or correct answers (e.g., idea, plan, or knowledge generation)

Table 1. 15 task frameworks proposed in the behavioral and organizational sciences. These 15 frameworks are not meant to be systematic or comprehensive, but they serve to illustrate the extent of the problem: that there are numerous competing (and at times conflicting) methods for categorizing team tasks.

Appendix B: Tasks Included in the Task Map

Table 2: 102 Tasks in Task Map			
Task Name	**Source**	**Stimulus Description**	**Goal Description**
9 Dot Problem	Maier 1930	Participants are given a 3x3 grid of nine dots, which is shared by everyone completing the task.	Within a limited amount of time, participants must connect all nine dots (e.g., draw a line that passes through them) using no more than four lines, and may not retrace any lines. Participants will be able to visually tell that they have successfully met these criteria. They are given full credit if they accomplish this task, and no credit if they fail. An important note is that there is really only one "correct" answer to this question — that is, the way to connect the nine dots actually requires the participant to go "outside" the boundaries of the 3x3 grid, or else it cannot be done. The true goal of this task is to find this optimal answer. Participants who do not think "outside the box" therefore usually have difficulty completing this task. Once they are shown that the answer involves going outside of the box, they usually have an "aha" moment and do well on the task in the future.
Abstract grid task	Adams et al. 2021	Participants are given a shared grid, in which some squares in the grid are colored in, and other squares in the grid are not colored in. The grid is interactive, and offers the opportunity to change the squares' colors by either "adding" color where a square is not filled in, or "subtracting" color where a square is already filled in (removing the color from the square). Clicking a particular square on the grid "toggles" the color — shifting it from colored to non-colored, or vice versa.	The goal is to make the grid appear exactly symmetrical from left-to-right and top-to-bottom using the fewest number of clicks. (Each click represents one manipulation — either adding or subtracting color from the grid.) Each problem has a known "lowest" number of clicks, which represents the best possible answer. However, participants are not given any feedback about whether their answer was the best. Finally, participants must complete the task within the time limit.
Advertisement writing	Whiting et al. 2019	Participants were given the description of a product from Kickstarter. For example, they could read about "Soapstone: the Sustainable Travel Soap Dispenser," and learn about its features.	The goal was to write an online text advertisement of no more than 30 characters for the Kickstarter product. The advertisement had to be creative, informative, and get people to click. When working in teams, participants needed to submit a single collective writeup, produced within the time limit.
Aerospace Investment (Role-playing)	Sanpietro 2019	Participants are randomly paired for a scored negotiation over a venture capital investment. Each person is randomly assigned the role of either the venture capitalist investor or the founder of the aerospace company. Both individuals are scored on their ability to negotiate favorable investment terms for themselves and on the quality of the relationship they develop with their potential business partner. Each negotiator's Total Score is a sum of Substantive Points, awarded according to the agreed-upon terms of the investment, and Process Points, awarded according to each partner's perception of the negotiation process. Specifically, each negotiable term of the investment is quantified as specific point values. After the term sheet has been agreed upon, participants independently fill out questionnaires that ask them to evaluate each other on five attributes to determine the amount of Process Points they are awarded. The five process attributes are: trust, respect, equitability, regard for the other's interests, and interest in future collaboration. These attributes serve as proxies for assessing the future of the business relationship based on their experiences during the negotiation.	Participants are explicitly instructed that their goal is to maximize their Total Scores (Substantive Points + Process Points), and are given a limited amount of time (e.g., 30 minutes) to conduct their negotiation. The total score reflects the overall success of balancing both substantive and relationship concerns. For instance, the founders may be able to negotiate all the substantive terms in their favor, but if their potential business partners have no desire to work with them again, they would have left significant value at the negotiating table, and this will be reflected in a lower Total Score.

Table 2: 102 Tasks in Task Map

Allocating resources to programs	Whiting et al. 2019	Participants are given a list of complex competing programs, along with details of those programs. All participants get a list with the same programs. For example, participants could receive the following list: (1.) To purchase a new computer system for the county government in order to hold local taxes constant. (2.) To establish a community arts program featuring art, music, and dance programs for children and adults. (3.) To establish an additional shelter for the homeless in the community.	The goal is to collectively decide how to allocate $500,000 between three competing programs within the time limit. As each program is designed to appeal to personal values, there is no right, wrong, or optimal way of allocation, and the purpose of the task is to resolve different conflicts in people's values and interests.
Apache helicopter flight simulator (Longbow2)	Marks et al. 2002	Participants are randomly assigned to groups of three to play the Apache helicopter flight simulator game called Longbow2. Each person is randomly assigned the role of either the pilot (person flying the aircraft), the gunner (person operating the weapons system), or the radar specialist (person monitoring and interpreting radar systems containing critical enemy information). Participants are given a limited amount of time to complete various tasks (e.g. destroying enemy tanks) and stay alive. Before performing the task, they would receive training in role-specific task skills.	The goal is to maximize the score of the team. The score is calculated from the sum of the number of targets acquired during the course of each mission and the team's average "alive time". Team performance scores could range from 0 (no targets killed, helicopter killed by enemies immediately) to 30 (killed 15 targets, stayed alive for the duration of the 15-minute mission).
Architectural design task	Woolley et al. 2010	Participants receive written and video instructions of a task to design a house, a garage, and a pool in a computer program. Within the program, they have a limited number of building blocks of different color and size to use. For example, participants can view a table of available materials, in which they can see twenty-seven 1x1 red blocks, twelve 2x4 blue blocks, zero 1x6 orange blocks, etc. Participants receive scoring information for each object. For example, participants can see that each square foot used to build a pool will bring them a $50 bonus and that following certain design restrictions (e.g., using certain colors for the pool and making it symmetric) will give them a $1500 bonus. Participants also receive rules to follow for building objects. For example, a pool must include a diving board and have blue floors.	The goal is to design a house, a garage, and a pool of as high value as possible. Three objects must be designed from a limited set of building blocks within a fixed amount of time. The task can only be considered completed if participants adhere to the provided rules for each building. If participants don't follow the building rules, their design project fails. In other words, participants have to maximize the value of objects they are designing but they have to do so in a specific way, balancing tradeoffs of following the building rules and adding value to each object. Within these restrictions, however, they are free to be creative and produce any design they wish.
Arithmetic problem 1	Shaw 1963	Participants see a hypothetical story problem about five people who operate five machines. The problem reads as follows: You are a five man team whose job it is to manufacture a product, the completion of which requires the operation of five machines. In the past, you have rotated positions to avoid boredom, but each man has spent most of the time operating the machine that he prefers. John prefers machine 3, Steve machine 2, Walt machine 4, Robert machine 1, and Dennis machine 5. The Methods man has been around checking the time each man requires to complete the operation on one product when he is operating each of the five machines. He has come up with the following results: Machine 1 Machine 2 Machine 3 Machine 4 Machine 5 John 3 min 3 min 4 min 3.5 min 4.5 min Steve 2 min 2 min 5 min 2.5 min 3.5 min Walt 1 min 2 min 5 min 2 min 1.5 min Robert 4 min 1 min 3 min 3.5 min 3 min Dennis 5 min 3 min 2 min 5 min 3 min	The goal is to pair each man with one of the machines, such that the amount of time it takes to manufacture the product is minimized. There is exactly one right answer to this problem — it is to put Walt on machine 5, Steve on machine 4, Dennis on machine 3, Robert on machine 2, and John on machine 1. In the optimal answer, the total time to create the product is 10 minutes. Participants have a limited amount of time to submit their answers. Those working in groups can discuss their answers before submitting in an online portal. The task is graded based on how close a submission is to the true minimum time for creating the product. Submitting the correct answer (10 minutes) will give participants the highest possible score; a solution that builds the product in 11 minutes will get more points than a solution that builds the product in 15 minutes.

Table 2: 102 Tasks in Task Map

		Your foreman has noticed that when each man runs the machine that he most prefers, the total time spent on each product is 16 minutes. It seems to him that a different method of operation would result in substantial savings. He believes in letting his workers make their own decisions, as far as possible, and has asked that you consider the problem and try to come up with a plan that will be more efficient than the present mode of operation.	
Arithmetic problem 2	Shaw 1963	All group members receive different clues and have to collaborate to reach a solution to an arithmetic problem. The problem involves figuring out what time a plane arrives at its final destination given the set of clues. No one of the participants can solve the problem alone as no one has all the information needed to solve it. The clue is randomly distributed among group members and involves essential information such as plane stopover time, the distance between the source and destination, and the speed of the plane.	The goal is to solve the problem in the shortest possible time. In order to do so, they should communicate with each other about their clues in order to reach solutions. They must give an answer within the time limit.
Battle of the sexes	Rabin 1993	Participants are randomly paired with a different anonymous opponent for a predetermined number of decision-making rounds. In each round, both players must make a decision. The final outcome for the players is determined by the decision of both parties. Throughout the rounds, the players have access to a shared table that displays the possible outcomes, depending on each player's decision. For example, suppose that each player has 3 options. Then the rows might represent Player 1's options, and the columns might represent Player 2's options. There would be 3x3 = 9 possible outcomes, and the 9 cells of the table would contain the number of points that the players would earn based on their choices. For example, if Player 1 chooses their first option and Player 2 chooses their second option, then the final outcome would be determined by the number of points listed in row 1, column 2. Some outcomes are better than others. Thus, in order to get a better outcome, players have the option to use non-binding one-way communication to signal that they may choose a particular row or column, so that the other player can attempt to coordinate. The process of communication is as follows: • The participant is asked if they would like to make an announcement. They may select either yes or no. • If they select yes, they will be asked what they would like to announce. They may select any of the actions that will be available to them during the decision-making period (for example, "I intend to select option 2"). The opposing participant will then be told of their selection. • If they select no, the opposing participant will be told an announcement was not made. Players can choose to communicate one or more times. After all communications are complete, the participants will then independently select their actions, and each will receive the number of points corresponding to this combination of actions in their table. Players are not required to actually select the action they previously announced.	The goal for each participant is to maximize their earnings (points) through the decision-making rounds. Therefore, they should wisely make announcements and selections in order to obtain the highest value in the payoff table. The best outcome is when both players make choices that lead to the highest possible payoff. In all rounds, participants are required to make choices for their announcements (if applicable) and their actions in a fixed amount of time.

Table 2: 102 Tasks in Task Map

Best job candidate (hidden-profile)	Schulz-Hardt et al. 2006	Participants read information about three candidates who supposedly had applied for an assistant professorship. Each candidate's profile included positive, neutral, and negative characteristics, and the characteristics are rated on a 7-point scale ranging from 1 (negative) to 7 (positive). Not all information, however, is shared among group members. The information is distributed to create a hidden profile. For example, each group member knows only two of candidate A's six positive characteristics, and only one of candidate B's negative characteristics. For candidate C, they knew only one positive and one negative characteristic. Thus, based on the shared information, candidate B seemingly has more positive and fewer negative characteristics and therefore would be the preferred candidate. Only by exchanging information through group discussion could the group members detect the best alternative, candidate A.	The goal for each group is to reach an agreement on the best candidate they would hire. The task is designed so that the full information would show that a candidate among the three is objectively the best (the "correct answer"). Through discussion and sharing of information, the group should find out the optimal candidate within the time limit.
Biopharm Seltek	Bhatia and Gunia 2018	Participants are randomly paired for a negotiation about the sale of a biotechnology plant. Each person is randomly assigned the role of either the buyer, the CFO of BioPharm, or the seller, the CFO of Seltek. BioPharm needs to either buy or build a plant to produce a genetically engineered antibiotic compound called Depox. Depox is a promising pharmaceutical product, and BioPharm needs a plant with special manufacturing equipment. There are two choices for BioPharm: (1) build a new plant; (2) buy a plant that is already set up to manufacture genetically engineered compounds. For this second option, Seltek seems like an ideal "turnkey" facility. BioPharm possesses public information about Seltek, including previous appraisal value, accounting statements, and so on. Meanwhile, Seltek is a medium-sized pharmaceutical company that (unbeknownst to BioPharm) has been struggling financially, and it is desperate to sell both the plant and the patent for its chemical compound. Seltek possesses detailed information about its own financial state, and also knows basic information about BioPharm from public documents.	The goal of each party is to reach an agreement that is as beneficial as possible for their own side. In other words, Biopharm should buy the plant for as low a price as possible, and Seltek should sell the plant for as high a price as possible. Since each side possesses different information, participants must strategically share (or choose to withhold) what they know in order to get a good deal. For example, if Seltek divulges the information that they are desperate to sell, they would likely obtain a much lower price than if they did not divulge such information. There are two possible outcomes of the negotiation: either the two parties reach some consensus (to sell the plant at an agreed-upon price) or they are unable to reach an agreement, and BioPharm must build a new plant instead (a costly option).
Blocks World for Teams	Butchibabu 2016	Participants work together in a virtual world to complete a search-and-deliver task. There are nine rooms, designated A1 through C.3, and each room contains various colored blocks. A 10th room, designated the "drop zone," is located at the bottom of the map. A special sequence of colored blocks is depicted below the "drop zone." We can call this special sequence the "target sequence," and it represents the blocks that participants will bring from the main rooms into the drop zone. In addition to the 10 rooms, there are hallways allowing participants to travel from room to room. Each participant controls an avatar and can move their avatar to pick up a block. The participant then navigates to the drop zone to successfully deliver the block. If the participant delivers a block into the drop zone that is not of the requested color, this incorrect block would be automatically and randomly placed into one of the other nine rooms. As they move from room to room, participants are only able to see a limited amount of information. For example, they are only able to see the blocks in the room they are currently occupying, and they are not able to see their teammates. However, participants can see whether or not an adjacent room is occupied.	The goal is to deliver the correct sequence of colored blocks (the target sequence) from the virtual rooms to the drop zone. Participants should complete the task as quickly as possible, as well as finish within a limited amount of time.

Table 2: 102 Tasks in Task Map

Bullard Houses	Cohen, Leonardelli, and Thompson 2010	Participants are assigned to play in one of three formats: one-on-one, two-on-two, or three-on-three. In other words, all teams play with another team of equal size. One side serves as the seller(s) on behalf of a retail company representing the Bullard Family. The other acts as an agent representing a blind trust company operated by a hotel group. The seller is instructed to only sell the property to a known buyer who fully discloses their planned use of the property, while also seeking to get a fair deal for the family. The buyer is instructed to keep the buyer's identity hidden and not reveal the intended use of the site while also trying to reach a fair deal for the property. Both parties are to attempt to reach an agreement while acting in their roles. After negotiations cease, participants should report whether they have reached an agreement or an impasse. These should be coded, i.e. agreement(0) and impasse (1).	Participants should report whether the negotiation is settled or not. This requires careful analysis of the available information both before and during the negotiation. Beforehand, negotiators should work through a variety of simplified, but reasonably realistic financial structures (bonds, mortgages, loans, etc.) to make a judgment about the relative worth of the various offers and possible alternatives. During the negotiations, while much information is instructed to be kept hidden, the information that does get shared may have important and even unforeseen implications for the other side. In this game, the best outcome of a negotiation is to avoid reaching an agreement. In other words, an impasse (1) is the optimal answer.
Carter Racing	Brittain and Sitkin 1986	Participants get several pages of description about a dilemma facing a racecar team, which involves deciding whether to go ahead with the race that would begin in the immediate future. The description mentions that the team has been experiencing a series of engine failures and that an engine failure during this race on national television will present a danger to the driver and team's sponsorship. However, if the team does well on the race, it will get a lucrative sponsorship deal. Participants are given a chart with information about the temperatures during the last 7 engine failures, which shows a range of temperatures 53-75 degrees and a mean temperature of 64 degrees. The chart is misleading because it does not contain information about the air temperature when the car does NOT experience an engine failure (therefore, the information is biased). Participants are also given instructions that mention that they can ask for any additional information during the task, but that the request must be precise.	The goal is to come up with reasons and ultimately decide to terminate the participation in the race within a fixed amount of time. The data in the description is gathered from the Report of the Presidential Commission on the Space Shuttle Challenger Accident in 1986. The relationship between ambient temperature and O-ring failure is disguised as the relationship between air temperature and car engine failures. In other words, participants that decide to go ahead with the race make a decision that is parallel to the decision to go ahead with the Space Shuttle Challenger launch. Participants have to recognize that the data they have is inconclusive and ask the experimenter for the air temperatures when the engine has not failed. If they do so, the experimenter provides them with an additional chart of the last 17 races without an engine failure. The chart shows that races are clearly warmer when engine failures do not occur with the range of 66-82 degrees and a mean temperature of 73 degrees. Therefore, the only correct and precise solution to the task is not to proceed with the race, while participants that come up with a conclusion to proceed with the race fail the task. Participants also have to explain their reasoning behind this decision by finding a relationship between engine failures and air temperatures.
Carter Racing (Experimenterless Version)	Sellier, Scopelliti, and Morewedge 2019	Participants get several pages of description about a dilemma facing a racecar team, which involves deciding whether to go ahead with a race that would begin in the immediate future. The description mentions that the team has been experiencing a series of engine failures and that an engine failure during this race on national television will present a danger to the driver and team's sponsorship. However, if the team does well on the race, it will get a lucrative sponsorship deal. In particular, this race day has unusual temperatures, and participants are told that the outside temperature could be related to engine failures in the past. Participants are given charts with information about the outside temperatures during previous races in the season, and whether or not there was an engine failure during those previous races. Some of these charts may focus just on the successful races, while others may focus only on the failed ones — as a result, the presentation of the data requires the participant to do work to piece the full picture together.	The goal is to come up with reasons and ultimately decide to terminate the participation in the race within a fixed amount of time. Everyone working together submits their decision in a shared interface. The data in the description is gathered from the Report of the Presidential Commission on the Space Shuttle Challenger Accident in 1986. The relationship between ambient temperature and O-ring failure is disguised as the relationship between air temperature and car engine failures. In other words, participants that decide to go ahead with the race make a decision that is parallel to the decision to go ahead with the Space Shuttle Challenger launch. Participants have to recognize that, once they are able to piece the full picture together from the information they are given, there is a clear relationship between temperature and engine failure. Based on the temperature of the day, they should predict that the car's engine will fail, causing them to lose the sponsorship. Therefore, the only correct and precise solution to the task is not to proceed with the race, while participants that come up with a conclusion to proceed with the race fail the task. Participants also have to explain their reasoning behind this decision by finding a relationship between engine failures and air temperatures.

Table 2: 102 Tasks in Task Map

Categorization problem	Abimbola 2006	Participants receive 16 cards with pictures of two objects on each one (32 total objects). Half of the objects (16) on the cards are related to each other, and can be grouped into four categories. For example, the categories could be "vehicles," "animals," "furniture," and "shapes." There are four objects per category. The remaining 16 objects are distractors. There is one category object and one distractor on each card. Participants see all the objects depicted on the cards in a random order. They do not know which objects on the cards belong to the four categories and which are distractors, which makes this task difficult. For example, "furniture" may be one of the categories, and participants' cards may include a picture of an armchair, bar stool, office chair and table. However, there may also be distractors, such as an alarm clock and wall clock, which could "throw off" someone attempting to figure out the categories — they may think that the alarm clock, wall clock, and office chair are all related to a category of going to work, and therefore miss the true category of "furniture."	The goal is to identify the 16 cards belonging to categories, and correctly form four categories with them, within a fixed amount of time. The task is designed in such a way that it is impossible to form four categories if participants start using distractor objects; there is only one objective way of forming categories so that all cards are used. Participants either complete this task by forming the four objectively correct categories, or they fail.
Checkers	Woolley et al. 2010	Participants are given an online checkers board. This is an 8x8 board, in which the squares are colored dark and white in alternating colors. Participants start out with 12 round, flat pieces on each side; these are placed on the dark squares on their side of the board. The rules of checkers are as follows: • The two sides take turns playing. One side completes a move, and then the other side plays. • You can only move a piece diagonally across the board (the game uses only the dark squares). • A piece can only move diagonally into a square where there are no existing pieces. • Pieces can only move forward (the only exception is rule #6). • You can take (or "capture") your opponent's pieces by moving diagonally, jumping over the piece being captured, and landing on an unoccupied square. Multiple enemy pieces can be captured in a single turn by jumping forward several times in a row. • If a piece reaches the other side, it becomes a "king" and has an additional piece stacked on top of it. Kings have the additional ability to jump backwards. • The game ends when one party loses all of its pieces (everything is captured). Participants control one side of the board, and a computer controls the other side. When playing as a team, participants discuss their strategy and make their next move together.	The goal is to win — to beat the computer (the opponent) in checkers. This involves capturing all of the computer's checkers pieces by following the rules of the game. Participants need to win the game within a fixed time limit, or else they automatically lose.
Chicken	Camerer 2017	The game of Chicken simulates a potential confrontation between two players in which each player has the choice to either "Dare" the other player or to "Chicken out." Participants are given a set of payoffs in advance, and depending on each participant's choices, they can obtain different outcomes. In general, participants get a large benefit if they "Dare" the other player and the other player "Chickens out." If both players "Chicken out," participants each get a small benefit. If both players choose to "Dare" the other, then they get a penalty.	Each participant's goal is to maximize their own payoff. The participants must balance the temptation of "Daring" the other person and the risk of their opponent doing the same to minimize their losses. Participants must also make each decision in a fixed amount of time.

Table 2: 102 Tasks in Task Map

		Each participant independently makes their choice without discussion. This process is repeated for a predetermined number of rounds.	
Computer maze	Aggarwal and Woolley 2010	Participants are given a computer monitor with a virtual maze environment and either a keyboard or a joystick. The maze consists of a long winding corridor with many hallways branching off. The hallways are populated by complex unfamiliar objects called "greebles," which are difficult to distinguish from one another. A maze contains a certain number of pairs of identical greebles and a certain number of distractor greebles.	The goal is to navigate through the entire maze and tag as many identical greeble pairs as possible within a fixed amount of time. Participants are evaluated based on maze navigating and greebles tagging. Participants receive points for navigating through the entire maze and don't receive, but also don't lose, any points for not going through the entire maze. Participants get points for tagging identical greeble pairs and lose points for tagging distracting greebles or mixing greebles from other pairs. The maximum score occurs when participants both navigate through the entire maze and tag all greeble pairs correctly.
Crisis mapping	Mao et al. 2016	Participants are given a "crisis mapping" tool on a computer. They view a collection of pre-processed 1567 tweets about Typhoon Pablo in the Philippines, which were posted on December 4-5, 2012, and they receive detailed instructions on how to navigate the crisis mapping interface. Some of the tweets contain information relevant to the typhoon, such as evacuation centers, damaged infrastructure, or number of deaths, while other tweets are irrelevant to the disaster.	The goal is to select relevant tweets and geo-locate, time-stamp, and categorize damage/flooding photos and videos in them, creating a precise and accurate crisis map for Typhoon Pablo within a fixed amount of time. Experimenters pre-define a "gold standard" map with 49 distinct events in different regions. Participants only get credit based on how close their map is to the "gold" map. Therefore, their goal is to identify as many relevant events as possible without including irrelevant events, and the highest possible score is achieved when they identify precisely the same events as those in the "gold standard."
Desert survival	Mayo et al. 2020	Participants view information describing a survival scenario in which a plane crashes in the desert. The survivors are able to salvage a list of 16 items (e.g., water, machete, compass, cosmetic mirror) from the wreckage. Everyone sees the same list of items.	The goal is to rank all 16 items in the order of most important to least important for survival. For example, if water is the most important for survival, then it would receive the rank of 1. If the cosmetic mirror is the least important for survival, then it gets the rank of 16. Participants' rankings are evaluated based on how close their rankings are to those created by a panel of survival experts. That is, the expert ranking is the "true" answer, and the closer a participant is to the true answer, the more points they earn. The highest possible score is achieved when participants rank all 16 items in the exact same order as the experts did. If working in a team, participants are able to discuss the scenario with each other and come up with a single set of rankings as a team.
Dictator game and its variants	Dana et al. 2006	Participants are placed into pairs and randomly assigned the role of "dictator" or "receiver." In each round of the game, the dictator sees an amount of money (the "endowment") on the screen, as well as an option to give part of the money to the receiver. After choosing how much to give to the receiver, the dictator keeps the remaining funds, while the receiver keeps the amount that was given to them by the dictator. The receiver is required to accept whatever the dictator chooses to give them and has no option to reject the offer. The same pairs play together for a predetermined number of rounds and roles are reassigned each round.	Each participant has the goal of maximizing their own earnings. In the dictator role, participants must decide how much they will offer (from $0 to the full endowment). Since the dictator knows that the receiver cannot reject the offer, it is in the dictator's interest to give the receiver nothing ($0), because they would then keep the full amount. However, there are a few reasons why dictators may choose to give more than $0. Giving nothing may conflict with social norms (because the dictator would then be seen as self-serving), or simply feel morally "unjust." If there is more than one round being played, dictators may worry that their relationship and reputation with their partner may be harmed in later rounds. The dictator must weigh these considerations and decide on a final amount to give to the receiver. Participants must also make their decisions in a fixed amount of time.
Divergent Association Task	Olson et al. 2021	Participants are given a task sheet with instructions, a blank space to write 10 words in, and a list of rules.	The goal is to come up with 10 nouns as different from each other as possible within a fixed amount of time. For example, words "cat" and "dog" are considered similar to each other because they are often used in the same context, while words "cat" and "thimble" are different because they are rarely used together.

Table 2: 102 Tasks in Task Map			
			Participants are evaluated on how semantically different the words are. The minimum score occurs when all the words are the same, while the theoretical maximum score occurs when words have the least possible relation to each other. Participants maximize their score by coming up with as many of the most semantically distant words as possible.
Estimating Factual Quantities	Silver, Mellers, and Tetlock 2021	Participants are given a series of questions on estimating certain factual quantities. For example, participants can be asked to estimate the year when DNA was discovered or the current price per share of Twitter stock.	The goal is to answer as many questions from the series as correctly as possible within a fixed amount of time, and without looking up the solutions online. Participants are evaluated based on the percent difference of their estimation from the correct answer. Participants receive the maximum score if all their estimations are exactly like the correct answers and the percent difference between them is 0. Participants maximize their score by submitting estimations that are as close to the right answers as possible. For example, because the DNA was discovered in 1869, participants get the maximum score for this estimation if their guess is 1869, while an estimation of 1860 receives a higher score than an estimation of 1900.
Estimating geological metrics	Whiting et al. 2019	Participants are given a series of questions on estimating geological metrics. For example, participants can be asked to estimate the number of miles from New York to Los Angeles or the number of U.S. states that border the Gulf of Mexico.	The goal is to answer a series of questions within a fixed amount of time as close to the correct answer as possible without looking up the solutions online. Participants are evaluated based on the percent difference of their estimation from the correct answer. Participants receive the maximum score if their estimation is exactly like the correct answer and the percent difference between them is 0. Participants maximize their score by submitting an estimation that is as close to the right answer as possible. For example, because the distance between New York and Los Angeles is 2446.3 miles, participants get the maximum score if their estimation is 2446.3, while an estimation of 2400 miles receives a higher score than an estimation of 2000 miles.
Estimating pages of a book	Engel et al. 2014	Participants see images of the page edges of a book.	The goal is to guess the exact number of pages that the book has, using only the image of the page edges. Participants are rewarded for being as precise and as close to the true number of pages as possible. They get points proportional to how many pages "off" they were, corrected for the total number of pages in the book. For example, 10 pages wrong out of a 400-page book is a smaller error than 10 pages wrong out of a 50-page book. Participants are not told if their judgment is exactly correct. Finally, participants have a limited amount of time to make their judgment.
Estimating social quantity	Woolley et al. 2010	Participants are given a series of questions on estimating quantities of real-life social phenomena. For example, participants can be asked to estimate the median age in the U.S. in 2009.	The goal is to answer a series of questions within a fixed amount of time as close to the correct answer as possible without looking up the solutions. Participants are evaluated based on the percent difference of their estimation from the correct answer. Participants receive the maximum score if their estimation is exactly like the correct answer and the percent difference between them is 0. Participants maximize their score by submitting an estimation that is as close to the right answer as possible. For example, because the median age in the U.S. in 2009 is 36.7, participants get the maximum score if their estimation is 36.7, while an estimation of 35 years receives a higher score than an estimation of 30 years.

Table 2: 102 Tasks in Task Map			
Euclidean traveling salesperson	Bernstein et al. 2018	Participants get a map with multiple cities, which are displayed as dots on the screen. These dots can be connected by different "paths" by clicking on two dots one after the other. For example, clicking on point A, then point B, would draw a "path" between A and B. All participants, who are working on individual screens, are shown a shared map of the cities and can click and interact with it simultaneously.	The goal is to find and select the precise shortest path through the cities within a fixed amount of time. This is the sequence of cities corresponding to the shortest possible path to traverse through all of the cities, forming a cycle (or loop). In other words, the participant should click through the dots in a sequence, thus drawing a path that touches every city exactly once, and the first city twice: you can think of it as a "traveling salesman" who visits each city on the map before returning to his home city. Imagine that the shortest path involves going from A, to B, to C, to D. In that example, the participant's goal would be to find this path, and then click the cities in that order: A, B, C, D, A. Participants are evaluated on how close the path they found was to the optimal shortest path. They will be able to see the length of the current path they found, but they will not be told that they have reached the optimum.
Find the common symbol	Leavitt 1951	Each participant is given a set of five symbols. For each participant, four of their five symbols are different from those of the other participants, and only one of the symbols is common among everyone. Participants cannot see each others' symbols, but they can send messages to one another.	The task is to work together to find the common symbol, and to do so (1) as quickly as possible; (2) with as few communications/messages as possible; and (3) with as few mistakes as possible. Their performance would be assessed by their time (in seconds) spent for a correct answer, the number of messages sent by each group member, and the number of wrong answers that had to be corrected during a trial. Participants should provide an answer within the time limit.
Find the maximum	Weidmann and Deming 2020	Participants see an empty text box where they can enter numbers. Behind the scenes, there is a hidden function. This hidden function has a maximum value — for example, the function y = -x2 reaches its maximum when x = 0. Participants know nothing about the hidden function; they only see the text box. When participants enter numbers into the text box, the system will give them feedback about the value of the function at that point. For example, suppose that the hidden function is y = -x2 and the participant enters 2. Then the system will tell them that the value is -4, since this is the value of -x2. If the participant enters -5, the system will tell them that the value is -25. As they make guesses, participants see a log, which shows their guesses so far, the feedback they previously received, and how many guesses they have remaining.	The goal is to correctly guess the maximum value of the hidden function after a limited number of guesses (for which there is feedback after each guess). Every problem has an objectively correct answer — the mathematical maximum of that function. However, participants are not told that they have found the correct answer even when they submit it as their guess. For example, if the hidden function is y = -x2, the correct answer is 0. A participant who submits 0 will simply be told that the value of the function at that point is also 0; they will not be told that this is the true maximum. Instead, the participant will just get their fixed number of tries, after which they are required to submit their final guess. Participants are scored based on how close their final guess is to the true maximum. If the answer is 0, then the participant gets the highest possible number of points if they guess 0; a participant who guesses 2 will get a higher score than someone whose final guess is 10. Participants aim to get as close as possible to the correct answer. Finally, if participants are working together, each one gets a limited number of guesses, and then they decide their final guess as a team.

Table 2: 102 Tasks in Task Map			
Game of Clue - Terrorist Attack	Shore et al. 2015	Participants work in a shared web interface in which they solve a 'who-done-it' game like Clue or Cluedo, but involving a terrorist threat. Everyone in the game is solving the same puzzle. Each participant has a personal "inbox" where they have their own list of facts and clues. In addition, at specified intervals of time, participants are able to use a "search engine" tool to look for more clues, which then appear in their inbox. For those who are working in groups, participants also have the ability to share clues by sending them to each other and to add comments to each other's clues.	The goal was to solve for four facts: • Who would carry out the terrorist attack • What would be the target • Where the attack would take place • When the attack would take place. The game is designed so that there is exactly one correct answer for each of these four facts. Participants submit their final answers via the web interface. Since there are multiple facts that need to be solved, participants are able to submit one "solved" fact at a time. They are also able to return to the submission page to change their answers at any point before the time limit ends. Participants are not given any feedback about whether their submissions are right or wrong, but when playing in a group, they are able to view what other people have been guessing so far (which can help provide a hint). Participants are evaluated not only for correctly finding the exact facts, but also for solving the mystery as quickly as possible — they earn more points for finding each fact faster. Therefore, someone who finds all 4 facts in 2 minutes will earn more points than someone who submits the same answers, but takes 5 minutes to reach their conclusion.
Graph coloring task	Kearns et al. 2006	Participants are given one or more "graph coloring problems". These are pictures of different topology/network graphs. For context, these look like many different nodes (points), which are connected by various edges. Imagine, for example, a picture of a social network: each person in the network is a "node," and each of their friends is an "edge." In this case, all participants have access to and can interact with a shared graph. Participants will also be given a limited set of colors, with a larger number of colors than the minimum required to complete the task successfully.	The goal is to use the colors given to "color" the entire network without conflicts. This means that participants need to assign each node in the graph a color, ensuring that no two connected nodes share the same color. Using our social network example, if Person A and Person B are friends (which means they are connected), they cannot be assigned the same color. Participants must successfully color all nodes and resolve all conflicts within a specified amount of time, or receive no credit.
Guessing the correlation	Almaatouq et al. 2020	Participants receive a series of scatter plots which are a series of points plotted on a grid. In any round, all participants see plots with an identical correlation.	The goal is to estimate the correlation of the points on the grid. (For example, if all the dots are in a straight line pointing upwards, the correlation is 1; if the dots are completely scattered randomly with no pattern at all, the correlation is 0). There is a time limit to make an estimation. Participants are evaluated on the accuracy of their judgment. They receive the highest score by guessing the exact correlation but they have to try to get as many points as possible by being as precise as they can. After participants submit their guess, they can see the true correlation and see whether they are right or wrong as well as how close their guess is to the true correlation.
Hidden figures in a picture (Recall Task)	Finlay, Hitch, and Meudell 2000	Participants see an incredibly busy and complex image that contains many hidden animals. Imagine a picture that looks like "Where's Waldo" — with lots of people and animals in one detailed, colorful image. Participants are then given a list of animals that they are supposed to find in the picture: for example, lion, penguin, dolphin, elephant. If working together, participants in a team see the same picture and have the same list of animals to find. Participants then engage in a task in which everyone working together attempts to find as many animals from the list as they can	The goal of the task is to be able to recall as many animals as they can from the original list. Participants type and submit the words using the webpage. They have a limited amount of time to submit all the words that they can remember. If working together, only one participant on a team needs to remember the word in order for everyone to receive credit. Participants are scored based on the number of animals from the original list that they are able to recall. For example, if there were 10

Table 2: 102 Tasks in Task Map

		in the provided image. For example, if "elephant" is on the list, they need to look for the elephant in the picture, and then click on it. After a delay, participants are taken to a page where they can type and submit words.	animals on the list, they would get the maximum score if they recalled all 10 animals.
Hidden figures in a picture (Searching Task)	Finlay, Hitch, and Meudell 2000	Participants see an incredibly busy and complex image that contains many hidden animals. Imagine a picture that looks like "Where's Waldo" — with lots of people and animals in one detailed, colorful image. Participants are then given a list of animals that they are supposed to find in the picture: for example, lion, penguin, dolphin, elephant. If working together, participants in a team see the same picture and have the same list of animals to find.	The goal of the task is to successfully identify the locations (in the image) of as many animals as possible from the list. Participants must find the most animals they can within a limited period of time. Participants submit their response by pointing and clicking on the correct location of the animal, and they get points for every animal that they are able to find. Thus, they would get the highest score if they find all animals on the list. If participants are working together, only one person on the team needs to click on the animal in order for it to be counted, which means that having members of a team search separately is advantageous.
Husbands and wives transfer	Lorge and Solomon 1959	Participants are given a problem that reads as follows: "On the A-side of the river are wives (W1, W2, W3) and their husbands (H1, H2, H3). All of the men but none of the women can row. Get them across to the B-side of the river by means of a boat carrying only three at one time. No man will allow his wife to be in the presence of another man unless he is also there."	The goal is to achieve the fewest number of trips that get all six people to the B-side, and to solve this problem within the time limit. There are several possible "shortest" solutions to this problem. In general, participants must realize that solving the problem requires people to row back and forth — that is, some of the people who go from the A-side to the B-side have to go BACK to the A-side — and then figure out the right combinations that meet the requirements (W1 cannot be in the presence of H2 or H3 unless H1 is also there, and so on). Therefore, participants must satisfy the constraints of the problem while also thinking creatively. As a final piece of helpful context, this task is actually a well-known one that many people have built algorithms to solve.
Image rating	Engel et al. 2014	Participants see pictures of a series of products and accompanying slogans for them. They then have the opportunity to rate the product and slogan on a scale from 1 (worst) to 10 (best). Participants see these pictures one at a time.	The goal is to give a rating that is as close as possible to how other Americans viewed the product and slogan. The challenge is to be exact in predicting how other Americans think. Participants receive points based on how close their ratings are to the "truth." The "truth" is calculated based on the average poll response of 100 American users on Amazon Mechanical Turk. Participants are NOT told whether their guesses are correct or how many points they earned. Finally, participants have to make their decisions within a limited amount of time.
Intergroup Prisoner's Dilemma	Bornstein 2003	All participants are given a fixed amount of individual money ("endowment") and divided into two groups, A and B. Participants are told that their group is competing with the other group to see which one "invests" the most money. The winning group receives a larger cash prize. Every individual player in the group must make an independent decision to either invest their personal endowment or keep it for themselves. The group wins or loses depending on the number of players that chose to invest relative to the other group. For example, if everyone in Group A invests and no one in Group B invests, Group A would win the entire cash prize and split it among its members. If some people from both groups invest, but more people from Group A invest than people from Group B, then the two groups would split the cash prize (but Group A would get more of it). Finally, if the same number of people in Groups A and B invest (a tie), then the two groups would split the cash prize 50-50.	Each participant's goal is to maximize their own payoff. Therefore, when making a decision, participants have to weigh up the cost of losing their individual endowment with the expected gain from splitting the cash prize if their group wins. Since the outcome depends on not only the participant's personal choice, but also the choices of all other players, there is a substantial amount of uncertainty and risk. For example, a participant could choose to invest, only to find that their group has lost (not enough other players invested), and they end up with less money than when they started. On the other hand, there is a greater chance of winning the maximum cash prize if everyone in the group invests. Participants must make their decisions in a fixed amount of time.

Table 2: 102 Tasks in Task Map

Investment Game (aka Trust Game)	Berg et al. 1995	Participants are assigned to one of two rooms: Room A or Room B. They then play a two-stage game. In the first stage, participants in Room A are given some money and have to decide how much to send to participants in Room B. In the second stage, the recipients in Room B get triple the amount of money that was sent to them. They then must decide how much money to keep for themselves, and how much to send back to those in Room A.	Participants should maximize their own pay-off (get as much money in the end as possible). Those in Room A must balance the possible benefit of earning more money (since all the money they give to Room B is tripled) with the risk that the people in Room B would keep all the extra money for themselves, leaving those in Room A with nothing. Conversely, those in Room B must decide whether they will keep all of the money for themselves, or give a portion of it back to Room A. Participants are given a limited amount of time (e.g., 10 minutes) to send money back and forth.
Investment game (hidden-profile)	McLeod et al. 1997	Participants are asked to take the role of the board of directors of an investment company. Their task is to evaluate three companies available for acquisition using a specified set of investment criteria and to provide a rank ordering of the companies' desirability. Profiles on each of the three companies include the financial opinions of in-house and external financial analysts, information on the company's business strategy, strength of its management team, market position, and human resources practices. The investment criteria for evaluating the companies include long-term financial return, the degree of risk tolerated, the ability of the company to stand on its own and fit with the company's general business philosophy. Each participant is given different information about the investment options, which means that initial majority and minority opinions are likely to be divided. Participants are then asked to discuss the investment strategy. After they discuss, the group must come to an agreement on how to rank-order the three companies from best to worst investment.	The goal for each group is to reach an agreement on the correct rank ordering of the three companies based on the information they receive. The task is designed so that the full information would show that one company is objectively the best, one is in second place, and one is objectively the worst. Through discussion and sharing of information, the group should find out the "correct decision" within the time limit.
Iterated Snowdrift Game (With Punishment)	Jiang, Perc, and Szolnoki 2013	Participants are randomly paired in a group acting as two drivers who are caught in a blizzard and trapped on either side of a snowdrift. They can either get out and start shoveling ("cooperate") or remain in the car ("defect"). There are two stages in this game. During the first stage, participants choose to cooperate or defect. If both cooperate, they have the benefit of getting home while sharing the labor. Thus, Rewards = (benefit - labor)/ 2. If both defect, they do not get anywhere and hence incur the punishment, meaning Rewards = 0. If only one shovels, however, they both get home; however, the defector is able to free-ride and avoids the labor cost while reaping all the benefits (Rewards = benefit). In contrast, the cooperator gets fewer Rewards, as they have to put in all the labor (Rewards = benefit - labor). We assume benefit is greater than labor, as both parties want to get home. During the second stage, the cooperator is given the chance to punish the defector for their free-riding behavior. Punishment works as follows: the cooperator has to incur a fixed fee (1 unit) in order to punish defectors. If they choose to incur the fee, then the payoffs for the defectors are reduced by a fine. Note, however, that while the fine can take different values, the fee for punishing does not change. The possibility that someone might punish you if you defect therefore makes the option of defection more risky.	Participants are instructed that their goal is to maximize their individual payoff (Rewards) and are given a limited amount of time (e.g., 10 minutes) to make a decision during the first stage. Letting the opponent do all the work is the best option for your individual payoff, but if the other player stays in the car, it is better to shovel. The worst outcome is if both players choose to stay in the car. For example, imagine that there are three groups in the game. In Group 1, both player A and player B decide to cooperate. In Group 2, player C remains in the car while player D gets out and starts shoveling. In Group 3, neither player E nor player F is willing to cooperate. The payoff values of different players rank in order: player C > player A/player B > player D > player E/player F. However, participants should take the fine and potential cost of punishment into account in this game. The cooperator has the chance to punish the defector, which means if the fine is heavy, it is not worth it to defect.

Table 2: 102 Tasks in Task Map

Iterated Snowdrift Game (Without Punishment)	Jiang, Perc, and Szolnoki 2013	Participants are randomly paired in a group acting as two drivers who are caught in a blizzard and trapped on either side of a snowdrift. They can either get out and start shoveling ("cooperate") or remain in the car ("defect"). If both cooperate, they have the benefit of getting home while sharing the labor. Thus, Rewards = (benefit - labor)/ 2. If both defect, they do not get anywhere and hence incur the punishment, meaning Rewards = 0. If only one shovels, however, they both get home; however, the defector is able to free-ride and avoids the labor cost while reaping all the benefits (Rewards = benefit). In contrast, the cooperator gets fewer Rewards, as they have to put in all the labor (Reward = benefit - labor). We assume benefit is greater than labor, as both parties want to get home.	Participants are instructed that their goal is to maximize their individual payoff (Rewards) and are given a limited amount of time (e.g., 10 minutes) to make a decision. Letting the opponent do all the work is the best option for your individual payoff, but if the other player stays in the car, it is better to shovel. The worst outcome is if both players choose to stay in the car. For example, imagine that there are three groups in the game. In Group 1, both player A and player B decide to cooperate. In Group 2, player C remains in the car while player D gets out and starts shoveling. In Group 3, neither player E nor player F is willing to cooperate. The payoff values of different players rank in order: player C > player A/player B > player D > player E/player F.
Iterative lemonade stand task	Sommer, Bendoly, and Kavadias 2019	Participants are given an interactive online interface that represents managing a lemonade cart. The program allows users to tweak five variables: (1) price, (2) lemon content, (3) sugar content, (4) lemonade color, and (5) location of lemonade cart. There are limited options for lemonade color and cart location that are predefined within the task. For example, the lemonade can only be either yellow or orange, and a cart can either stand in the park or near the supermarket. Price can be any number the participant decides to use, while lemon and sugar content can be any value between 0 and 100 percent.	The goal is to maximize the cart's profit by adjusting the five given variables within a fixed amount of time. For example, participants can decide to sell their orange lemonade in the park with 20% lemon content and 5% sugar content for $3 a glass. There is a formula in the program that quantifies the cart's profit and shows participants instant feedback. Participants will not be told that they have reached the maximum possible profit; rather, they simply see feedback about the current profit and are able to adjust as many times as they want before they submit. Participants are evaluated based on their profit. The more profit their cart makes after they tweak the variables, the better their performance is on the task.
Letters-to-numbers problems (cryptography)	Laughlin et al. 2006	Participants are given instructions to map letters A, B, C, D, E, F, G, H, I, and J, which have been initially randomly assigned without replacement, to one of the 10 digits (0 through 9). For example, because letter "A" maps onto digit "3" and letter "B" maps onto digit "5", an operation "A + B" would correspond to "3 + 5". If letter "C" would map onto digit "8", then A + B = C.	The goal is to identify the mapping of the 10 letters to the 10 numbers in as few trials as possible within a fixed amount of time and a limited number of trials. Participants must also get the correct answer as quickly as possible. There is a specific process participants have to follow to map letters onto numbers. Participants are first asked to come up with an addition or subtraction operation (e.g., "A + B") and then given an answer in the letter form (e.g., "A + B = C"). Then participants propose a mapping for a letter (e.g., "A = 3") and receive feedback (e.g., "True, A = 3"). Participants have to follow this process in proposing the mapping of all letters to all digits. To submit their final answer, participants have to indicate which digits each of the letters correspond to. Participants are evaluated based on several criteria, such as the number of trials they used, letters identified per trial, and letters used per equation. There is a theoretical maximum score, and participants get the highest score by finding the correct mapping as fast as they can, with as few trials as possible.
Logic Problem	Littlepage 1991	Participants receive a logic problem, which is a story problem with a correct answer that must be deduced via logic. This problem is shared among everyone who is working together. Here is an example question: There are five couples who vacationed at five different ski resorts. Based on the clues, figure out which couple vacationed where. Participants would get a set of clues to solve this problem, such as "Neither Tammy nor Sue vacationed in Innsbruck," "Both Rita and Mrs. Owens returned from their vacations with broken wrists," "no one at Tahoe was injured that weekend," and so on. Participants are provided with sufficient clues to be able to solve the question.	The goal is to use the clues provided to achieve the correct answer in the logic problem. Participants must solve the problem and submit their answer within a limited amount of time. In the above example, the participants would use logic to reason that neither Rita nor Mrs. Owens went to Tahoe, since no one at Tahoe was injured. They would then use other clues to match the correct people to the correct vacation locations. Participants are scored based on the number of correct matches in the final solution that they get. In this case, since there are five couples and five ski resorts, participants would get credit for every pair that they are able to match, and they would achieve the maximum score if they get all five pairs correct.

Table 2: 102 Tasks in Task Map

Mastermind	Bonner, Baumann, and Dalal 2002	Participants get a series of pegs, or points, of different colors to use in a logic game played on the computer. For example, participants can get 6 colors to use during the game: blue, red, green, orange, yellow, and purple. As the game begins, an algorithm randomly designates a color pattern as the correct answer to the game. For example, the correct pattern to the game can consist of 4 colors, in this order: yellow, orange, yellow, red.	The goal is to identify all the colors and their positions from the correct pattern, using as few trials as possible and within a fixed amount of time. As the game begins, participants have a limited number of trials to propose color patterns. Participants receive immediate feedback in the form of black and white pegs after each trial. A white peg means that one of the colors they chose is in the pattern (but in the wrong position); a black peg means that one of the colors they chose has both the color and the position correct. For example, if the correct pattern is yellow, orange, yellow, red, and a participant proposes blue, green, orange, red, then this participant correctly identifies two colors from the pattern and a position of one of them correctly. Therefore, they would receive a white peg for orange (correct color, wrong position) and a black peg for red (correct color, correct position). However, participants don't know from the feedback exactly which color and position they identified correctly; instead, they have to try to deduce it from feedback for each trial. Participants try to guess the pattern using as few tries as possible, and get the highest score by receiving as many pegs, especially black ones, as early in the game as possible.
Minimal Group Paradigm (study diversity)	Tajfel 1970	Participants are randomly and anonymously divided into two groups (e.g., "Group A" and "Group B"). After they are divided into groups, participants receive an anonymous list of players (e.g., "participant number 34 of Group A," "participant number 12 of Group B"). They are then asked to distribute a valuable resource (e.g., money or points) between the participants on the list. Participants are told that, after the task is finished, they will receive the total amount of the resource that has been allocated to them by the other participants.	The goal for participants is simply to allocate the resources however they like. There are no right or wrong answers for how participants should divide up the resources. Participants do NOT need to maximize the resources of people on their own team, nor maximize their personal resources.
Minimum-effort tacit coordination game	Van Huyck, Wildenthal, and Battalio 2002	Participants play in a large group with others. Each participant is given a list of numbers from 1 to 7. In every round, participants must independently choose and submit one of the numbers as their "guess." Each participant's guess will then be compared to the median guess across all participants. Before making their guesses, participants receive a payoff table that explains how much money they would earn for their guess, depending on how close it was to the median, and on the value of the median. In general, participants earn more money if their guess is closer to the median, and if the median is higher (i.e., closer to 7). Participants win less money if their guess is farther away from the median (e.g., the median was 4 and they guess 5), and if the median is lower (i.e., closer to 1). Additionally, participants suffer a greater penalty for "overshooting" (guessing higher than the median) than they do for "undershooting" (guessing lower than the median). This process is repeated for a predetermined number of rounds.	Each participant's goal is to maximize their own payoff. The highest possible payoff is if the participant correctly guesses the median, and the median is 7. However, if the true median is lower than 7, the best payoff in the round is to correctly guess the median value and to avoid overshooting. Guessing too high (for example, choosing 7 when the median is 4) would result in a penalty. Therefore, participants must balance the tradeoff between guessing a large value, which boosts the median and the overall payoff, and guessing a value that is too large, which results in the overshooting penalty. Additionally, there is a tradeoff between individual gain and collective benefit: if a participant chooses 6 or 7, they would boost the median value (which would increase the overall payoff), but they would be more likely to personally overshoot (leading to the penalty). On the other hand, if the participant chooses 1 or 2, they would decrease the overall payoff, but reduce their own chances of overshooting. Participants must also make each decision in a fixed amount of time.

Table 2: 102 Tasks in Task Map

Mock jury	London and Nunez 2000	Participants are to act as jurors in a mock jury. Once they have read the summary of the trial, they will be given a brief pre-deliberation questionnaire. They should take into account all facts of the trial and any specific instructions given by the judge before filling out the questionnaire and continuing with the process. In the questionnaire, participants are asked to record their personal verdict for the trial, and their assuredness of this verdict, which are indicated on a 9-point scale. The scale ranged from 1 (definitely not guilty) to 9 (definitely guilty), and 1 (not at all) to 9 (very confident). After they finish the questionnaire, all the jury members will be instructed on the laws surrounding the case. All participants will then deliberate and attempt to reach a unanimous verdict. The jury is allowed to deliberate for a limited time (e.g. 1 hour) until they reach a verdict. Then all jurors will once again be given questionnaires where they will render their individual verdict, and record their assuredness of the verdict.	Participants should reach a unanimous verdict with their fellow jury members within a limited amount of time.
Moral Reasoning (Disciplinary Action Case)	Woolley et al. 2010	Participants read a fictitious, controversial case in which a college basketball player bribed an instructor to change his grade on an exam in order to maintain his eligibility on the team. Participants are also given a list of potential actions that the college could take in response, in which the student and/or instructor could be given a punishment or consequence. Examples include: (1.) Lowering Student's Grade (2.) Student Suspension (3.) Punishment from Student's Basketball Team (4.) Instructor Punishment (5.) Preventing the Instructor from Getting Future Position and/or Promotion These possible courses of action are further challenged by conflicting interests of the faculty, college administration, and the athletic department.	Although there is no right or wrong course of action among the options, the goal is to select one of the courses of action. Participants had to try to make sense of the different stakeholders' conflicting interests, get agreement from team members, and make a final choice about how to handle the disciplinary situation.
NASA Moon survival	Yetton and Preston 1983	Participants view information describing a survival scenario in which a spaceship crashes on the moon. The survivors are able to salvage a list of 15 items (e.g., matches, food concentrate, a 50-foot nylon rope, a portable heating unit, two 100-pound tanks of oxygen) from the wreckage. Everyone sees the same list of items. The survivors now need to carry the most important items on a 200-mile trek to get to "home base."	The goal is to rank all 15 items in the order of most important to least important for survival, as the astronauts travel from the wrecked spaceship to home base. For example, if the oxygen tanks are the most important for survival, then they would receive the rank of 1. If the 50-foot rope is the least important for survival, then it gets the rank of 15. Participants' rankings are evaluated based on how close their rankings are to those created by a panel of experts from NASA. That is, the expert ranking is the "true" answer, and the closer a participant is to the true answer, the more points they earn. The highest possible score is achieved when participants rank all 15 items in the exact same order as the experts did. If working in a team, participants are able to discuss the scenario with each other and come up with a single set of rankings as a team.
New Recruit	Overbeck, Neale, and Govan 2010	Participants are randomly paired for a negotiation over a mock job recruitment scenario. Each person is randomly assigned the role of either the recruit (person being hired) or the recruiter (person doing the hiring). The different roles have different objectives: for example, the recruit may want a higher bonus, while the recruiter would rather pay as low a bonus as possible. The possible settlement options for each participant are associated with point values, which are described in a private "payoff schedule" for each participant (explained in more detail below). Here is an example to help you imagine what the negotiation might look like. Participants would receive a list of settlement options for	Participants are explicitly instructed that their goal is to maximize their (individual) payoff, and were given a limited amount of time (e.g., 30 minutes) to conduct their negotiation. In order for any agreement to be binding, participants need to reach an agreement on all the issues.

Table 2: 102 Tasks in Task Map

		each issue, and the payoff (in points) associated with each option. They would see only their own payoff schedule, and would not be aware of their counterparts'. Participants would then conduct negotiations over the employment package.	
Object based generalization for reasoning (Phyre)	Bakhtin et al. 2019	Participants get a series of puzzles related to properties of physical objects in a simulated 2D world on a computer. Access is given to all puzzles at once and participants can choose which puzzles to work on. For example, in one of the puzzles, participants can view an environment with two blue and green balls on little platforms floating in space, with a slanted floor at the bottom. Considering this scenario, participants are asked to make the green ball and blue ball touch. Participants are expected to solve puzzles by adding physical objects to the environment so that when the physical simulation is run, the puzzle gets solved. For example, the participant could use an object to knock the blue and green balls off of their floating platforms, and cause them to roll down the ramp and touch (thus completing the puzzle). There are many ways to creatively add objects to the environment and cause different physical consequences.	The goal is to solve as many puzzles in as few attempts as possible within a fixed amount of time. The maximum score occurs when all puzzles are solved during their first attempts. After each unsuccessful attempt, the environment resets to its initial state and participants can try again until they run out of time, or decide to try a different puzzle.
Oligopoly game	Hemenway et al. 1987	Participants experience a game with a pre-determined number of rounds. In each round of the game, participants must independently decide whether to "compete" or "collude." Players get a different amount of payoff in each round based on a combination of how they decided and how everyone else in the round decided. The players know the different possible payoffs ahead of time. Here is an example. Suppose that the four possible pay-offs are as follows: [You choose - "Compete", Majority chooses - "Compete", Your Payoff - 10], [You choose - "Collude", Majority chooses - "Compete", Your Payoff - 0], [You choose - "Compete", Majority chooses - "Collude", Your Payoff - 40], [You choose - "Collude", Majority chooses - "Collude", Your Payoff - 20]. In this example, you would get the most payoff if you choose to compete, but the majority chooses to collude. Let's say that, in the first round, more than half of the players decide to "compete." Then those who choose to "compete" would receive 10 points. Those who choose to "collude" would receive 0 points. Then we will play round 2 ,and everyone will vote all over again. Suppose in round 2 that a majority of players choose to "collude". Then those who choose to "compete" receive 40 points and those who "collude" receive 20 points.	Each participant's goal is to maximize their own payoff. Therefore, when making a decision, participants will have to weigh individual benefits against the unknown actions of other players. Notice that, even though the best possible individual outcome (Payoff = 40) is for you to choose "Compete" and for the majority to choose "Collude," if everyone follows this same logic, then the majority will end up being "Compete," and the payoff will only be 10. Participants must make their decisions in a fixed amount of time.
Organization Game	Krackhardt and Steele 1988	Participants are randomly assigned to one of four equal size but physically separated divisions, each having its own complement of operating (line and staff) units. With the minimal structure and only unit heads and resource controllers assigned, participants must decide how to further develop and staff the organization in order to achieve an effective system for the division and coordination of work. These processes must include a way for the organization to learn and change as it adapts to internal and external forces.	The goal is to maximize the organizational effectiveness score based on several objective performance indicators. Notably, the goal is for participants to coordinate and perform well as an entire organization, even though each individual's control is limited to their own unit. Factors that will impact the overall score include the productivity of line units, investments in organizational development programs, and events such as members' absence, strikes, dismissals, resignations, vacations, and underutilization. The indicators for each session of play are calculated by the coordinator and delivered to the information processing unit at the beginning of the next session.
Pharmaceutical Company (hidden-profile)	Kelly and Karau 1999	Participants are randomly assigned to triads and participate in a role-playing simulation of managers in a pharmaceutical company trying to decide which of two drugs to market. They would be given information sheets about the drugs and be given a limited amount of time (e.g., 10 minutes) to familiarize themselves with the information. However, the information sheets are constructed such that some of the facts were shared (provided to all three members)	The goal for each group is to reach an agreement on the drug they would like to market. The task is designed so that the full information would show that one drug of the two is the best. They should find out the optimal drug within the time limit.

Table 2: 102 Tasks in Task Map			
		and some were unshared (provided to only one group member). They would be also told that the company is in financial trouble and that their marketing decision is therefore especially important. After they finish reading the information sheets individually, they would be asked to recall as many of the facts about each drug as they could and be given limited time for the recall task. Following the recall task, the group would be asked to discuss the problem and come to a group consensus about which drug to market.	
Prisoner's Dilemma (various versions)	Dawes 1980	Participants are given a choice to "cooperate" or "defect" and are given, in advance, the payoff to be expected for each choice, and when appropriate, the combination of choices between participants. The expected payoffs are the same for all participants. Each participant independently makes their choice without discussion. This process is repeated for a predetermined number of rounds.	Each participant's goal is to maximize their own payoff. For example, let us consider a simple two-player game. If both participants cooperate, they'll both lose a small amount, and if both defect, they will lose a large amount. If one participant cooperates while the other defects, the cooperator will lose a large amount while the defector loses nothing. In this scenario, the participants must balance the temptation of defecting and the risk of their opponent doing the same to minimize their losses. Participants must also make each decision in a fixed amount of time.
Public goods game	Tomassini and Antonioni 2020	Participants receive the same amount of tokens and they secretly choose how many of their private tokens to put into a public pot. The tokens in this pot are multiplied by a factor (greater than one and less than the number of players, N) and this "public good" payoff is evenly divided among players. Each participant also keeps the tokens they do not contribute. Here is an example to help you imagine what the game might look like. A group of four players participates in this game and starts with $20 in hand. Three players contribute their full $20 while the fourth chooses to contribute $0. The $60 is multiplied by a factor of 1.2 and the resulting $72 is distributed equally among the four players (this means everyone gets $18 from the pot). In this scenario, the three players contributing the most end the round with $18, while the fourth player gets as much as $38 (they kept their initial $20, and earned an additional $18).	Participants are instructed that their goal is to maximize their individual payoff as well as the group's total payoff. It means that one goal is to maximize their personal earnings, which means that they should not invest anything, and simply "free ride" off of the others, as in the above example. But another goal is to optimize the group's outcomes, which means that everyone should contribute all of their tokens to the public pool. If everyone in the example had contributed their full $20, then there would have been $80 in the pool, and each player would be able to take home $24. Participants should try to maximize both individual and group outcomes, which causes a tradeoff. They are given a limited amount of time (e.g., 10 minutes) to play the game.
Putting food into categories	Choi and Thompson 2006	Participants were provided with a shared list containing several (e.g., a dozen) fruit and vegetable items. For example, you might have a list with the words, "orange," "grapes," "peach," "apple," etc. Everyone in the team sees the same list, and the contents of the list stay the same for a given round.	The goal was to generate as many different ideas as possible for categories that could be used to divide the food items. These ideas are purely abstract concepts meant to demonstrate creativity: e.g., "how many ways can you think of to put these food items into different categories?" For example, they could divide the fruit into citrus vs non-citrus, tropical vs non-tropical, seeds versus no seeds, or any number of possibly wild or imaginative ideas (things you can feed an elephant vs things you cannot feed an elephant). Participants must come up with as many criteria for dividing the food items as they can within a limited amount of time. To submit, they simply type the ideas into the system. Those working together can see all ideas generated by everyone in the team so far. Finally, credit is awarded for the total number of non-redundant ideas that participants came up with. That is, participants would only get credit once even if they submit the same idea multiple times, but it is OK if a different team also submitted the same idea.

Table 2: 102 Tasks in Task Map			
Railroad Route Construction game	Traeger et al. 2020	Participants see a shared visual of a map, consisting of a grid in which each square is a place where players can place a "railroad track" piece. Some parts of the map are blocked off — for example, because there is a tree or other obstacle in that location — and therefore railroad tracks cannot be placed there. There is a "start" and "finish" marked on the map. Participants also get a collection of several different types of railroad tracks. Some track pieces are bent like an elbow, and there are versions that bend downwards, bend upwards, and so on. Other pieces are straight. Participants are given enough pieces to be able to build a railroad that connects the start and finish.	The goal of the game is to build a railroad using the given puzzle pieces to connect the starting point to the finish point. Participants have to avoid the obstacles (places where you cannot place railroad tracks) — therefore, they need to use a mixture of bent tracks and straight tracks to navigate the map. Any path that runs smoothly from the start to the finish counts as a success, so there is no specific final path or "ideal" route. Participants must finish the game before the time runs out, but are able to visually see their progress on the screen throughout the game (since everyone is clicking, dragging, and interacting with it directly). Participants are evaluated on whether or not they are able to connect the start to the finish: if they do not make it in time, they receive no credit. There is no partial credit for building only part of the railroad.
Railroad Route Construction game (Impossible Version)	Traeger et al. 2020	Participants see a shared visual of a map, consisting of a grid in which each square is a place where players can place a "railroad track" piece. Some parts of the map are blocked off — for example, because there is a tree or other obstacle in that location — and therefore railroad tracks cannot be placed there. There is a "start" and "finish" marked on the map. Participants also get a collection of several different types of railroad tracks. Some track pieces are bent like an elbow, and there are versions that bend downwards, bend upwards, and so on. Other pieces are straight. Importantly, participants do NOT have enough pieces to successfully build a railroad that connects from the start to the finish. A piece is missing, which makes this game impossible to complete. However, participants are not told that a piece is missing.	The goal of the game is to build a railroad using the given puzzle pieces to connect the starting point to the finish point. Participants have to avoid the obstacles (places where you cannot place railroad tracks) — therefore, they need to use a mixture of bent tracks and straight tracks to navigate the map. Any path that runs smoothly from the start to the finish counts as a success, but since there is a piece missing, there is actually no solution to this puzzle. Participants aim to finish the game before the time runs out, and are able to visually see their progress on the screen throughout the game (since everyone is clicking, dragging, and interacting with it directly). Participants are evaluated on whether or not they are able to connect the start to the finish: if they do not make it in time, they receive no credit. However, since there is no partial credit and the game is missing a critical piece, all participants lose. There is no way to win.
Random dot motion	Moussaïd et al 2017	Participants see an image with many moving dots on the screen. Some of the dots are moving consistently in the same direction ("correlated dots"); other dots are simply bouncing around in random directions ("random dots"). The percentage of dots moving in the same direction varies, which makes the task easier or harder. For example, if 50% of of the dots are correlated dots, then it is easy to observe the direction that the correlated dots are "flowing" in; if only 5% of the dots are correlated, than the picture looks mostly just like dots bouncing randomly, and it is harder to discern the direction that the correlated dots are moving in.	The goal is to correctly determine the direction that the correlated dots are moving in. Participants submit their answer by using their mouse to place an arrow that points in the main direction of motion. They must submit their answer in a limited amount of time. There is an objectively correct answer, since the system is designed for the correlated dots to be moving in a specific direction. Participants are evaluated based on how close they are to the objectively correct direction. Their score is calculated based on the angle that their direction forms with the true direction; if they get the direction exactly right, then the angle is 0 degrees, so they would get the highest possible score. If they are off by a small amount (e.g., the angle is 10 degrees), they would get more points than if the participants are off by a significant amount (e.g., the angle is 90 degrees).
Rank cities by population, rank words by familiarity	Shaw 1963	Participants receive a list of cities (e.g., South Bend, Little Rock, Jacksonville, Portland, Charlotte, Lowell) or a list of words (e.g., Uncle, Kennel, Effort, Money, Village).	The goal is to rank the cities by their population according to the most recent population census, and to rank the words by familiarity to people in the participants' country. The rank of 1 is given to the city with the largest population or to the most common word, the rank of 2 is given to the second largest city or second most familiar word, and so on. Participants must do this task as correctly and accurately as possible and within a fixed amount of time. Participants are evaluated based on how correct their ranking is compared to the "ground truth" data, which is either the population census or the data on familiarity with certain words within a nation. The maximum score is achieved if the ranking is fully correct and the score is maximized by assigning as many correct rank values as possible.

Table 2: 102 Tasks in Task Map

Ravens Matrices	Weidmann and Deming 2020	Participants are presented with a pattern, which looks like a matrix or grid with a white background. There is a missing piece to the pattern. Here is an example to help you imagine what the pattern might look like. Participants could see a grid with four spaces. In the first space they see a square, divided into four pieces, with one piece colored in. Then, in the second space, there is a square with two of four pieces colored in. In the third space, there is a square with three of four pieces colored in. After the first three steps in the pattern, there would be a blank space on the grid. Participants would then see a few multiple-choice options for what could fill the blank space. For example, they could choose between a square with one, two, three, or four pieces filled in. All participants working together see the same pattern.	The goal is to complete the pattern by identifying what the missing piece should be. In the example above, the correct answer is that the fourth space should have a square with four pieces colored in. (That is, successful participants would notice that the first square had one piece colored in; the second had two colored in; the third had three colored in; so, naturally, the fourth should have four colored in.) The problem is multiple-choice, so participants will choose just one correct answer from a variety of options. There is always exactly one correct answer in the multiple-choice problem. Participants must make their selection within the time limit.
Reading the mind in the eyes	Almaatouq et al. 2021	Participants were given 36 pairs of eyes. Each pair is a photo of a person, cropped so that only the eyes are visible. Each pair of eyes is intended to convey some sort of emotion — some of them are squinting with suspicion or wide-eyed with fear; some of them have their eyebrows raised in surprise; others have crinkles at the corners to suggest a smile. Below the picture of the eyes, participants see four words describing different emotions: for example, "playful," "sad," "irritated," and "comforted."	The goal is to correctly label the eyes with an associated emotion by selecting the right one out of the four words. Each pair of eyes has an objectively correct label out of the four given. Therefore, participants' goal is to answer all questions correctly. They achieve the highest possible score if they label every pair of eyes with the right emotion. Participants have a limited amount of time to finish the test.
Recall association	Takahashi 2010	Participants are given a list of several words to study. Example words on the list include "spider," "foot," "pen," "car," etc.	The goal is to recall as many words as possible from the studied list within a fixed amount of time. Participants receive the maximum score for recalling all words correctly and the minimum for recalling none. Participants maximize their score by recalling as many words as they can.
Recall images	Weidmann and Deming 2020	Participants receive a certain number of images to memorize on individual screens.	The goal is to memorize all target images within a fixed amount of time and to be able to identify them later from a lineup of old and new images. For example, participants can be given 20 seconds to memorize six target faces and then asked to identify those target faces at a later time. At the later time period, they could be consecutively shown 15 sets of three faces (6 previously seen, and 39 new), and asked to pick out which faces were the images that they had previously seen before. Participants are evaluated on the number of correctly recalled images. They earn the maximum score when they recall all images correctly (e.g., select the 6 original faces out of 45). This means they need to correctly identify all the images that they had seen before, and avoid mis-identifying a new image as something they had previously seen.
Recall stories	Johansson, Andersson, and Rönnberg 2005	Participants read one or more stories that are each approximately 4 pages in length. After reading each story, they advance to a different page (where they are no longer able to reference the story), and they see multiple (e.g., 12) reading comprehension questions accompanying the story. Every question has an objectively correct answer based on the text.	The goal is to correctly answer knowledge questions about the stories within a fixed amount of time. Participants are evaluated on the number of correctly answered questions and they earn the maximum score when they answer all questions correctly. Since they cannot reference the stories, they must answer the questions based on their memory of what they read.
Recall videos	Engel et al. 2014	Participants are shown a 90-second video several times.	The goal is to correctly answer a set of questions about what occurred in the video within a fixed amount of time. Participants receive points

Table 2: 102 Tasks in Task Map			
			for each correct answer and maximize their score by correctly answering as many questions as they can.
Recall word lists	Takahashi 2010	Participants hear about a list of words. For example, they could listen to a tape of a male voice reading several words at the rate of one word per second.	The goal is to remember the list of words and then write down as many words from the list as possible. Participants are required to start their list with the last word they heard, or else it is not counted. Otherwise, words can be written in any order. Participants have a limited amount of time to recall all the words they can. The maximum score occurs if participants recall all words correctly. They get zero points for recalling no words, or for failing to start the list with the final word. Participants lose points for incorrectly recalling words that were not on the list. If working in a team, participants can collaborate to submit the final list of recalled words.
Reproducing arts	Woolley et al. 2010	Participants are given a shared online spreadsheet tool that can be edited, as well as a shared, static copy of a spreadsheet with each cell colored in a different way (the "target" spreadsheet).	The goal is to reproduce the "target" spreadsheet in the editable spreadsheet. That is, within a fixed amount of time, participants have to make the shared, editable spreadsheet look like an exact copy of the pattern of colors in the "target" spreadsheet. Participants receive points for coloring a cell correctly and don't receive points for coloring a cell incorrectly. They get the maximum score on the task if they reproduce the exact copy of the "target" spreadsheet and they try to get the highest score by correctly coloring as many cells as they can.
Room assignment task	Almaatouq et al. 2021	Participants are given a set of people (students), rooms, and rules (constraining conditions).	The goal is to assign students to rooms within a fixed amount of time in order to maximize the students' utility as much as possible while also respecting all of the constraining conditions (for example, one rule might say that certain people are not allowed to be in the same room). Students get a specified amount of utility from being assigned to a particular room, which translates to points in the game. Participants will be shown the running total of how many points their current submission has earned, but they will not be notified if they have achieved the optimal arrangement. If participants' submission violates some of the constraining conditions, they will receive a score penalty.
Run a mini business	Shaw 1963	Participants are given information on running a tinker toy manufacturing company. The information is organized in the form of specialized tables that have parameters for the tinker toys, as well as details on assembling different models, order forms, costs, and prices. In addition, the tables reflect how costs and selling prices fluctuate over time. Participants can view these tables and run their business through an interface on the computer. In the beginning, participants are given a base pay to start off their business. As participants play, they can lose and gain money.	The goal is to make as large a profit as possible from running the business for a fixed amount of time. Based on the information provided to them, participants can "buy" parts, "manufacture" products, and "sell" them at the price they determine. Participants are evaluated based on how much money they have by the end of the game's time limit. The less money they lose and the more money they make, the more successful their performance is on the task.
Search for Oil Task	Isenberg 1981	Participants view a shared 12x12 grid of points, creating a digital "map". Each point represents a section of "land" in a theoretical oil field. A limited number of these points have been pre-determined by the makers of the game to contain "oil," while others contain nothing. Participants are also given some information about which points might be more likely (but not certain!) to contain the oil. For example, they are told that some parts of this grid represent land with the appropriate chemical composition, making it more probable that there will be oil there. (Other pieces of information that make oil more or less likely include information about the	The goal of the task is to decide which of the 144 sections of land is suitable for drilling oil, and to do so as accurately as possible before the game's time limit runs out. In other words, participants have to remember and account for all the information they were provided about where oil is most likely to be found, and then decide where they want to "drill" (by clicking on the point on the map). Participants aim to get the highest possible score, and they earn points for every location with oil that they successfully find. They lose points if they select locations that do not have oil. Therefore, one would get the maximum score of this game by correctly finding every point with oil without making any mistakes.

Table 2: 102 Tasks in Task Map

		surface hardness, surface mantle thickness, and geological stratification of the area.) After being shown the detailed information for a limited amount of time, participants have to remember as much information as they can and then decide where they want to "drill" completely from memory. During the main gameplay, participants see just the 12x12 map, where they are able to click on each point if they want to drill there. As participants select which locations they want to drill, they are given feedback after each decision. For example, after participants pick a point on the 12x12 grid to drill oil, they will be told whether they were right or wrong about oil being there, and they will see the running total of their score.	
Sender-Receiver game	Gneezy 2005	Participants are paired for a communication-based game where one player sends a message to another, and the second player then chooses an action that determines their payoff. Player 1 is told to send a message to Player 2 about two options for winning different payoffs. The two possible messages are: "Option A will earn you more money than Option B" and "Option B will earn you more money than Option A". Player 1 can choose to tell the truth or to lie to Player 2, and Player 2 can choose to heed or to ignore the advice of Player 1. The choice of the receiver will determine the payments of both players in the experiment. Here is a concrete example. Suppose that the true options are: (Option A) Player 1 gets $5 and Player 2 gets $6; (Option B) Player 1 gets $6 and Player 2 gets $5. Player 1 will see the true options, and they can either choose to tell Player 2 the truth (saying, "Option A will earn you more money than Option B") or lie (saying, "Option B will earn you more money than Option A"). If they tell the truth, and Player 2 listens to them, then Player 1 will end up with less money in the end ($5 rather than $6). On the other hand, if they lie, they can potentially influence Player 2 to (unknowingly) choose the option that is better for Player 1. Meanwhile, because Player 2 knows that Player 1 may be lying, they can choose to either follow the advice or to ignore it. In the end, both players receive the payoffs from Player 2's choice, but Player 2 is not informed of how much Player 1 received (so they do not ever find out whether they were "correct" or not.)	Each participant's goal is to maximize their own payoff. Therefore, Player 1 must decide which message will lead to Player 2's choosing an advantageous outcome; Player 2 must decide whether or not to trust Player 1, so that they can obtain their own advantageous outcome. Participants are given a limited amount of time (e.g., 10 minutes) to send a message and make a decision.
Shopping plan	Woolley et al. 2010	Each participant has a grocery list and a map that shows distances and times between each grocery store, as well as a list of potential items to purchase and how many points they are worth.	Participants have to plan a shopping trip as though they were all residents of the same house sharing the same car. The goal is to purchase as many high-quality items as possible in a fixed amount of time, and to get the highest number of points by considering tradeoffs between price, quality, and driving time. Participants gain points for every item that they are able to plan to purchase. Getting anything less than the maximum number of points will lead to partial credit.
Space Fortress	Arthur et al. 1993	Participants see a video game, in which a "space fortress" is at the center of the screen. The fortress can rotate in all directions and fire off "shells" at anyone who attempts to attack it. In this game, each subject controls a spaceship, which is able to fly around freely and shoot missiles at the space fortress. As they approach the space fortress, subjects must also dodge "mines," which periodically appear on the screen and chase the subject's ship. If a mine hits the subject's ship, the spaceship will suffer damage. However, some special floating objects — which appear to be mines — will actually give the subject a "power-up" and damage the space fortress instead.	The goal of the game is to maximize points by destroying the space fortress. Participants need to attack the space fortress as much as possible while avoiding damage from enemy attacks and mines. This game can be either single-player or multiplayer; in either case, the objective is the same. To summarize, participants earn points when: • They successfully hit the fortress; • The fortress is destroyed (which requires hitting it repeatedly and progressively weakening it); • Participants earn a bonus in the game (e.g., through a power-up);

Table 2: 102 Tasks in Task Map

			• The participants destroy or neutralize a mine. On the other hand, participants lose points when: • Their ship is damaged (e.g., by an enemy shell or a mine); • Their ship is destroyed (which occurs after the ship suffers too much damage); • The participant runs out of missiles. Participants are evaluated on their final game score.
Splitting a deck of cards	Laughlin and Shippy 1983	Participants are shown cards from a standard deck, one by one. As the participants see the cards, they are told whether the card belongs to a "rule" or not. From the instructions, participants know that there is a rule dividing the 52 cards into two categories, but they do not know what the rule is. The rule can be based on the numerical value of cards; their color, suit, logical connectives, alternation, or any combination of these. For example, a rule could be "the card is divisible by 3," or "two black cards and one red card alternate." The fist card card shown is always an example of the rule (it's an exemplar). As participants see more cards, they will see some exemplars, and some cards that do not belong to the rule (non-exemplars).	The goal is to correctly figure out the rule dividing the deck of cards in as few trials as possible and within a fixed amount of time. As participants keep seeing new cards one-by-one, they make proposals as to what the rule is and receive immediate feedback. Cards that have been shown get sorted into exemplars and non-exemplars and stay in front of participants so that they can see patterns. To get a better score on the task, participants have to understand what the pattern is as fast as they can.
Sudoku	Engel et al. 2014	Participants see a shared online system that displays a Sudoku puzzle. Sudoku is a logic puzzle, in which there is a 9x9 grid, which is divided into nine 3x3 subgrids (also known as "boxes" or "regions"). Each cell in a region has a true underlying value from 1-9; however, most of the cells appear empty to the player, and only a few of them have the correct numbers filled in.	The goal is to solve the Sudoku puzzle within a limited amount of time. This means that participants need to fill in each blank cell with the correct underlying number from 1-9; in addition, they must satisfy the constraint that every row, every column, and every 3x3 region can only show each digit from 1-9 exactly once. For a given Sudoku puzzle, there is only one correct answer. Participants either correctly identify the underlying value for every cell, or they fail the Sudoku game. Finally, participants working in groups may use a chat function to discuss which value should be placed in a given cell.
Summarize Discussion	Hackman, Jones, and McGrath 1967	For this task, participants are given an open-ended discussion statement about a specific topic. The statement may be a question like, "what makes for success in our culture?" or "should birth control be made available to anyone without a prescription?" Everyone in the same team gets the same question.	Participants are asked to talk about and submit a response to the discussion statement — that is, they need to turn in a written summary of their arguments or answers to the question given to them. (For example, "Yes, birth control should be made available to anyone for three reasons, Number 1…" or "No, we should not make birth control available to everyone. You should need a prescription because …") If people disagree, the summary needs to resolve and synthesize the pros and cons. Participants have a specific time limit for completing the writeup. They are graded on the quality of their final written summary.

Table 2: 102 Tasks in Task Map			
Target Search	Shaw 1963	Participants view a labeled spreadsheet with 100 squares on it. The columns are given letters and the rows are given numbers (so, each blank cell has a letter-number coordinate, like A5, E.10, etc.). While these cells on the spreadsheet appear blank to the participant, each one is actually connected behind the scenes to a specific value that only the experimenter can see. On the experimenter's side, there is essentially an invisible "target" — the cells are grouped into concentric circles. This "target" shape determines the point value of each cell. Cells located in the middle of the target are worth more points than cells near the edges of the target. The closer to the center of the target a cell is, the more points it is worth. If a cell is outside of the target entirely, it is worth 0 points. For example, imagine that E.10 is the center of the target. Then E.10 is worth the most points. A cell that is next to the target (E.11) is worth the second-most number of points, whereas a cell that is far away from the target (e.g., A5) would be worth fewer points (or zero, if it is outside the target). Of course, since the participants do not see the target, they have no idea how many points each cell is worth.	The goal is to get the most points. In other words, participants need to find the coordinates of as many high-value cells as they can within a fixed amount of time and under a limited number of allowed guesses. For each guess, participants submit the coordinate of a blank cell (e.g., A5) and then get feedback about how many points they received. Based on feedback, they would try to figure out where the experimenter's underlying "target" is and look for the high-scoring cells. If they are working in teams, participants work together to determine their next guess. After a few guesses, participants can often "decipher" the logic of the sheet and will then be able to name the cells corresponding to the largest values. This is why this game is called "target search" — participants are using each guess to learn about where the target is, and to search for the highest-value cells. Participants are evaluated based on the values of the cells they name. The score is maximized by naming as many cells with as large values as possible. Theoretically, the best possible score would be to find the center of the target on your first try and then only name the highest-value cells, but this is not likely to happen; most participants will experience some trial and error.
The beer game	Chen and Samroengraja 2009	Participants are assigned to a group of four, in which each player has a different role: (1) retailer, (2) wholesaler, (3) distributor, and (4) factory. Customer demand (in kegs of beer) arises only at the retailer, which replenishes its inventory from the wholesaler; the wholesaler gets its beer from the distributor; and the distributor gets its beer from the factory, which produces the beer. The customer demand in different periods is a random variable, independently drawn from a distribution that is known to all players. The beer flows upstream to downstream (i.e. from the factory, to the distribution center, to the warehouse, and finally to the retail store). In contrast, the information about the customers' orders flows in the opposite direction (it goes from the retail store back up to the factory). Both the material and information flows are subject to delays. For example, there is an order processing delay, or information lead time, between when an order is placed and when the supplier receives the order. Here is an example. Suppose the retail store orders 10 kegs from the warehouse on Monday. The warehouse receives this order on Wednesday. This information delay is due to the administrative steps in processing an order. On Wednesday, however, the warehouse only has 5 kegs of beer, so it ships 5 kegs to the retail store and backlogs the remaining 5. This shipment of 5 kegs arrives at the retail store on Friday. This delay is due to transportation. In each period, the channel members must decide how much, if any, beer to order from their respective suppliers. The factory must decide how much, if any, beer to produce.	Participants are instructed that their goal is to minimize the total holding and backorder costs incurred in the entire supply chain and optimize the system-wide performance. In other words, the goal is for each member to choose the right amount of beer to order or produce, despite the lag in information. If participants get the number wrong, they suffer a penalty. If they have too much beer, the wholesaler, distributor, and factory suffer a "holding cost," because it is expensive to store the extra inventory. If they have too little beer, the retailer suffers a "backorder cost," because it is expensive to lose customers when the store does not have the beer people want. In an ideal world, if the entire supply chain operates perfectly smoothly, the cost incurred would be zero.
The Fish game	Krafft et al. 2015	The participants see a shared virtual canvas with many different cursors moving around. Each participant controls their own cursor (which is denoted by a unique color and the word "YOU"). Participants are able to use the controls to move their cursor and "walk" around the virtual space. An underlying system sets a "point value" for each location in the space. That is, walking to specific locations in the virtual space will earn the participant more points. Participants will be able to see how many points their current location gives them, and they will also see the running total of how many points they earned so far. However, participants do not know where the high-point-earning locations are, so they must wander around the virtual room. Participants lose all their points if they touch the wall.	The goal is to maximize points in the game by discovering and moving your cursor to the highest point-earning locations in the virtual space. Participants try to get as many points as possible before the time limit expires. This essentially involves a bit of trial and error, since participants only see how many points their current location is worth. There is also an element of understanding risks and tradeoffs: participants have to decide whether to explore other locations in the room (at the risk of only finding lower-point areas) or just stay put and earn however many points their current location is worth (at the risk of not finding the highest point-earning location). When playing with others, participants may get a sense of where to go by observing where other people in the room are heading.

Table 2: 102 Tasks in Task Map			
The N light bulbs game	Yahosseini and Moussaïd 2020	Participants see 10 light bulbs. Some of the light bulbs can be turned "on" or "off" by clicking on them. Other light bulbs are stuck — they are either on or off, and they cannot be changed. The light bulbs that are stuck are marked with a big "X." Different patterns of light bulbs are worth different numbers of points. For example, maybe an alternating pattern — on, off, on, off, etc. — has the highest points for this round of the game. However, participants do not know that this is the best pattern.	The goal is to get the most points by setting the lightbulbs to a pattern that maximizes the point value. This means participants have to switch the light bulbs (that are not stuck) on and off to try to figure out which configurations have the highest payoff. For each guess, participants only get to change one light bulb (turn it "on" or "off") at a time. There are a limited number of guesses. If working in teams, everyone collectively determines the next guess. The computer will give feedback and tell the participant how many points they earned after each guess. While there is a theoretical "best" configuration, participants will not be told if they have achieved it; they only know how many points the guess is worth. It's possible, for example, that they make a guess that gets to the optimal point value, but then make another guess that performs worse. Participants also have the option to end the round early if they are happy with their guess. A final useful piece of information is that the lightbulbs can be thought of as binary variables (a 0 or 1), which would then allow this game to be played by an algorithm.
To evacuate or not to evacuate	Shirado et al. 2020	Participants read about a situation in which 'a disaster may or may not strike.' They are given information about the risk level of the disaster, as well as a button to "evacuate" from the disaster. Evacuating from the disaster costs money — participants lose half their bonus if they choose to evacuate. However, if they don't evacuate and the disaster happens, then they lose their entire bonus. Therefore, this task is about weighing up the different risks and rewards of clicking the "evacuate" button. Participants only find out whether the disaster happened or not at the very end of the game. All participants are equally affected by the disaster, but they make their selection for evacuation individually.	The goal is to make a decision regarding whether or not to evacuate. Participants have a limited amount of time to make their decision. Choosing NOT to evacuate means that there is high risk (you could get nothing), but potential for a high reward. On the other hand, choosing to evacuate has a low risk but a smaller reward.
TOPSIM - general mgmt business game	Frick et al. 2017	Participants are randomly assigned to a team and manage a fictitious company. They compete against other teams which manage other companies in the same simulated market. Over each period (the eight weeks), teams have to develop and adapt business strategies based on the basic endowments in the different functional areas of their company and based on the economic situation in their respective market. In each period, teams have to decide on a number of variables driving company performance: the sales price(s) of the product(s), marketing and sales activities, R&D expenses, investment in environmental facilities or process optimizations, and financial resources. At the end of each period, each company's periodic share price will be calculated based on how the company as well as its competitors decided on the above-described variables given the predetermined endowment and economic circumstances. The group then receives a full report documenting how the team performed as well as an updated economic forecast for the subsequent period in order to make new business decisions.	Participants are instructed that their goal is to maximize the firm's share price at the end of the eight-week period. The final share price translates into a team grade with the best possible grade going to each market's best-performing company. The remaining companies are graded relative to the best performers. In other words, the highly competitive setting incentivizes teams not to cooperate with competitors.
Trivia Multiple Choice Quiz	Woolley et al. 2010	Participants receive a series of multiple choice trivia questions. For example, participants can be asked about the average weight of an elephant and be given options of 9,000 lbs, 5,000 lbs, and 15,000 lbs.	The goal is to correctly answer as many questions as possible from a multiple choice trivia quiz within a fixed amount of time and without looking up the solutions online. Participants don't get any points if they answer a question incorrectly and they get points for choosing the correct answer. Participants maximize their score by answering as many questions correctly as they can and get the maximum score if they answer all questions correctly.

Table 2: 102 Tasks in Task Map

Typing game	Woolley et al. 2010	Participants see an image of a document — we can call this the "Document to Copy." This document contains several passages of relatively complicated text. The Document to Copy cannot be edited, copied, or pasted. Everyone working together sees the same document. Participants also have a shared, editable document, where they are able to type.	The goal is to exactly reproduce the words from the Document to Copy in the editable document. The focus is on typing words correctly: participants earn points for every word they correctly type, and they lose points for skipping words or for making typos. Participants have a limited amount of time to type as many words from the Document to Copy as they can. If working in a group, everyone types into the same shared document, and therefore team members must also coordinate to avoid typing over each other or missing certain sections.
Ultimatum game (various versions)	Bornstein and Yaniv 1998	Participants are placed into pairs and randomly assigned the role of "proposer" or "responder." The proposer will receive an endowment that they may split with the responder. The proposer will make a decision as to how they would like to split it, which is then shared with the responder. The responder can then choose to either accept or reject. If accepted, the sum is split as was proposed, but if rejected, both players receive nothing. Players are made aware of the payoff arrangement prior to making their decisions. The same pairs play together for a predetermined number of rounds and roles are reassigned each round.	Each participant has the goal of maximizing their earnings. In the proposer role, participants must consider how much they should offer in order to keep as much as possible while ensuring the offer will still be accepted. In the responder role, participants must consider how they will respond to "unfair" or "low" proposals and the impact this will have on their payoff in the remaining rounds. Both must also consider their reputation and relationship with their partner throughout the game and the impact of these factors on their earnings. Participants must also make their proposals and response decisions in a fixed amount of time.
Unscramble words (anagrams)	Engel et al. 2014	Participants see a shared list of 24 randomly scrambled words. For example, a word on the list could be, "SEEMTRIMUM." This is an anagram (a word created by scrambling the letters) of the word "SUMMERTIME."	The goal is to unscramble as many words from the list as possible in a limited amount of time. Each of the scrambled words has exactly one correct answer. Participants obtain the highest score if they are able to unscramble all the words; otherwise, they earn points for every word that they are successfully able to unscramble. Those working in a team are able to collaborate by discussing the words before they submit.
Visual Oddball Target	Hertz et al. 2016	Participants see several groupings of visual stimuli. For example, they might get a series of Gabor patches. These are circular-shaped patterns with black and white bars oriented in the same direction; the patterns are often used in psychological research because they can create an optical illusion, making for a challenging visual perception task. The participants might get several groups of 6 such patches. Sometimes, one of the visual items might be an "oddball" (it looks different than the rest). In this case, if participants see 6 patches at a time, the oddball would be a Gabor patch with different levels of contrast (it's darker) or have a different pattern than the other 5.	The goal is to visually identify when there is an oddball versus when there is not. The task is focused on visual perception: participants see a grouping, and then need to decide whether one patch looks different than all the others. They submit their decision by pressing a button using their mouse or a keyboard within a fixed amount of time. For example, if participants see a group of 6 Gabor patches, they would press a mouse or a keyboard button when there is an oddball and press nothing when all patches in the group look the same. There are only two possible outcomes of task performance. Participants can either correctly identify the oddball or not.
Volunteer Investment Game	Babcock et al. 2017	Participants are randomly and anonymously assigned to groups of three in each of the ten rounds. Members of the group are given a limited amount of time (e.g. two minutes) to make an investment decision. Individual earnings are $1 in the event that no one invests before the end of the game. Once one group member makes the investment, the round ends. The individual making the investment secures payment of $1.25, while the other two group members each receive $2. The investor is randomly determined in the event that multiple parties simultaneously invest. Participants are not allowed to communicate with other group members in any way.	Participants are instructed that their goal is to maximize their individual payoff as well as the group's total payoff. It means that one goal is to maximize their personal earnings, which means that they should not invest anything; rather, they should wait for others to invest. But another goal is to optimize the group's outcomes, which means that at least one person in the group should invest. Participants should try to maximize both individual and group outcomes, which causes a tradeoff. They are given a limited amount of time (e.g., 2 minutes) to play the game.

Table 2: 102 Tasks in Task Map

Wason's Selection Task	Wason 1968	Participants are given a conditional sentence in the form of, "If X, then Y." They also see four cards, which have either a letter or a number on the front. The cards are double-sided; those with a letter on the front have a number on the back, and vice versa. As an example, imagine that the conditional sentence says the following: "If there is a D on one side of any card, then there is a 3 on its other side." Then, imagine that the four cards are: • D (front), 3 (back) • 3 (front), K (back) • B (front), 5 (back) • 7 (front), D (back)	The goal is to identify which cards, if turned over, would allow the participant to determine if the conditional sentence is true or false. Participants make the selection by clicking on the appropriate cards on the interface. Consider the example above: "If there is a D on one side of any card, then there is a 3 on its other side." This means that if there is a card with D on one side, and something other than 3 on the other side, it would make the sentence false. Thus, the answer for each card is as follows: • Card #1: We should select this card; if there is something other than 3 on the back, then the statement would be false. • Card #2: We should NOT select this card; whether the other side is D or not does not affect the truth of this statement. If it is D, then the statement is true; if it is not D, it is simply neutral evidence and does not disprove the statement. • Card #3: We should NOT select this card; the statement does not say anything about cards with letters other than D. • Card #4: We should select this card; if there is a D on the back, then the statement would be false (since D would be paired with a number other than 3). Selecting Card #1 is an easy decision: since it has a D on the front, it is easy to see that we should check for a 3 on the other side. However, some of the other choices are more difficult and require skillfully applying logic: for example, participants may not notice that they need to check Card #4 to make sure there is not a D on the other side. Participants may also mistakenly choose cards #2 or #3. For a given sentence, there is a single correct solution (in the example above, choosing both #1 and #4). If participants correctly select exactly the right cards, then they solve the task correctly. If they do not select enough cards (e.g., picking #1 without picking #4) or if they select too many cards (e.g., picking #1, #2, and #4), then they fail the task.
Whac-A-Mole	Naber, Pashkama, and Nakayama 2013	Participants see a shared screen with a bunch of colored moving targets (circles) and an ability to click or "hit" the targets as they move around. Each participant can control their own individual mouse, but everyone sees the same moving targets.	The targets are worth different numbers of points based on their color. Participants' goal was to hit (e.g., click on) targets based on their point value, thereby maximizing their points. They have a limited amount of time to get as many points as possible.
Wildcam Gorongosa (Zooniverse)	Straub, Tsvetkova, and Yasseri 2023	Participants view a series of pictures of animals taken by motion-detecting cameras in Gorongosa National Park in Mozambique. Participants view and interact with the pictures through an online interface on the Wildcam Gorongosa site.	The goal is to classify as many pictures as correctly as possible according to five criteria within a fixed amount of time. For each picture, participants are tasked with: (1) detecting the presence of the animal(s), (2) identifying the species of the animal(s), (3) counting how many animals there are, (4) identifying the behaviors exhibited — specifically, identifying whether the animal(s) is (a) standing, (b) resting, © moving, (d) eating, or (e) interacting (multiple behaviors may be selected), and (5) recognizing whether any young are present. Each picture has a "ground truth" classification done by scientists, which is used to calculate correctness. Participants are evaluated based on the total number of pictures they classified out of all possible pictures in the task, as well as the number of correct classifications out of all the classifications they managed to do. This means that both quantity and quality matter. In other words, participants maximize their score by classifying as many pictures as correctly as they can.

Table 2: 102 Tasks in Task Map

Wildcat Wells	Mason and Watts 2011	Participants see a realistic-looking 2D desert map with dots on it that represent hidden oil fields. All dots look the same to participants, but some "wells" are wider and deeper than others and contain more oil, which earns the participant more points. As they play, participants can see the total number of points that they have accumulated throughout the game.	The goal is to drill as much oil as possible within a fixed amount of time by choosing the dots corresponding to "wells" that the participant wants to select. Participants earn points proportional to the amount of oil they discover. Thus, the goal is to try to find the "wells" with the most oil before the time runs out.
Wolf, goat and cabbage transfer	Kennedy 2009	Participants are given a problem that reads as follows: "Once upon a time a farmer went to a market and purchased a wolf, a goat, and a cabbage. On his way home, the farmer came to the bank of a river and rented a boat. But crossing the river by boat, the farmer could carry only himself and a single one of his purchases: the wolf, the goat, or the cabbage. If left unattended together, the wolf would eat the goat, or the goat would eat the cabbage. The farmer's challenge was to carry himself and his purchases to the far bank of the river, leaving each purchase intact. How did he do it?"	The goal is to achieve the fewest number of trips that get the wolf, goat, and cabbage from one river bank to another, and to solve this problem within the time limit. There are a few different possible solutions that get to the lowest number. All of the solutions require the participants to take the goat over first, because any other action will lead to something being eaten. Once the farmer comes back, though, there is a dilemma, because if he takes the wolf over, the wolf would eat the goat on the other side; if he takes the cabbage over, the goat would eat the cabbage on the other side. The "trick" to this problem is to realize that the farmer can actually take the goat BACK on his return trip: that is, he takes either the wolf or cabbage with him on the second trip. Then, he takes the goat BACK to the ORIGINAL side. He then picks up the remaining one of the cabbage or wolf, and finally, he goes back with the goat. Realizing that you can actually bring things back and forth (in both directions), rather than just in one direction across the river, usually requires some out-of-the-box thinking. Therefore, participants must satisfy the constraints of the problem while also thinking creatively. As a final piece of helpful context, this task is actually a well-known one that many people have built algorithms to solve. To summarize, the solution to this task is: • Take the goat to the other side • Return to original side (where the wolf and cabbage are) • Take either the wolf or cabbage to the other side • Return to the original side with the goat [THIS IS THE TRICK! Realizing this prevents either the wolf eating the goat or the goat eating the cabbage.] • Take the cabbage or wolf over to the other side • Return to the original side, where only the goat remains. • Take the goat over
Word completion given part of word	Woolley et al. 2010	Participants get a set of 36 words with 2-3 letters missing. Each word has a specific correct answer. For example, " _ ech_ _ que" would be "technique".	The goal is to complete as many words as possible within a fixed amount of time. Participants get the maximum score for completing all words correctly and maximize their score by completing as many of the words as they can.
Word completion given starting letter	Woolley et al. 2010	Participants are given a starting letter and ending letter for words. For example, the start letter could be "S" and the end letter could be "N." The starting and ending letters are selected such that there is always a nonzero set of English words that can be created.	The goal is to generate a list of valid English words using the starting and ending letter provided. Using the previous example, participants could be asked to generate words like "SPIN," which starts with "S" and ends with "N." Participants must generate the maximum number of valid words they can within a fixed amount of time. Nonvalid words will be rejected by the system.

Table 2: 102 Tasks in Task Map			
Word construction from a subset of letters	Finlay et al. 2000	Participants are presented with nine letters. For example: A, F, H, B, E, J, N, K, D.	The goal is to produce as many valid four-letter English words as possible from a given set of nine letters within a fixed amount of time. For example, out of the set of letters above, words such as "bean," "head,", "khan," "dean" and many others can be formed. Although there is a theoretical maximum score if participants exhaust all the possible combinations of letters to form four-letter words, the goal is simply to generate as many words as participants can within the time limit. Participants submit their words through an online system. Words that are not valid in English or that are longer than four letters will be rejected by the system. Those working in teams can brainstorm in any way they wish; the total number of non-repeated words submitted by the team will be counted.
Writing story	Hackman, Jones, and McGrath 1967	Participants receive a writing prompt. Examples include, "Write a story about this inkblot," or "Describe this mountain scene." Everyone in the same team gets the same writing prompt.	The goal is to write a response to the prompt. If working in teams, participants discuss and collectively produce the same story. The work must be submitted online within a fixed time period. Since this is an open-ended writing task, there are no right or wrong answers, and anything that addresses the writing prompt is a valid answer. Participants are evaluated based on the quality of their responses (for example, the length, writing quality, and level of creativity).

Table 2. Tasks included in the Task Map repository, with the source, stimulus, and goals.

Appendix C: Details of Task Rating Process

This appendix describes the procedure for rating the 102 tasks along 24 dimensions. We begin with some context of the early iterations of the task rating process (Appendix C.1), which led to key design decisions (Appendices C.2 and C.3) and informed our final procedure (Appendix C.4).

C.1. Early Iterations

C.1.1. Initial Reliability Issues. In an early version of this project, we began with a set of 71 task dimensions from our five focal frameworks. To test different operationalizations of theoretical constructs, we measured the same concept using different phasings. For example, we tested both “there is an **objective** solution to this task” and “there is an **optimal** solution to this task” — two wordings that express the notion of having a “best” answer. We asked pairs of trained research assistants to separately answer 71 questions about each task, then arrive at a consensus for the final rating.

However, we soon noticed that these ratings suffered from reliability issues:

1. **Subtle changes in phrasing made an enormous difference**: raters’ responses to the two questions above (“there is an objective solution” and “there is an optimal solution”) had significant but very low correlation ($r = 0.313$, N = 117, Kendall’s tau = 0.233, $p = 0.005$); just 52.14% of tasks had the same rating for both questions. In other words, depending on the wording a researcher happens to use, half of the tasks would have been placed into a different category.

2. **Similar tasks were projected into different locations.** The initial version of the Task Space did not pass an important sanity check: tasks that were intuitively similar (the NASA Moon Survival task and the Desert Survival task, which are the same activity with only minor details altered) were positioned far apart in the Task Space (Figure 1, left). The results suggested that the rating system could not reliably produce the same task ratings, given a similar input.

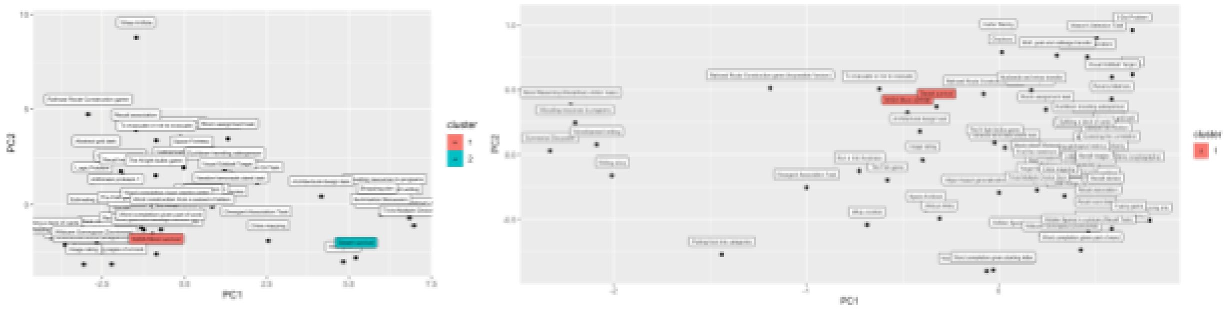


Figure 1. A two-dimensional projection (using PCA) of an early version of the Task Space (left), which included 71 dimensions, compared to a later version of the Task Space (right), with 23 dimensions. In the early version, a pair of similar tasks (NASA Moon Survival Task and Desert Survival Task) were positioned very far apart. In later versions of the map (right), these tasks were positioned much closer together.

C.1.2. Insights from Interviews with Raters. To better understand the source of the reliability issues, we interviewed each of the raters, asking them to explain their evaluation process step-by-step. Our interviews resulted in four key insights:

1. **Raters had inconsistent understandings of tasks:** In the early stages of our project, raters were instructed to read the description of a task directly from the Methods or Supplementary Material

sections of a research paper. However, research papers do not describe tasks consistently. Some embed the description of their task with information about the research design (e.g., experimental conditions, hypotheses) or information about the participants being studied; others describe tasks at only an abstract level; still others show images of the user interface, with step-by-step instructions. This unstructured and uneven presentation of task information made it difficult for raters to develop a consistent understanding of tasks, and a key reason why raters could not provide reliable ratings was because they did not have reliable inputs.

2. **Raters had different interpretations of words:** Our experiments with using variations of the same question revealed that different words suggest different connotations to raters. For example, we asked raters to explain their understanding of two questions, for which agreement was just 51.28%:

    1. The task can be completed by **adding together** the individual group members' efforts.

    2. The group's performance on the task is determined by the **sum** of the performance of each member.

    In a notable moment, two mappers confidently provided opposing interpretations of this question pair. Rater 1 argued that "sum" requires group communication and "add" involves pooling answers without communication:

    > "Yeah, I try to think of [sum] as like the group is able to communicate and work together to find a solution versus [adding], where they might be working individually and then pooling their solution in the end."

    Rater 2 argued that "add" requires group communication and "sum" involves pooling answers without communication (the opposite interpretation):

    > "Sum feels like there's points for everybody's answer, and then the group overall. … I feel like at the end, like this group needs this many points, and this person got this many with theirs at this point. So like, that's what sum feels like to me. But adding together individual efforts feels more like it includes the idea of communication within the group and people participating."

    Thus, a second explanation for low reliability was that the words used to operationalize the concepts were insufficiently precise, leaving raters to make their own assumptions. Over time, raters grew more confident in these assumptions, but most failed to realize that their underlying interpretations of questions did not align with those of their peers.

3. **The Frame Problem:** More generally, since it is impossible to perfectly describe every detail of a task or dimension, raters found themselves mentally "filling in the blanks," but doing so inconsistently. This is a phenomenon known as the *frame problem* (Watts 2014). The frame problem was especially salient for task dimensions relating to participant behavior — for example, "The task can be completed by adding together the individual group members' efforts." Here, raters struggled to imagine how participants completing the task would "add together" their contributions. They encountered similar problems when evaluating task dimensions involving the participants' intrinsic characteristics (such as their skill level or degree of interest in the task), since they had to place themselves in the mindset of an (imagined) research participant.
This insight from the interviews highlighted the fact that some task dimensions rely on characteristics of the participants and their behaviors, which are best evaluated by the *actual participants*, and not by a third party. Recognizing the challenges with such dimensions led us

eventually to adopt Larson's task definition as a combination of *stimulus* and *goals*, thus excluding any dimensions related to the participants and their behaviors.

4. **Disagreements in ratings are more fundamental than traditional inter-rater reliability.** Given these insights, we realized that our initial diagnosis of reliability issues only scratched the surface of the problem: operationalizing theoretical constructs is difficult because words always leave some room for imagination, and disagreement may not only reflect variance in applying the scale, but may also reflect *incompatible interpretations of the underlying construct* — and two raters may be reading the same words while applying entirely different ideas.

C.2. Key Design Improvements in Rating Process

Thus, our guiding principle in redesigning the rating process was to create as much consistency as possible, leaving minimal room for the raters' imagination. In the following version of the Task Space, we implemented four key changes:

1. **Using a consistent task format**: In this stage, we adopted a format that described tasks explicitly in terms of their stimulus and goals. We also used a checklist to ensure that each task description included consistent information (Appendix D).

2. **Removing task dimensions that cannot be answered using only a *task class-level* description of the stimulus and goals:** We removed questions that required information about the group process, the participants' characteristics, or any other information not contained in the stimulus and goals. For example, we removed questions about whether a task benefits from "adding together" contributions, or whether participants would find the task interesting.

3. **Using one question per concept:** Raters were confused by the seemingly redundant questions, and assumed that the subtle changes in language had different meanings. In this stage, we ensured that each concept or construct had exactly one question. Our final set included 23 questions.

4. **Using longer elaborations to clarify misconceptions:** We integrated insights and common misconceptions from raters into a long "elaboration" for each question, in an effort to publicly clarify assumptions.

C.3. Iteratively Improving Questions and Elaborations

To identify additional misconceptions and improve the general clarity of our questions, we piloted questions among both a sample of 8 researchers and 50 Amazon Mechanical Turk workers. For each question, we tracked three key metrics:

1. **Mean rater agreement** (the percentage of raters who selected the same answer for a given task);

2. **Rater confidence** ("How confident do you feel in your answer," 5-point scale);

3. **Open-ended suggestions** ("Did you have any trouble answering the previous question, or wish for more clarification in order to make answering the question easier? If so, please tell us more.")

We audited questions with the lowest agreement and confidence to improve the wording or expand the elaboration, and we incorporated suggestions from the open-ended box when relevant. Within the researcher sample, we were also able to hold follow-up conversations to discuss ratings and incorporate feedback in more detail.

In usability design, Neilsen (2000) claims that the majority of usability issues in a software application can be uncovered with just five users. Our process of uncovering "usability issues" in the rating questions exceeded this threshold more than ten times over, and gave us confidence that the resulting questions left minimal ambiguity.

C.4. Rating Process of Final Task Space

Our final rating process proceeded in two phases. Due to our substantive interest in group advantage, we initially limited the scope of our research to only the activities that could be played by both individuals and groups (so that individual performance and group performance could be directly compared). This limitation eliminated McGrath's Type 6 (Mixed-Motive), because it describes tasks in which each group member is assigned a unique role, interest, or motive. A researcher on the team hand-coded each task to determine whether the task belonged to this category, and a second researcher independently confirmed these ratings. This process eliminated 30 tasks from consideration. As a result, we conducted the first phase of ratings using 72 tasks and 23 dimensions.

In a second phase, we re-introduced the 30 tasks that had initially been eliminated, and our panel rated them along the same 23 dimensions. We also included the 24th dimension, which is 0 for the 72 tasks included in the first phase, and 1 for the tasks introduced in the second phase. This variable is effectively a binary indicator that separates tasks rated in the first phase from tasks rated in the second phase, and it is also an instantiation of Mixed-Motive tasks as defined by McGrath.

In this section, we describe in detail the rating process for the 23 dimensions that were rated by the Amazon Mechanical Turk panel.

C.4.1. Rater Pre-Screen. In the first stage of screening, we identified workers who had passed a lab-internal screener and submitted thoughtful responses and high reading comprehension skills. In the second stage, we asked potential raters to undergo a training program, developed using "gold standard" ratings performed by expert researchers. Specifically, four researchers independently completed the full 23-dimension rating for four tasks, and their responses were averaged to determine the "correct" answers. During training, workers could not advance the survey without selecting the right answer and reading an explanation for why it was correct. In the third screening stage, workers completed a 16-question pre-test.

We asked 155 MTurk workers to complete the training module and pretest. 61 of these workers also completed a full 23-question "validation task," which we used to ensure that the pretest was effective, as well as to select a cutoff score for becoming a qualified rater. Figure 2 illustrates the process of selecting a score cut-off for the test (75%). 121 workers passed the cutoff and were included in our rater pool (among these, 112 ultimately submitted at least one rating). The average score for workers who passed the pre-test was 84.56%.

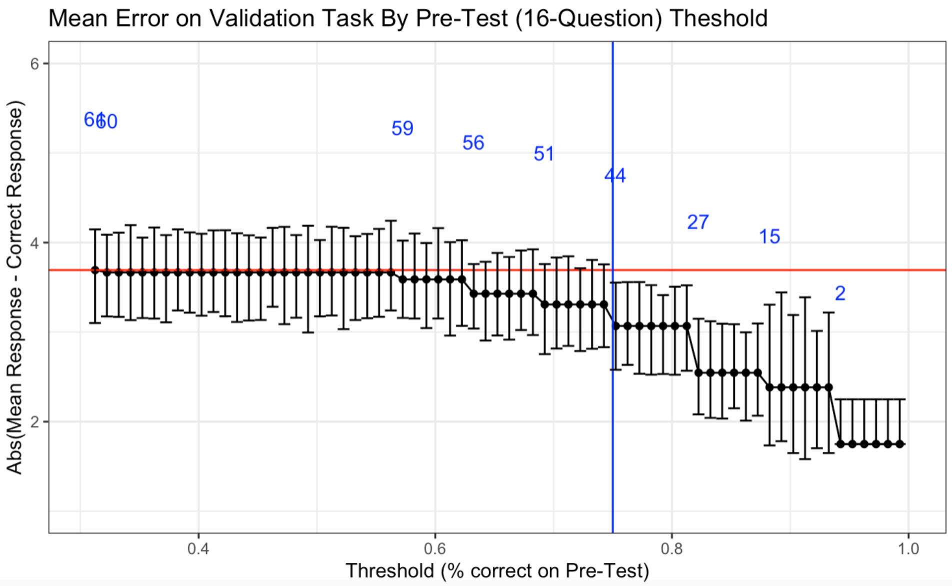
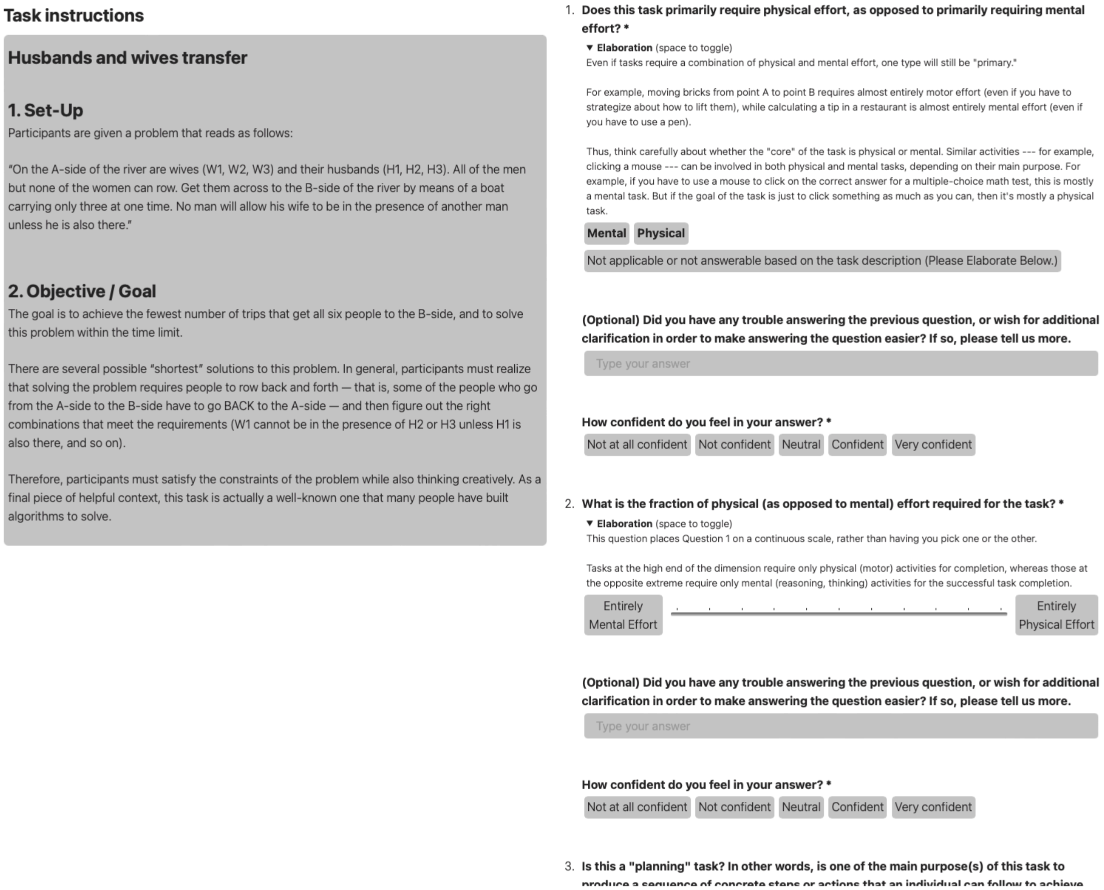


Figure 2. A plot of 61 raters who completed the training module, pre-test, and the validation task. The red line shows the mean score across the 61 participants. The blue line shows the cut-off threshold (75% correct). The error bars are bootstrapped 95% confidence intervals. As shown in the plot, as participants score higher on the pre-test, their deviation from the "gold standard" response decreases. Beyond the cut-off threshold, using the mean response of only the highest-scoring participants performs significantly better than using the mean response of all participants. Since we aggregate ratings using the mean, our task-rating process is itself a task in which the average individual determines the outcome (one of Steiner's dimensions!). Thus, increasing the competence of our average rater is critical.

C.4.2. Task Rating Process. After loading completed task summaries into a database, we notified our pool of raters, who were able to rate each task on a first-come, first-served basis (rating interface shown in Figure 3). Tasks that received at least 20 ratings by unique individuals were removed; due to a lag, in which tasks were removed slower than raters could claim them, each task had an average of 23.20 ratings.

During the rating process, workers were also asked to provide a level of confidence for each score (using a 1-5 scale anchored at "Very confident" and "Not at all confident"), and were also given the opportunity to refuse to answer questions that they felt were inapplicable to the task. These 'NA' ratings constitute a very small fraction (474/47,800 ratings, or 0.991%) of the total.



Task instructions

Husbands and wives transfer

1. Set-Up

Participants are given a problem that reads as follows:

"On the A-side of the river are wives (W1, W2, W3) and their husbands (H1, H2, H3). All of the men but none of the women can row. Get them across to the B-side of the river by means of a boat carrying only three at one time. No man will allow his wife to be in the presence of another man unless he is also there."

2. Objective / Goal

The goal is to achieve the fewest number of trips that get all six people to the B-side, and to solve this problem within the time limit.

There are several possible "shortest" solutions to this problem. In general, participants must realize that solving the problem requires people to row back and forth — that is, some of the people who go from the A-side to the B-side have to go BACK to the A-side — and then figure out the right combinations that meet the requirements (W1 cannot be in the presence of H2 or H3 unless H1 is also there, and so on).

Therefore, participants must satisfy the constraints of the problem while also thinking creatively. As a final piece of helpful context, this task is actually a well-known one that many people have built algorithms to solve.

1. **Does this task primarily require physical effort, as opposed to primarily requiring mental effort? ***

▼ **Elaboration** (space to toggle)
Even if tasks require a combination of physical and mental effort, one type will still be "primary."

For example, moving bricks from point A to point B requires almost entirely motor effort (even if you have to strategize about how to lift them), while calculating a tip in a restaurant is almost entirely mental effort (even if you have to use a pen).

Thus, think carefully about whether the "core" of the task is physical or mental. Similar activities --- for example, clicking a mouse --- can be involved in both physical and mental tasks, depending on their main purpose. For example, if you have to use a mouse to click on the correct answer for a multiple-choice math test, this is mostly a mental task. But if the goal of the task is just to click something as much as you can, then it's mostly a physical task.

Mental **Physical**

Not applicable or not answerable based on the task description (Please Elaborate Below.)

(Optional) Did you have any trouble answering the previous question, or wish for additional clarification in order to make answering the question easier? If so, please tell us more.

Type your answer

How confident do you feel in your answer? *

Not at all confident | Not confident | Neutral | Confident | Very confident

2. **What is the fraction of physical (as opposed to mental) effort required for the task? ***

▼ **Elaboration** (space to toggle)
This question places Question 1 on a continuous scale, rather than having you pick one or the other.

Tasks at the high end of the dimension require only physical (motor) activities for completion, whereas those at the opposite extreme require only mental (reasoning, thinking) activities for the successful task completion.

Entirely Mental Effort — Entirely Physical Effort

(Optional) Did you have any trouble answering the previous question, or wish for additional clarification in order to make answering the question easier? If so, please tell us more.

Type your answer

How confident do you feel in your answer? *

Not at all confident | Not confident | Neutral | Confident | Very confident

3. **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**

Figure 3. The interface for rating tasks. The Stimulus and Goals of the task description (which are labeled as "Set-Up" and "Objective / Goal" in order to be more accessible to raters) are presented in a panel on the left-hand side. The questions and elaborations appear on the right-hand side.

C.4.3. Question Elaborations. The elaborations we provided to raters are presented in Table 3.

Table 3: Dimension Elaborations Provided to Raters

Dimension Name	***Elaboration***
Conceptual-Behavioral	Even if tasks require a combination of physical and mental effort, one type will still be "primary." For example, moving bricks from point A to point B requires almost entirely motor effort (even if you have to strategize about how to lift them), while calculating a tip in a restaurant is almost entirely mental effort (even if you have to use a pen). Thus, think carefully about whether the "core" of the task is physical or mental. Similar activities — for example, clicking a mouse — can be involved in both physical and mental tasks, depending on their main purpose. For example, if you have to use a mouse to click on the correct answer for a multiple-choice math test, this is mostly a mental task. But if the goal of the task is just to click something as much as you can, then it's mostly a physical task.
Intellectual-Manipulative	This question places Question 1 on a continuous scale, rather than having you pick one or the other. Tasks at the high end of the dimension require only physical (motor) activities for completion, whereas those at the opposite extreme require only mental (reasoning, thinking) activities for the successful task completion.
Type 1 (Planning)	These are tasks "having emphasis on action orientation." Answer 'yes' if this task involves writing down a series of concrete steps that someone will follow in order to achieve a goal. Examples include organizing an event, or figuring out a plan for buying things at the store. This plan may later be executed by the person making it, or could be given to someone else. Answer 'no' if this task just involves doing an activity without writing the steps down. (For example, if you write a recipe and then cook, you are planning; if you cook without a recipe, that is NOT planning.) Also answer 'no' if there is no intention to actually use the plan in the "real world," or if it's just a hypothetical math problem. Sometimes math problems involve a story or context (e.g., "calculate how much time it takes for Jimmy to go to the store," or "find the shortest path that a traveling salesman can take". However, these are math problems rather than plans.
Type 2 (Generate)	These are tasks based around generating ideas, examples, or concepts. The examples being generated can be either abstract (generating words and colors) or concrete (generating ideas for how to spend $10,000 or ways to use a paperclip). If the ideas are concrete, they should NOT suggest a specific plan or course of action. This question is therefore different from the previous question; the participants are NOT generating a plan. Note that you should only answer “yes” to this question if one of the main outcomes of this task is to generate such ideas. Many tasks require people to discuss ideas (for example, a jury task), or come up with creative ways to solve a problem (for example, a math problem), but the primary goal of such tasks is not just to generate the ideas.
Creativity Input	You can think of this, in a way, as the continuous version of the previous question, but it is also intended to capture a wider range of "creative" activity than just purely generating ideas. Descriptors that might suggest creativity include: "new", "novel", "unique", "come up with as many ideas as possible", "invent", "create," etc. A purely creative task is solely based around such activities (for example, writing advertisements). Descriptors that do not suggest creativity include: "add these numbers together", "click your mouse as fast as possible", "find the optimal allocation." Such tasks might involve simply following the rules or executing a task without thinking creatively. Finally, some tasks are in between — it may be possible to invent a unique way to solve a math problem. This is more creative than a problem that simply requires adding numbers.

Table 3: Dimension Elaborations Provided to Raters

Dimension Name	***Elaboration***
Type 5 (Cognitive Conflict)	Within a group, a difference in viewpoint or opinion may arise either because members interpret the same information in different ways, assign different importance to that information, or both. Resolving those differences may take many forms, such as having a discussion, building consensus, holding a vote on the options, or simply thinking through ways to account for the viewpoints (e.g., making a judgment call that balances the perspectives). For individuals, resolving opinions may involve adjudicating real or hypothetical disagreements for others. For example, a game in which you pretend to be a judge and decide court cases would primarily involve resolving differences in perspective or opinion. Another example of a 'yes' is an activity where you run a funding organization and different projects have applied for access to your funds. A decision-making task for allocating these funds will involve making sense of the different opinions and perspectives that parties may have.
Type 7 (Battle)	This question captures whether a task is a "competition" or "battle," in which your outcome is relative to someone or something else (e.g., another player, team, or an AI agent). A battle is "where the focus is on conquest of an opponent and winner-take-all distribution of payoffs," and a competition is "where there is a lot of emphasis on standards of performance excellence over and above the reckoning of winners and losers." Competitions between sports teams fit this category (e.g., basketball, soccer, ice hockey), but so too do military and street gang battles (e.g., the spoils of war go to the victor no matter how honorably the vanquished may have fought). Answer 'yes' ONLY if you "win" or "lose" relative to somebody else who is playing the game. If someone performs poorly in the task, but they are not being compared to another player, you should answer 'no.' For example, if you are being asked to come up with as many ideas as possible, you may think that you "lose" if you generate zero ideas or that you "win" because you generated many ideas. However, you should still answer 'no,' because this question is asking about "winning" and "losing" in relative terms. A task in which you need to come up with more ideas than another player, on the other hand, would be a 'yes,' since you would win or lose relative to that other person.
Type 8 (Performance)	These tasks are about doing just enough work to meet some threshold or standard, but once you reach that standard, precision does not matter. As an example, a pass-fail class (e.g., "anything above 70% is a pass") is all-or-nothing. On the other hand, a letter grade (A+ for 100%, A for 90, and so on) is not. Another example is a task in which people simply need to move objects from Point A to Point B: if they meet the outcome, they succeed. Even if they moved the object part of the way there, they don't get any credit for their effort. Moving it an extra 10 feet does not get them any brownie points. Counter-examples include games in which you try to earn as many points as possible, and you end up with a numeric score instead of simply a success/fail. For example, trying to generate as many ideas as possible, or solve as many problems as possible, would NOT be all-or-nothing tasks. If participants moving an object get credit for every foot that they manage to move it, then it's NOT all-or-nothing.
Divisible-Unitary	Ask yourself: is it efficient or useful for a team of people to "divide and conquer" this task, or is this really something where one person should be doing most of the work? A key heuristic is whether the different sub-parts of the task are interdependent or not. If the sub-parts are not dependent on each other, it often makes sense for different people to work on each part separately (so answer 'yes'). However, if the sub-parts are interdependent, one person can't start their part without waiting for another person to finish, so dividing and conquering doesn't make sense (and you should answer 'no'). For example, if you are solving 10 simple arithmetic problems, each of ten people may work on one of them. If there is only one problem to be solved, it's not efficient for one person to do the thinking and another to do the writing. Similarly, for a task where you want to generate as many ideas as possible, everyone can separately come up with ideas and combine them in the end. For a task where everyone needs to come up with a shared plan, the parts of the plan depend on each other, so it is difficult to divide and conquer.

Table 3: Dimension Elaborations Provided to Raters	
Dimension Name	***Elaboration***
Maximizing	Sometimes the goal that is to be achieved entails doing as much as possible of something, or doing it as rapidly as possible. Thus, if an individual or group is asked to exert a maximum force on a rope, a strong pull is regarded as a more successful performance than a weak pull. If a team of mountain climbers is asked to ascend a cliff as rapidly as possible, maximum speed is the criterion against which performance is evaluated. Look for tasks asking participants to score the most number of points, get the most utility, for everyone to generate as many ideas as they can, and so on.
Optimizing	Here, there is a specific standard to meet, but precision matters. (This is unlike the previous question about all-or-nothing outcomes, for example, where precision did not matter.) Look for tasks with a specific, most preferred outcome. If the task is to estimate the temperature of a room, the goal is to agree with the value indicated by the thermometer. Another example is moving exactly 10 objects — no more, no less — or to exactly reproduce something (such as a task where people have to exactly copy a work of art). Also look out for concepts like "exactly," "precisely," "optimal," or "best." For example, if you are asked to generate exactly 10 ideas, generating both 9 ideas and 11 ideas would be considered a failure. Another clue might be terms like "constraints," "requirements," or "rules" that someone has to follow, as long as there is a specific best answer within the rule set. For example, if you want to buy the most number of items using a budget of $100, the "constraint" of $100 also serves as a precise goal. That is, the closer you get to $100 (the more precise you are), the better you do at the task.
Outcome Multiplicity	For example, an arithmetic problem, such as summing a bunch of numbers, will have only one correct answer. On the other hand, a creative writing task has many valid solutions. In cases where there are many possibly correct answers, or answers with partial solutions, there could still be one "best" solution, or a specific "best" solution that the experimenter is looking for. For example, in a game in which one needs to win in as few moves as possible, there may be an optimal lowest number, even if there are less efficient solutions.
Solution Scheme Multiplicity	Is there only one right way of solving the problem? Answer 'no' if there is more than one possible course of action or process to attain the group's goal. Answer 'yes' if there is only ONE process or action that will lead to the correct answer or achieve the goal. As an example, answer 'yes' if the instructions specify exact steps, e.g., "first you write all the 3-letter words, then the 4-letter words, then the 5-letter words," or if there's really only one way to do the task (e.g., perform long division). Actions might also be limited by the environment: participants may be working on an electronic system that restricts their communication or the steps that they are allowed to take. However, answer 'no' if the task is open-ended in terms of course of action: e.g., if participants are simply asked to "come up with a solution" or "give as many ideas as possible" without specifying how they should achieve the goal.
Decision Verifiability	This item refers to the "degree to which acceptable solutions can be demonstrated to be correct," via logic and rules as opposed to having a general consensus. For example, the solution to a math problem can be verified via the rules of algebra. In a task where participants are asked to buy the items with the best value, one can also use logic to list out every possible combination of items and show that their solution is the best one. Another type of demonstration could involve showing facts: for example, if the task is to estimate how many people live in the United States, the "ground truth" is the statistic from the U.S. Census. Examples of tasks where you should answer 'no' include answering the question, "should we ban all guns?" — this cannot be verified via logic and rules. Evaluating whether something is "creative," as well as other subjective judgements, is also not demonstrable.

Table 3: Dimension Elaborations Provided to Raters

Dimension Name	***Elaboration***
Shared Knowledge	Some tasks can be written as a formal model or math problem, expressed with rules and syntax that the problem-solvers share. One way to think about this question is to ask yourself, could a robot or algorithm do this task? Here are examples of a 'yes' answer: - A specific set of rules and outcomes: Robots are really good at following rules. Answer 'yes' if you can input the rules and the desired outcome, and have a robot follow predetermined steps to get the result. For example, you can tell the robot, "buy as many items as you can with a budget of $100." (In fact, online shopping websites have exactly this tool!) - You can write it like a math problem: Robots are really good at doing math. Something like, "find the best teams under the constraints," combined with a list of how much "utility" people get from being put together, can easily be put into a computer and solved. Similarly, if you're trying to find the shortest path from Point A to Point B, you can think of this in terms of geometry. - You can break the problem down into small units of meaning: Examples of these units include the alphabet (26 letters), colors (which can be broken down into units of Red, Blue, and Green), or coordinates on a graph (X,Y). An algorithm could use these units of meaning to solve the task. For example, when generating words, the letters are the units. You can imagine an algorithm in which you use "brute force" to find every possible combination of letters that makes a valid word. Another example is Wordle — many people have built bots that consistently solve the puzzle! Here are examples of a 'no' answer: - Creative ideas that cannot be generated just by following basic rules. Whereas you can generate words just by using the rule of trying out all the different letters in the alphabet, you can't generate stories in the same way. There are no "rules" for what makes a good story. (We ask you not to think about advanced models like GPT-3 for the purposes of this question.) - Subjective tasks that require a judgment call. Deciding whether or not we should ban all guns can't really be done by an algorithm: people have to debate these ideas. - Tasks relying on social dynamics. Anything focusing on relationships between participants should be labeled 'no.'
Within-System Solution	This item refers to whether there is sufficient information to obtain a solution within the system. In other words, this question is about whether it is possible — using all of the information provided to the participants — to get a valid or acceptable answer. By "the system," we mean the entire set of stimuli given to the participant(s) for solving the problem. If the person is rating images, it is the set of images and the rating scale/survey; if it is a game, then it is the entire self-contained "system" of the game and the rules for playing it. If it is an optimization problem, it's the set of object(s) that need to be optimized and the list of constraints. Examples of 'no' answers: - The only way to get a valid solution is outside of the "system" — e.g., looking the answer up, asking someone else. For example, a trivia quiz (where you have to look up the right answer) would be a 'no' for this question. - There is no right answer, or the question is fundamentally unsolvable (e.g., it's a trick question). - There's not enough information to solve the question. In math, for example, you need two equations to solve for two unknown variables. If you don't have enough information, you can't do the problem.

Table 3: Dimension Elaborations Provided to Raters

Dimension Name	***Elaboration***
	Examples of 'yes' answers: - Any question where you can get a right answer from the information given is a 'yes.' This is true even if you're not sure of your answer. For example, in a multiple-choice quiz, you may not be sure that your answer is correct. But if one of the choices was the right answer, and the problem gave enough information for you to select it, then answer 'yes.' - If there are multiple valid answers in response to the problem (e.g., a creative writing task where any valid story is accepted), the answer should be 'yes.'
Answer Recognizability	This question is about whether participants who are not themselves able to solve the problem have sufficient knowledge of the system to recognize and accept a correct solution if it is proposed by someone else. In other words, are you informed enough to know the right answer when you see it? If there is no well-defined "correct" answer to begin with, the answer to this question is always "no."
Time Solvability	If someone is able to solve the problem (e.g., find the best solution), will that person "have sufficient ability, motivation, and time to demonstrate the correct solution to the incorrect members?" In other words: Assuming infinite time and resources, can someone both (1) solve the problem AND (2) show that the solution is right? You should answer “yes” to this question only if a participant can do BOTH steps. Otherwise, answer "no." If there is no well-defined "correct" or "best" answer to begin with, the answer to this question is always "no." Finally, remember that this question is about whether it is possible to come up with ANY "best" solution and prove that it is right; if the proof is so complicated that others would not recognize it, your answer to the previous question (about whether someone would recognize the correct answer if told it) should be "no," but you should answer “yes” to this question, because a proof of a correct solution still exists.
Type 3 and Type 4 (Objective Correctness)	If you answer 'yes' to this question, the task should have "a demonstrable right answer, and the group task is to invent/select/compute that correct answer." This right answer can be found in multiple ways: - By using intuition - By relying on general norms or accepted ideas - By using logic or applying rules - By consulting an expert (for example, the answer to, "is climate change real?" can be determined by an expert.) If you answer 'no' to this question, "there is not a demonstrably correct answer, and ... the group's task is to select, by some consensus, a preferred alternative." The alternatives might involve: - Figuring out consensus on cultural values or moral courses of action - Figuring out consensus by sharing relevant information

Table 3: Dimension Elaborations Provided to Raters

Dimension Name	***Elaboration***
Intellective-Judgmental	Imagine a continuous scale from judgemental (subjective) to intellective (objective/logical) tasks. You can think of this as the continuous version of the previous question. Intellective tasks are problems or decisions for which there exists a demonstrably correct answer within a verbal or mathematical conceptual system. The criterion of successful group performance is the achievement of this correct answer. Judgmental tasks are evaluative, behavioral, or aesthetic judgments for which there does not exist a demonstrably correct answer. The criterion of successful group performance is the achievement of consensus on a collective decision. If the solution is demonstrable by numbers or statistics, but must be done relative to other participants (for example, you want to know your score relative to how other players did), your answer should be in the middle (e.g., 0.5).
Conflicting Tradeoffs	Some tasks provide a variety of inputs to help complete the task, but these inputs don't always agree or point to the same solution, and participants must evaluate or navigate tradeoffs in executing the task. For example, a task involving writing an ad for a product may provide individuals with a variety of product features, customer reviews, and market information that may not all point to the same answer or angle for the ad. Similarly, a problem asking you to devise a business plan may require you to manage conflicting pieces of information, such as inconsistent needs for different consumer segments and tradeoffs between different courses of action.
Solution Scheme Outcome Uncertainty	This question is asking whether a player knows for sure that their strategy or answer is right BEFORE the activity ends, or whether they have to wait until AFTER to find out. Answer 'yes' if participants submitting their answers may feel unsure that their submission or approach is correct, with no way to "check" their answer until the task is done. Here are tasks where you would want to answer 'yes:' - Tasks where you are playing against an opponent in a dynamic game that is not determined until the end. For a soccer game, you can come in with a "game plan" and a strategy and play your best, but that doesn't guarantee that you win. You have to play the whole game to find out who wins. Even if you're 6 points ahead, there's still a chance that the other team makes a comeback, so you do not know for sure that you won until the end. (The same goes for other types of games, like video games.) - Tasks where there's a mathematical limit or objectively correct answer, which does not get revealed until the end. An example is guessing the number of jellybeans in a jar. You can try your best to guess, and you can even be fairly confident that you're right, but until the final answer is revealed, there is some uncertainty about whether your answer is "correct." Similarly, a task where you try to buy the most number of items with $100 has a mathematical maximum; however, you may not be sure that you found that maximum until the end. On the other hand, examples of 'no' are: - Tasks where you can check your work easily. If you solve an algebra problem, you can solve it using a well-defined series of rules and steps. And you will know your answer is right, because you can substitute your solution for X back into the equation to confirm that it works. - Tasks where you get information, BEFORE the game is over, that you already won. For example, if you need to earn at least 200 points and you can see your score update until it reaches 200, answer 'no.' Another example is if you need to move a small object from Point A to Point B — you can just carry it out and confirm (by looking at its new location) that it worked. - Tasks where the goal is to just do your best. If the task is to generate as many ideas as possible, there is no mathematical limit to how many you can come up with. As long as you do your best, you can be certain that

Table 3: Dimension Elaborations Provided to Raters

Dimension Name	***Elaboration***
	you have an acceptable answer. (On the other hand, if the goal is to generate more ideas than another player, then you would be uncertain.)
Eureka Question	Some questions have a special trick in them, such that, if you know the trick, the question is easy, but if you do not know the trick, the question may be quite difficult. An example of this is the Sphinx's Riddle (e.g., "what has 4 legs in the morning, 2 in the afternoon, and 3 in the evening?"). This puzzle is confusing if you have never heard it, but if you have heard the answer (a human -- since babies crawl, adults walk, and the elderly use canes), you'll never be confused again. In other words, does this task cause a "Eureka!" or an "Oh, I get it!" moment when the answer is explained? If there is no trick, answer 'no' to this question.

Table 3. The elaborations for each dimension.

C.5. Discussion of Ambiguity Remaining in Task Space

Despite our best efforts to clarify each task dimension, however, some ambiguity may remain. Figure 4 presents a heatmap in which each row represents a task and each column represents a dimension. The heatmap is ordered such that dimensions with lower average values (that is, most people gave a rating of 0) are at the far left, and dimensions with higher average values (that is, most people gave a rating of 1) are at the far right.

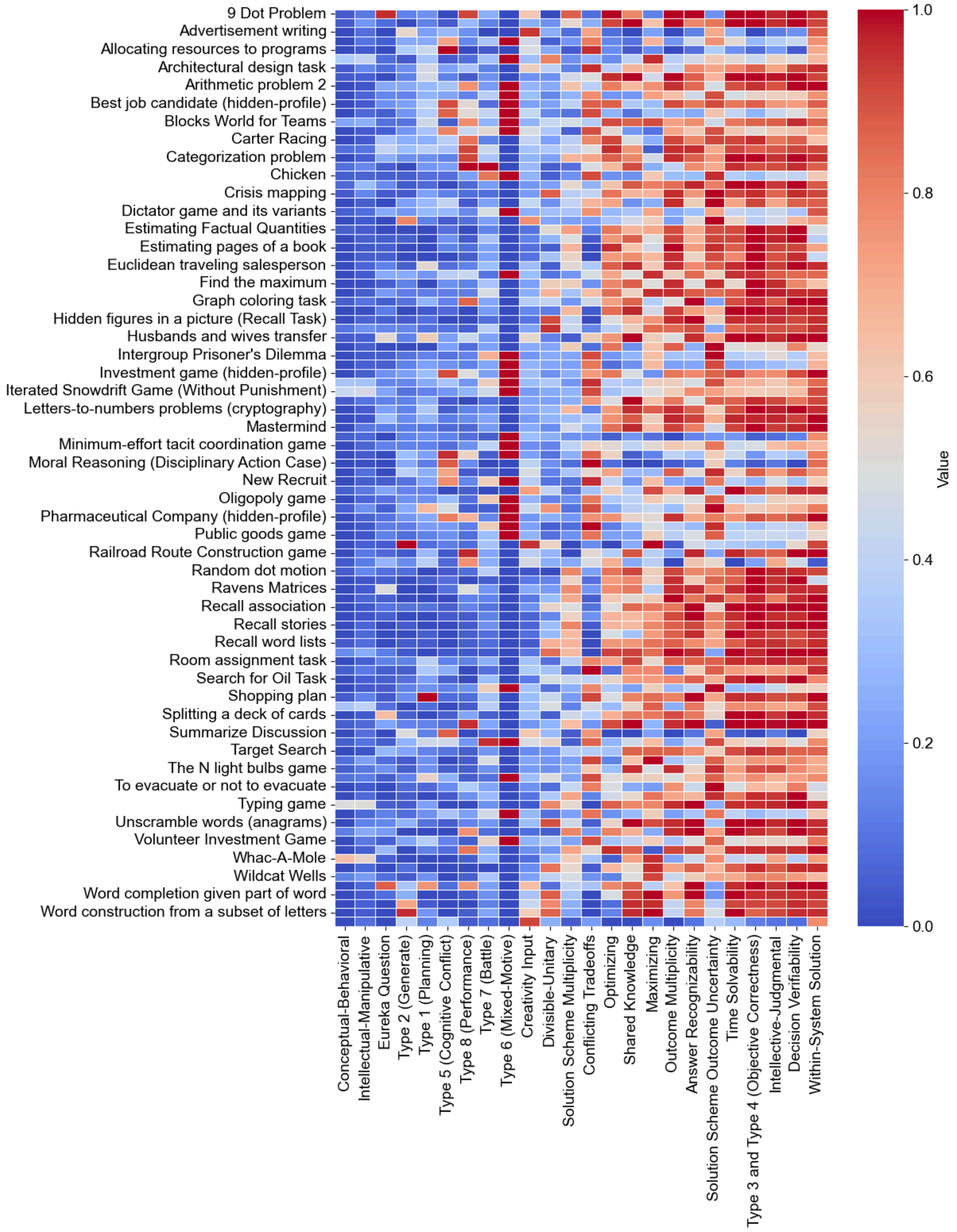


Figure 4. A heatmap in which each row represents a task, and each column represents a task dimension. The color represents the aggregated rater response; darker colors — red and blue — represent higher rater agreement (red indicates that the mean rating is close to 1, while blue indicates that the mean rating is close to 0) while lighter colors represent lower rater agreement (the mean rating is close to 0.5).

Since the ratings were performed on a 0-1 scale (with most questions being binary), values close to 0 and 1 represent high agreement, and values close to 0.5 represent low rater agreement. We interpret low agreement as "the task only partially displays the feature" — in other words, because the task does not clearly display the attribute, raters were uncertain about whether the dimension is relevant.

However, this interpretation also embeds raters' subjective understanding of the task dimensions, and it is possible that some *constructs* involve more subjective interpretation than others. For example, examining the columns of Figure 4 from left to right, the leftmost and rightmost columns represent those for which raters had the highest agreement. The questions at these extremes include Conceptual-Behavioral ("Does this task primarily require physical effort, as opposed to primarily requiring mental effort?") and a collection of questions about whether there was a right or wrong answer (e.g., Within-System Solution, Decision Verifiability).

Questions with the most disagreement — represented by the columns in the middle of Figure 4 — include Solution Scheme Multiplicity and Conflicting Tradeoffs. These questions involve constructs that are less straightforward; for example, what constitutes a "conflicting tradeoff" or a "divisible" task are more open to interpretation than whether or not the task involves achieving a correct answer.

To some extent, the frame problem will always be present, try as one might to be precise with one's words. (As Bertrand Russell once said: "everything is vague to a degree you do not realize till you have tried to make it precise.") We acknowledge this as a limitation of our current 24-dimensional Task Map, in which the annotations rely on subjective judgments. However, we also see this limitation as an opportunity for future work — for introducing new ways of quantifying task dimensions to perhaps resolve more of the ambiguity. The Task Space is designed to incorporate these improvements in measurement over time, and to serve as a living artifact of what we can know about tasks.

Indeed, we view this transparency as a key strength of the Task Space: when deciding whether to assign a given task to "Type 4" or "Type 5," researchers still make subjective judgments, but their decisions are not recorded; for example, a task may be right on the border between the two types, but it will only be encoded as one type or the other. In the Task Space, the subjectivity in our dimensions is fully transparent — evident through the numeric values in the dimensions — and it can help formalize and quantify what we don't yet know in the process of continuous learning.

Appendix D: Task Writing Checklist

To ensure that tasks were presented as consistently as possible, we used the following checklist to guide the writing of task descriptions. Many of these checklist points specifically mention dimensions that raters will evaluate (e.g., “Maximizing,” “Optimizing,” etc.); by clarifying goal directives in the task description, we hoped to leave as little room as possible to the raters’ imagination.

- □ **Articulate goal directive clearly.**
  - □ **Does the goal directive highlight the true purpose of the task?** E.g., are we trying to **maximize** (points or actions), to **optimize** (get to a specific best number), or just **do your best / state your opinion**? Is it a combination of these things — for example, you have to the the maximum points, but there is an “optimum” because there is a limit?
  - □ **Make clear if we are looking for a maximum/optimum/both/neither.** If it’s written in a vague or ambiguous way, pick a stance.
    - □ *Maximum:* Use words like “maximize” and “as much as possible” if there’s a maximum
    - □ *Optimum:* Use words like “precisely” “exactly” if there’s an optimum
    - □ *Both:* If there is both a maximum and optimum, try to make this explicit, e.g., “participants have to get as many points as possible, but there is a maximum/best possible score/limit …”
    - □ *Neither:* State explicitly “there is no right or wrong answer,” or that participants simply “give their opinion,” “produce the best possible writeup,” etc. to signal that there is no maximum or optimum.
  - □ **Is there an exact threshold that participants have to clear?** (e.g., come up with *at least* 10 ideas …)
- □ **Mention that the time is fixed.**
- □ **Mention, if unclear, that there is sufficient information to complete the question.** For example, participants must solve the puzzle in at least N moves, and the puzzle is solvable in N-1 moves.
  - □ Make clear that the question is or is not a “trick” question.
  - □ Make clear if the question causes an “aha” moment, and **make sure to add additional context about the right answer that will be useful for the map**.

    Example from the 9 Dot Problem:
    “An important note is that there is really only one "correct" answer to this question — that is, the way to connect the nine dots actually requires the participant to go "outside" the boundaries of the 3x3 grid, or else it cannot be done. The true goal of this task is to find this optimal answer.

    Participants who do not think "outside the box" therefore usually have difficulty completing this task. Once they are shown that the answer involves going outside of the box, they usually have an "aha" moment and do well on the task in the future.”

- □ **Give a concrete example**. If there’s a card, what’s on it? If participants get instructions or a stimulus, what does it say?
- □ **How is the task graded?**
  - □ Do you get to see the number of points you earn along the way? How much progress do you get to see?
  - □ Do you get partial credit?
  - □ Do you get told if you’re right?
  - □ Does the system reject wrong answers?
  - □ How many tries do you get?
- □ **Mention what resources are shared.** (e.g., “Participants have access to the same list …” or “Each participant has their own list …”)
- □ **Are any parts of the task dependent on other parts?** Make sure you are explicit if this is the case.

Appendix E: Intuitive Interpretation of the Task Space

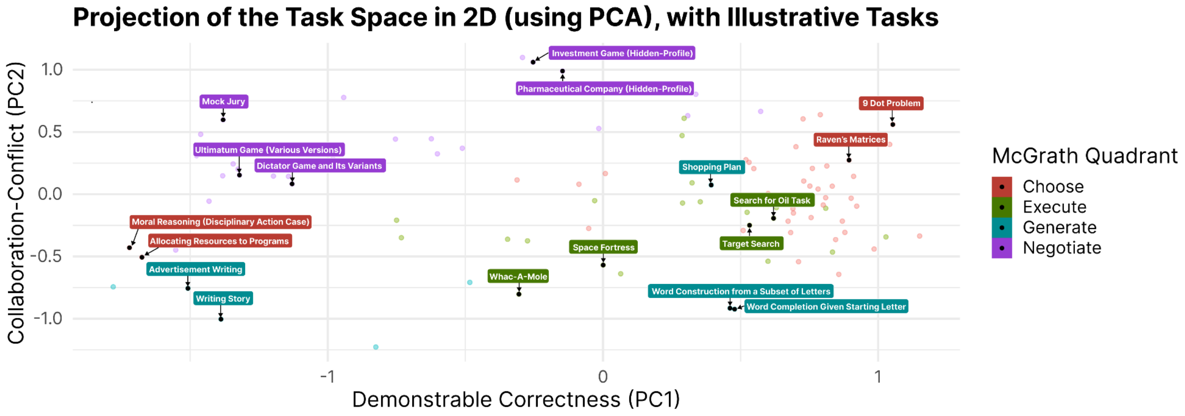


Figure 5. A two-dimensional projection of the Task Space using PCA. The first two components capture 60% of the variance in the data, with 47% in the first component and 13% in the second component. The axis labels provide a rough interpretation of each component by examining the weights assigned to each dimension. Such a projection is conditional on the specific sample of tasks and dimensions in the current Task Map; thus, we caution against over-interpreting such a diagram, as the principal components will change as new tasks and dimensions are added. However, the map-like projection is useful for achieving an intuitive understanding of the relationships between tasks.

In Figure 5, we visualize the Task Space in terms of its first two principal components, which together capture 60% of the current data's variance (47% in the first component and 13% in the second). These components illuminate some of the underlying dimensions shared across the Task Space's five constituent frameworks, though with an important caveat: the principal components are sensitive to the specific tasks and dimensions in the Task Map, and may change when different tasks and dimensions are under consideration. Nevertheless, we use this analysis to demonstrate that our multidimensional approach generates useful insights. It can recover known concepts and relationships, as well as identify cross-cutting themes among taxonomies.

Preserving Intuitive Concepts and Relationships Between Tasks. In Figure 5, we observe, encouragingly, that intuitively similar tasks are clustered together: points representing "Advertisement Writing" and "Writing Story" appear together in the bottom left; those representing "Word Construction from a Subset of Letters" and "Word Completion Given Starting Letter" (a pair of closely related word puzzles) are adjacent in the bottom right; points representing "Search for Oil Task" and "Target Search" (which both involve identifying a target within a defined region) are also adjacent. The points representing two hidden-profile tasks — "Investment Game" and "Pharmaceutical Company" — appear at the top center. Finally, points representing economic games ("Ultimatum Game" and "Dictator Game") jointly appear at the top left. These relationships capture our intuitive understanding that similar tasks should be proximate in the Task Space, and qualitatively demonstrate that our rating and aggregation method reliably assesses task dimensions. Despite being rated independently, similar tasks yield similar results.

Identifying Cross-Cutting Themes. Having passed this "sanity check," we next directly compare the Task Space to the well-known McGrath Task Circumplex (1984). Importantly, all of the dimensions in the framework belong to the task qua task approach, which makes it especially apt for comparison. We assign each task in our repository to one of the four "quadrants" from McGrath's circumplex (Generate, Choose, Negotiate, and Execute),[1] and we observe that, in addition to preserving the relationships theorized by

[1] The first author assigned each task to a quadrant by hand using the definitions from McGrath (1984). Where possible, we relied on existing precedent and McGrath's own examples to categorize the tasks; however, given

McGrath, our inclusion of additional task dimensions reveals novel insights that are not captured by the standalone framework.

For example, McGrath categorizes Intellective Tasks (Type 3) and Decision-Making Tasks (Type 4) as members of the Choose Quadrant, and thus one would expect these groups of tasks to be proximate. In our Task Map, however, Type 3 and Type 4 tasks appear at opposite poles — with Intellective Tasks ("9 Dot Problem" and "Raven's Matrices") at the far right, and Decision-Making Tasks ("Moral Reasoning" and "Allocating Resources to Programs") at the far left. The larger-than-expected distance between Type 3 and Type 4 reflects the fact that the Task Space incorporates seven dimensions from Laughlin and Ellis (1986), which measure a task's *demonstrability* (the extent to which there is a provably correct answer). Demonstrability is primarily loaded onto the first principal component (the *x*-axis), yielding a split across which more demonstrable tasks are on the right and less demonstrable ones are on the left.[2]

Similarly, demonstrability splits Creativity Tasks (Type 2) into two subcategories: *open-ended creative tasks* (such as "Advertisement Writing" and "Writing Story," in which participants freely generate responses with minimal restrictions) and *close-ended creative tasks* (such as word completion tasks, in which content must be constrained to valid English words containing specific letters). The former is less demonstrable — there are no rules for a "provably" good story — while the latter is fully demonstrable, since one can simply check the generated words against a dictionary. Here, then, incorporating task dimensions from another framework provides an insight that the original McGrath framework missed: not all creativity tasks are equal, and demonstrability is a key dimension along which they differ.

We also observe that some tasks are more similar than McGrath originally credited them for. For example, McGrath's framework positions Creativity Tasks and Competitive Tasks relatively far apart, yet Competitive Tasks in our repository ("Whac-A-Mole," "Space Fortress") are located near the close-ended Creativity Tasks ("Word Construction from a Subset of Letters" and "Word Completion Given Starting Letter"). Upon further examination, an important shared attribute between these tasks is that they are *maximizing* — a dimension described by Steiner (1972) as asking participants to perform some action as much as possible. It turns out that this attribute is shared between some Competitive Tasks and some Generate Tasks. The former involves earning as many points as possible, and the latter involves coming up with as many words or ideas as possible. Incorporating Steiner's feature thus yields a commonality that McGrath had not accounted for.

Importantly, the Task Space still reflects concepts and relationships central to the original framework. In line with McGrath's predictions, "Shopping Plan," a Planning (Type 1) task, is remarkably close to the Performance (Type 8) tasks, "Target Search" and "Search for Oil Task." Tasks in the Negotiation Quadrant are clustered in the top left corner, and they include those in which participants have different payoffs (such as in economic games), different viewpoints (such as in jury tasks), or different information (such as in hidden-profile tasks). As McGrath wrote, "these subtypes shade from one border of the category to the other" — and our quantitative representation of these tasks illustrates the shading. Indeed, the purple tasks in the Negotiation Quadrant meet the Type 4 tasks ("Moral Reasoning" and "Allocating Resources to Programs") at the border, exactly as McGrath had indicated in his original circumplex.

ambiguity in definitions, classifying tasks into discrete categories is inevitably a subjective exercise in judgment. Given this ambiguity, we limited the analysis to the quadrant level, as many tasks cannot be clearly assigned to a specific type.

[2] Considering that seven dimensions from Laughlin and Ellis are related to demonstrability, in addition to one dimension from McGrath (Type 3 - Type 4) and one from Steiner (Optimizing), 9 of our 24 task dimensions describe demonstrability to some extent. Thus, it is unsurprising that this underlying feature composes the first principal component, and the fact that PC.1 captures 47% of the variance is partially a function of its over-representation in the Task Map's current dimensions. For this reason, we take care not to argue that demonstrability is inherently the most important task feature — rather, it is the most important feature *in our data*, conditional on this particular current set of tasks and dimensions.

Indeed, the second principal component (the y-axis) encodes a version of McGrath's Collaboration-Conflict axis. Dimensions assigned large positive weights in PC.2 include those in which individuals are in conflict with one another (Mixed-Motive, Cognitive Conflict, Conflicting Tradeoffs) or in which only one individual needs to submit a correct answer (Optimizing). Dimensions assigned large negative weights include those describing tasks that can be divided and conquered among teammates (Divisible-Unitary), and those that benefit from doing something as much as possible (Maximizing). Taken together, one can interpret PC.2 as encoding the notion of task interdependence (Straus 1999; Marlow et al. 2018), describing the extent to which the task requires collaboration and coordination among team members.

Because the Task Space's dimensions encompass those of any constituent taxonomy, it contains strictly more information than any standalone framework. As our case study of McGrath (1984) illustrates, this additional information can be used to identify novel insights about tasks, while retaining classic ideas from the original taxonomy. This logic extrapolates to comparisons between the Task Space and any one of its constituent frameworks. For example, Laughlin and Ellis (1986) focus solely on constructs related to demonstrability. However, as Figure 5 makes clear, two tasks with comparable levels of demonstrability (encoded as similar positions on the *x*-axis) can nevertheless differ in other meaningful respects. While "Target Search" and "Word Completion Given Starting Letter" are similarly demonstrable, "Word Completion" was evaluated by raters as requiring far more creativity (0.42) than "Target Search" (0.15), as well as being more amenable for a divide-and-conquer strategy (0.78, versus 0.19). The ability to directly compare how different frameworks would evaluate the same set of tasks makes it easy to synthesize different perspectives. Consequently, we anticipate that researchers can transition seamlessly from using traditional task frameworks to using the Task Space.

Appendix F: A Large-Scale Integrative Experiment of Group Advantage – Study Design

F.1. Experiment Overview

Our study aims to examine how attributes of the task at hand can predict whether a team will outperform an equivalently sized collection of individuals working alone. We call this performance boost *group advantage*, and it is our primary dependent outcome. (Additional details on computing group advantage are provided in Appendix G.1.)

Across three waves of data collection, participants complete a series of tasks in an online laboratory built through Empirica and are randomly assigned to do so in one of three conditions: (1) alone; (2) in groups of three; or (3) in groups of six. The assigned tasks differ in their attributes (measured using the Task Space), and each task is presented repeatedly at different levels of complexity ("Low," "Medium," or "High"). These exogenously-varied factors—the size of the group, the attributes of the task, and the level of complexity of the task—serve as the primary predictors (independent variables) in our analysis, with special attention placed on the Task Space attributes.

By collecting data in three "waves," which introduce 10, 5, and 5 unique tasks, respectively, we demonstrate how our models can be trained on an initial set of tasks and used to generalize our findings to unseen tasks.

Our primary unit of analysis is the *condition*—a tuple of task × complexity × group size. When, in aggregate, those assigned to a given *condition* out-perform an equivalent number of individuals assigned to independently complete the same task at the same level of complexity, we say that group advantage is present.

F.2. Data Collection Waves and Task Selection

Our experiment proceeded in three waves, enabling us to make novel predictions by iteratively training on prior waves of data and making predictions about how teams would perform on subsequent waves. Furthermore, each wave consisted of a different set of tasks, allowing us to make predictions not only on unseen data ("out of sample"), but also on data with different task attributes ("out of distribution").

F.2.1 Task Selection Methodology

We proceeded in three data collection "waves," which involved 10, 5, and 5 tasks, respectively. To select tasks for the waves, we applied the following criteria:

- We focused on tasks that had been studied in both team and individual settings in prior literature (for example, the Room Assignment Task);
- We restricted the possible sample to tasks that could be conducted in an online team or individual experiment setup, specifically using the Empirica platform;
- We excluded tasks that involved requirements for formal roles, such as manager, leader, or subordinate.

With these constraints in mind, we selected tasks for waves using the following procedures:

Wave 1: We first selected 5 tasks for which we had a prior expectation of the presence of group advantage. We then selected 5 additional tasks that maximized cosine distance over the entire Task Space.

The selected tasks with an expectation for group advantage: *Divergent Association*, *Guessing the Correlation*, *Room Assignment*, *Sudoku*, *Whac-A-Mole*.

The selected tasks with no expectation: *Allocating Resources*, *Moral Reasoning*, *Wolf, Goat and Cabbage Transfer*, *Word Construction*, *Writing Story*.

Wave 2: We selected 5 tasks at random from the remaining tasks that satisfied our inclusion criteria: *Logic Problem*, *Random Dot Motion*, *Recall Word Lists*, *Typing game*, *Unscramble Words*.

Wave 3: We selected 5 tasks at random from the remaining tasks that satisfied our inclusion criteria: *Advertisement Writing*, *Putting Food into Categories*, *Recall Association*, *Search for Oil* (also known as *Wildcat Wells*), *Wildcam Gorongosa (Zooniverse)*. Note: Search for Oil and Wildcat Wells are two separate tasks in the Task Space. However, they are almost identical in practice.

In our analyses, we treat the first two waves as the "training" data, as they represent a mixture of randomly-selected and hand-selected tasks, from which we would be likely to learn meaningful information about group advantage. We evaluate our models on the "held-out" third wave.

F.2.2 Task Breakdown and Data Collection Timeline

The breakdown of tasks across the three waves took place as follows:

F.2.2.1 Wave 1 (10 Tasks)

Data collection began on 04/07/2023 and ended on 09/06/2023. Data included 115 individuals, 72 groups of 3, and 53 groups of 6.

1. **Sudoku** (Engel et al. 2014)
2. **Guess the Correlation** (Almaatouq et al. 2020)
3. **Writing Story** (Hackman, Jones, and McGrath 1967)
4. **Whac a Mole** (Naber, Pashkama, and Nakayama 2013)
5. **Allocating Resources** (Whiting et al. 2019)
6. **Word Construction** (Finlay et al. 2000)
7. **Divergent Association** (Olson et al. 2021)
8. **Wolf Goat Cabbage** (Kennedy 2009)
9. **Moral Reasoning** (Woolley et al. 2010)
10. **Room Assignment** (Almaatouq et al. 2021)

F.2.2.2 Wave 2 (5 Tasks)

Data collection began on 04/19/2024 and ended on 04/24/2024. Data included 42 individuals, 29 groups of 3, and 25 groups of 6.

1. **Unscramble Words** (Engel et al. 2014)
2. **Recall Word Lists** (Takahashi 2010)
3. **Typing** (Woolley et al. 2010)
4. **Logic Problem** (Littlepage 1991)
5. **Random Dot Motion** (Moussaïd et al 2017)

F.2.2.3 Wave 3 (5 Tasks)

Data collection began on 04/25/2024 and ended on 04/29/2024. Data included 36 individuals, 33 groups of 3, and 28 groups of 6.

1. **Wildcat Wells** (Mason and Watts 2011)
2. **Putting Food Into Categories** (Choi and Thompson 2006)
3. **Advertisement Writing** (Whiting et al. 2019)
4. **Recall Association** (Takahashi 2010)
5. **WildCam** (Straub, Tsvetkova, and Yasseri 2023)

F.3. Recruitment and Experimental Design

This section describes the shared experimental design and procedures for recruiting participants across all “waves” of data collection.

F.3.1 Population

Our population consisted of Amazon Mechanical Turk workers, whom we recruited to participate in a series of online, team-based tasks, described in Appendix F.5. Participants were selected from a filtered pool of MTurkers who had filled out psychometric surveys in a prior task.

F.3.2 Experimental Procedure

Participants were randomly assigned to conditions as follows:

- **Tasks:** Participants were assigned to complete five tasks from the pool of all possible tasks in a given experimental wave. Tasks were *block-randomized*, such that participants were randomly assigned one specific, pre-determined sequence of completing the tasks. This made it possible to create comparison groups between teams that were assigned to complete the exact same tasks in the exact same order, while also exploring different order variations. Table 4 lists the different block randomization options across each wave of data collection.
- **Group Size:** Participants were assigned to either groups of 1 member (independent work), 3 members, or 6 members.

Table 4: Block-Randomized Ordered Sequences for Tasks in Each Experimental Wave

Task Ordering	**Wave**
Writing Story, Moral Reasoning, Room Assignment, Wolf Goat Cabbage, Whac-A-Mole	1
Guess the Correlation, Whac-A-Mole, Word Construction, Wolf Goat Cabbage, Allocating Resources	1
Writing Story, Divergent Association Task, Guess the Correlation, Wolf Goat Cabbage, Whac-A-Mole	1
Room Assignment, Wolf Goat Cabbage, Guess the Correlation, Sudoku, Moral Reasoning	1

Table 4: Block-Randomized Ordered Sequences for Tasks in Each Experimental Wave	
Divergent Association Task, Sudoku, Moral Reasoning, Room Assignment, Writing Story	1
Whac-A-Mole, Writing Story, Word Construction, Allocating Resources, Divergent Association Task	1
Divergent Association Task, Sudoku, Guess the Correlation, Allocating Resources, Word Construction	1
Word Construction, Allocating Resources, Sudoku, Room Assignment, Moral Reasoning	1
Typing, Recall Word Lists, Unscramble Words, Logic Problem, Random Dot Motion	2
Recall Word Lists, Random Dot Motion, Typing, Unscramble Words, Logic Problem	2
Recall Word Lists, Logic Problem, Random Dot Motion, Unscramble Words, Typing	2
WildCam, Putting Food Into Categories, Recall Association, Wildcat Wells, Advertisement Writing	3
Putting Food Into Categories, Advertisement Writing, WildCam, Recall Association, Wildcat Wells	3
Putting Food Into Categories, Wildcat Wells, Advertisement Writing, Recall Association, WildCam	3

Table 4. Wave 1 involved 8 different task sequences; Wave 2 and 3 each involved 3 task sequences. Using a smaller number of task sequences makes it possible to make statistical comparisons within participants who experienced the exact same experimental set-up (e.g., identical tasks in an identical order) while also testing variation across different orderings.

For each task (e.g., Room Assignment Task) that participants were assigned to complete, participants were presented with three different instances that vary in their complexity (i.e., an easy version of the task, a moderately complex one, and a highly complex one) in a randomized order. This experimental set-up is summarized in Figure 6.

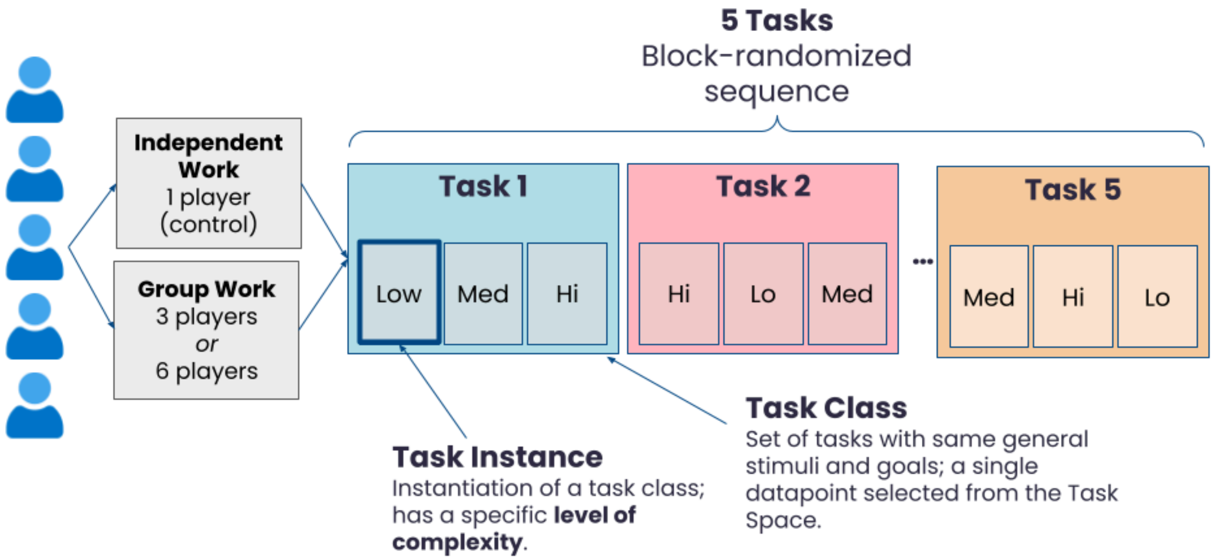

5 Tasks
Block-randomized sequence
Independent Work
1 player
(control)
Group Work
3 players
or
6 players
Task 1
Low
Med
Hi
Task 2
Hi
Lo
Med
...
Task 5
Med
Hi
Lo
Task Instance
Instantiation of a task class; has a specific level of complexity.
Task Class
Set of tasks with same general stimuli and goals; a single datapoint selected from the Task Space.

Figure 6. Experimental Set-Up. Participants were randomized into playing alone or with a team. Each collaboration unit (a solo player or team) played a sequence of five tasks, selected from the block-randomized ordering for a given wave. Within each task, they also played three versions varying in complexity, which were presented in a random order.

F.3.3 Experiment Mechanics

After receiving basic instructions, participants were presented with four versions of each task: one practice round (in which they familiarized themselves with the rules and interface), and the three graded versions (or "instances") of the task, which varied in their complexity.

Where possible, complexity levels were established by drawing 3 distinct versions of tasks used in prior literature. However, in some cases, prior literature had not established clear versions, so our research team generated parametric stimuli using the same methodology described in the papers introducing each task. For example, in Sudoku, the research team randomly sourced "easy," "medium," and "hard" puzzles from an online repository of Sudoku puzzles.

Participants played each of the five tasks sequentially; within each task, all members of the same team shared a persistent text-based chat at the right hand of the screen, allowing them to coordinate and discuss their strategy (in individual conditions, the text-based chat panel was grayed out, and the user saw the words "Chat Disabled"). Since participants played three graded versions for five tasks, we collected 15 graded task instances for each team. However, in rare cases of technical failure or participant drop-off, some teams left the study before completing all 15 task instances. When this occurred, we included all valid data from available instances.

To facilitate authentic collaboration, the game interface detected players who were using another tab or window for more than 60 seconds, or who did not use their keyboard or mouse for more than 120 seconds. If either of these conditions occurred, participants received a warning message. Participants who failed to respond to the message were removed from the game. Where relevant, the game platform disabled copy/paste to discourage cheating.

F.3.4 Participant Incentives

Compensation was linked to performance, incentivizing participants to achieve the best possible score. Participants received a base pay for showing up to the experiment, and an additional payment computed by (Payment Rate) × (Score) for each task. Participants were paid via a bonus at the end of the game.

F.3.5 Task Scoring

We operationalized performance for each task separately, and details can be found in Appendix F.5. The separate scoring of tasks took place because the tasks were heterogeneous by design; some tasks involved a specific correct answer, while other, often more open-ended tasks, required separate evaluation after the conclusion of the game. Still other tasks were dynamic, and involved accumulating an increasing number of points throughout the game. To create a consistent measure of performance, we min-max normalize all scores to a range of 0 (assigned to the lowest-performing individual or team for a given task) to 100 (assigned to the highest-performing individual or team for a given task). This formula worked as follows:

Scaled Score = 100 × Participant's Score / max(All Scores)

F.3.6 Recruitment Procedure

We used the online labor market Amazon Mechanical Turk (http://mturk.com, MTurk) to recruit participants. Participants originated from a pre-selected pool of workers who had previously completed qualifying tasks in a timely manner. The qualifying task included a battery of psychometric and demographic surveys (detailed further in Appendix F.4), including the Cognitive Reflection Test (CRT; Frederick, 2005; Appendix F.4.1); the Reading the Mind in the Eyes Test (RME; Baron-Cohen et al., 2001; Appendix F.4.2); and a measure of participants' Individual, Relational, and Collective Selves (IRCS; Brewer & Gardner, 1996; Appendix F.4.3), and demographic measures (Appendix F.4.4). The qualifying task thus served two purposes: first, it gave us psychometric information about the participants, from which we could analyze team composition; second, it allowed us to coordinate and re-contact participants from a pool that we knew to be of high quality.

Online games were conducted in a series of synchronous batches, which took place over the course of several weeks until we reached a desired sample size of at least 25 observations of each team size and task instance. The size goal was established by evaluating out of sample predictive accuracy in the first wave of data collection, which had indicated that, at around 23 observations per unit, predictive accuracy reached an asymptote and was not improving with more data samples. Counter to *p*-hacking, this method relied on a pre-set stopping rule which was then applied in future data collection waves once established.

Up to a few hours prior to each batch, participants received the following recruitment email informing them that a paid opportunity was due to take place. The email was titled, "**Join a task today at [TIME] EDT to earn at least $1 pay.**"

> Hi **[MTURK ID]** — join a task at **[TIME]** AM Eastern (**[TIME]** AM Central, **[TIME]** AM Pacific), today. You are welcome to participate again if this is not the first time you have received a message like this.
>
> IMPORTANT:

1. $1 MIN PAY — you will receive at least a $1 bonus if you just turn up and follow instructions before the task starts. The link has information about how pay works for the full task.
2. THE FULL TASK CAN TAKE AROUND 1 HOUR TO COMPLETE — please only join if you can stay for the entire time, but also be aware that you may not have the opportunity to participate for the entire time, or may finish early.
3. THERE ARE LIMITED SPACES — there are a limited number of spots for the full task. If you don't get a spot in the full task, but follow all the directions at the link, you will still receive a $1 bonus.
4. YOU MUST USE A COMPUTER (and CHROME works best) — The task will not work on a mobile phone or tablet. Several browsers will work but the Chrome browser tends to work best.
5. THERE WILL BE MORE OPPORTUNITIES TO PARTICIPATE — If you can't make it now, or can't make it for the entire time now, there will be many more opportunities in the future. Also, if you join in, but are not assigned a spot in the full task, you will still be able to join in a future task.

Join at **[TIME]** AM EDT on today by clicking here: **[URL]**

At the appointed time, participants who joined via the provided URL were directed to a series of online games hosted with the Empirica (https://empirica.ly/) platform. Once sufficient participants had joined to fill a game, participants were directed to begin the task.

F.4. Group Composition Variables from Qualifying Task

Prior to being recruited for the main experiment, participants had completed a battery of psychometric and demographic surveys, and we treat these variables as the core compositional attributes of the group. The present section describes each of these surveys in detail, as well as provides basic summary statistics about the participants along these compositional attributes.

F.4.1 Cognitive Reflection Test (CRT)

Participants completed a Cognitive Reflection Test (Table 5), in which they were asked to answer numerical questions with an incorrect "fast lure" but a correct answer that required slower, deliberate thinking (Frederick, 2005). Our questions were inspired by the more recently updated CRT-4 (Toplak et al., 2014).

Table 5: Cognitive Reflection Test (CRT) Questions

No.	Question	Answer
1	A drill and a hammer cost $330 in total. The drill costs $300 more than the hammer. How much does the hammer cost?	$15
2	Rachel is the 10th tallest and the 10th shortest girl in her class. How many girls are in her class?	19
3	When on sale for 20% off, a toaster costs $100. What does it cost when not on sale?	$125

Table 5: Cognitive Reflection Test (CRT) Questions		
4	In a pail of 60 apples, red apples are 3 times more common than green ones. How many are green?	15
5	After hatching from its egg, a baby bird doubles in weight every day. By day 12 it weighs a pound. On what day does the bird weigh half a pound?	11
6	A dog and a cat weigh 100 pounds in total. The dog weighs 86 pounds. What is the difference in weight between the dog and the cat?	72

Table 5. Questions in the Cognitive Reflection Test.

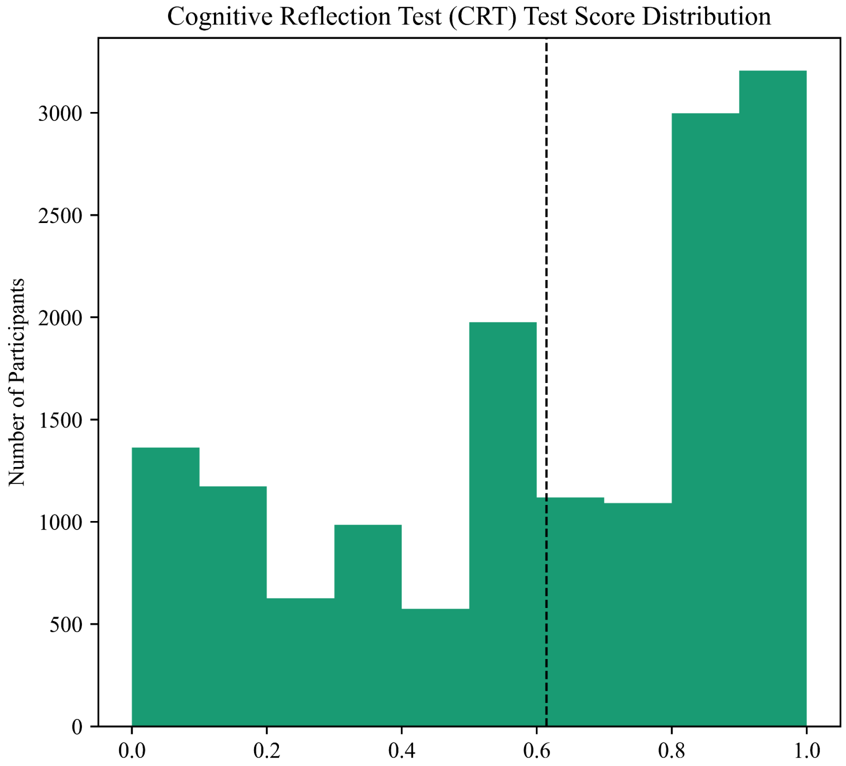


Figure 7. The distribution of CRT scores among participants. The mean score was 61%, as indicated by the dotted line. There was a relatively high standard deviation (32%), as many participants appeared to either get all of the questions right (100%) or all of the questions wrong (0%).

F.4.2 Reading the Mind in the Eyes Test (RME)

Participants were also asked to complete a 36-question "Reading the Mind in the Eyes" (RME) test, a measure of social perceptiveness (Baron-Cohen et al. 2001). The full RME test used can be found at this link: https://raw.githubusercontent.com/Watts-Lab/surveyor/prod/surveys/RME.csv.

In the test, participants are shown a black-and-white image of a pair of eyes, and are asked to select one of four words to describe what the person is thinking or feeling. A sample pair of eyes is provided in Figure

8; for this question, the options were “jealous,” “panicked,” “arrogant,” and “hateful,” with “panicked” being the correct answer.

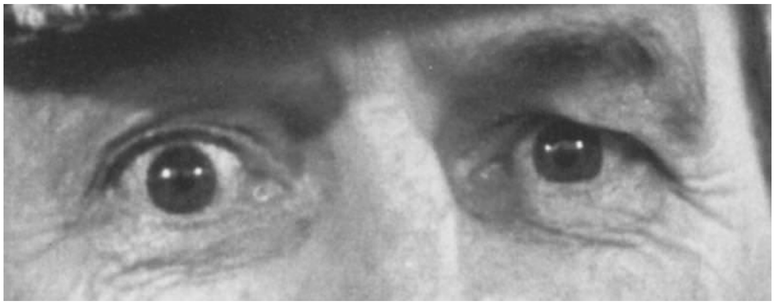


Figure 8. Sample image from the Reading the Mind in the Eyes (RME) Task. Here, the correct description is 'panicked.'

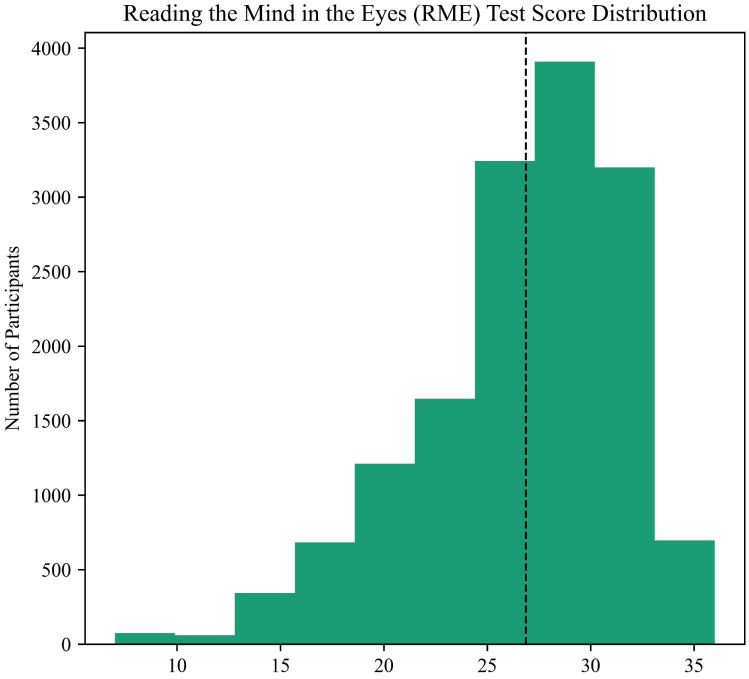


Figure 9. The distribution of RME scores among participants. Participants answered 26 (out of 36) questions correctly on average, as indicated by the dotted line; the standard deviation was 5 questions.

F.4.3 Individual, Relational, and Collective Selves (IRCS)

The Individual, Relational and Collective Selves (IRCS) inventory captures the extent to which a person is individualistic versus collectivistic. It consists of the following six subscales: Individual Self (IS); Relational Self (RS); Group Self (GS); Individual Agency (IB); Group Value (GV); and Individual Value (IV), which are detailed in Table 6.

Table 6: Individual, Relational, and Collective Selves (IRCS) Inventory

No.	Prompt	Subscale	Coding

Table 6: Individual, Relational, and Collective Selves (IRCS) Inventory

1	I enjoy being unique and different from others in many ways.	IR	Positive
2	I often do “my own thing.”	IR	Positive
3	I am a unique individual.	IR	Positive
4	My happiness depends very much on the happiness of those around me.	RS	Positive
5	I often have the feeling that my relationships with others are more important than my own accomplishments.	RS	Positive
6	If a coworker got a prize, I would feel proud.	RS	Positive
7	To me, pleasure is spending time with others.	RS	Positive
8	The well-being of my coworkers is important to me.	RS	Positive
9	I feel good when I cooperate with others.	RS	Positive
10	Overall, my group memberships have very little to do with how I feel about myself.	GS	Positive
11	The social groups I belong to are an important reflection of who I am.	GS	Positive
12	In general, belonging to social groups is an important part of my self-image.	GS	Positive
13	The social groups I belong to are unimportant to my sense of what kind of a person I am.	GS	Positive
14	What happens to me is my own doing.	IB	Positive
15	I tend to do my own things, and most people in my family do the same.	IB	Positive
16	Individuals should be judged on their own merits not on the company they keep.	IB	Positive

Table 6: Individual, Relational, and Collective Selves (IRCS) Inventory			
17	When faced with a difficult person problem, it is better to decide what to do yourself rather than follow the advice of others.	IB	Positive
18	People should be aware that if they are going to be part of a group, they sometimes will have to do things they don't want to do.	GV	Positive
19	I usually sacrifice my self-interest for the benefit of the group I am in.	GV	Positive
20	It is important to me to respect decisions made by the group.	GV	Positive
21	If the group is slowing me down, it is better to leave it and work alone.	GV	Positive
22	I will stay in a group if they need me, even when I'm not happy with the group.	GV	Reverse
23	One should live one’s life independent of others as much as possible.	IV	Positive

Table 6. Items in the Individual, Relational, and Collective Selves (IRCS) Inventory.

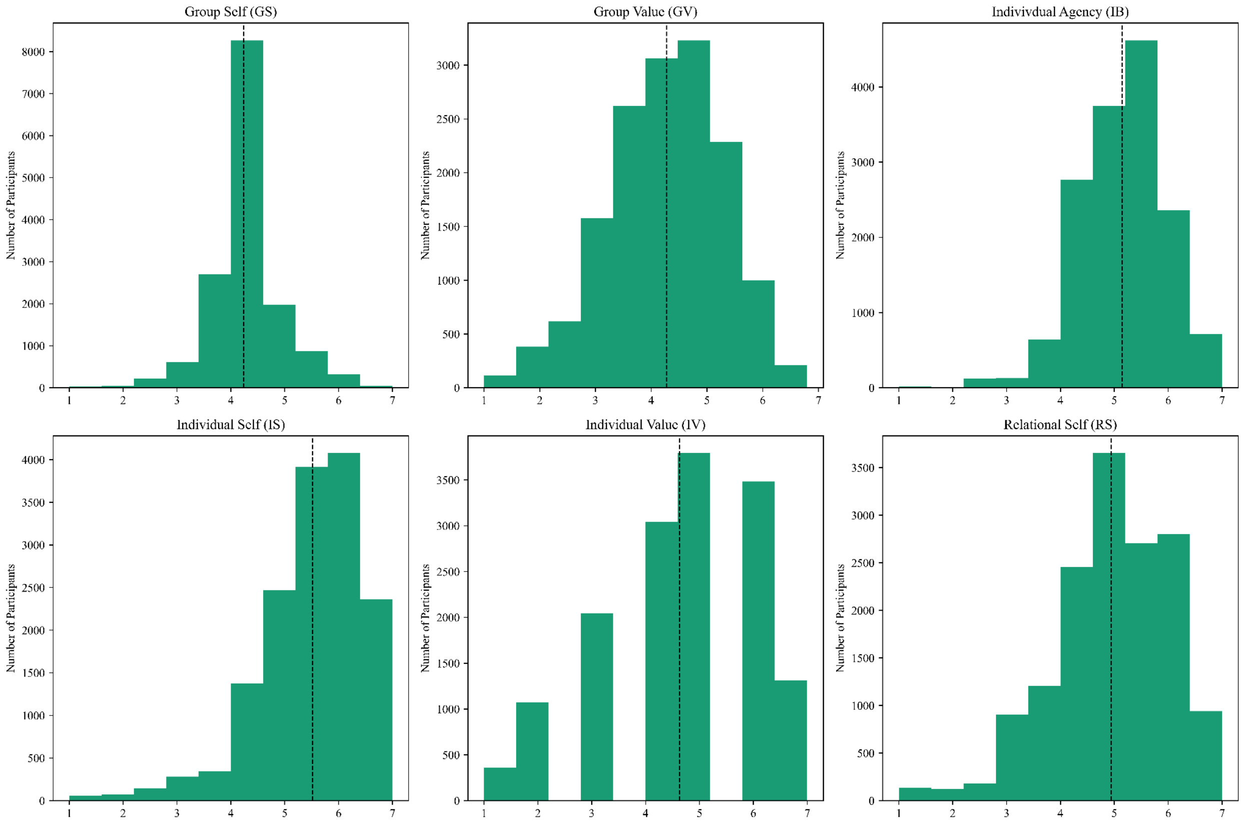


Figure 10. The distribution of IRCS scores among participants. The dotted line indicates the mean.

F.4.4 Demographic Measures

Finally, participants completed an 11-question battery of demographics questions, which are detailed in Table 7.

Table 7: Demographics Questions

Question ID	Prompt	Response Options
birth_year	In which year were you born?	Number
education	What is the highest level of education you have completed?	Less than a high school diploma, High school diploma, Some college or vocational training, 2-year college degree, 4-year college degree, Post-college degree, Other

Table 7: Demographics Questions		
income	Thinking back over the past year, what was your family's annual income?	Less than $10,000, $10,000-$19,999, $20,000-$29,999, $30,000-$39,999, $40,000-$49,999, $50,000-$74,999, $75,000-$99,999, $100,000-$149,999, $150,000 or more
political_fiscal	Thinking about economic issues, which of the following best describes your views?	Strongly Liberal, Somewhat Liberal, Moderate, Somewhat Conservative, Strongly Conservative
political_social	Thinking about social issues, which of the following best describes your views?	Strongly Liberal, Somewhat Liberal, Moderate, Somewhat Conservative, Strongly Conservative
political_party	Generally speaking, do you consider yourself a...	Democrat, Republican, Independent, Other Party
gender	What is your gender?	Male, Female, Other
race	Please choose one or more races that you consider yourself to be	White, Black or African-American, American Indian or Alaska Native, Asian, Native Hawaiian or other Pacific Islander, Other
country	In which country do you reside?	Text
zip_code	In which Zip Code do you reside?	Text
marital_status	What is your marital status?	Single Never Married, Married or Domestic Partnership, Widowed, Divorced, Separated

Table 7. Demographic Questions in the MTurk Panel.

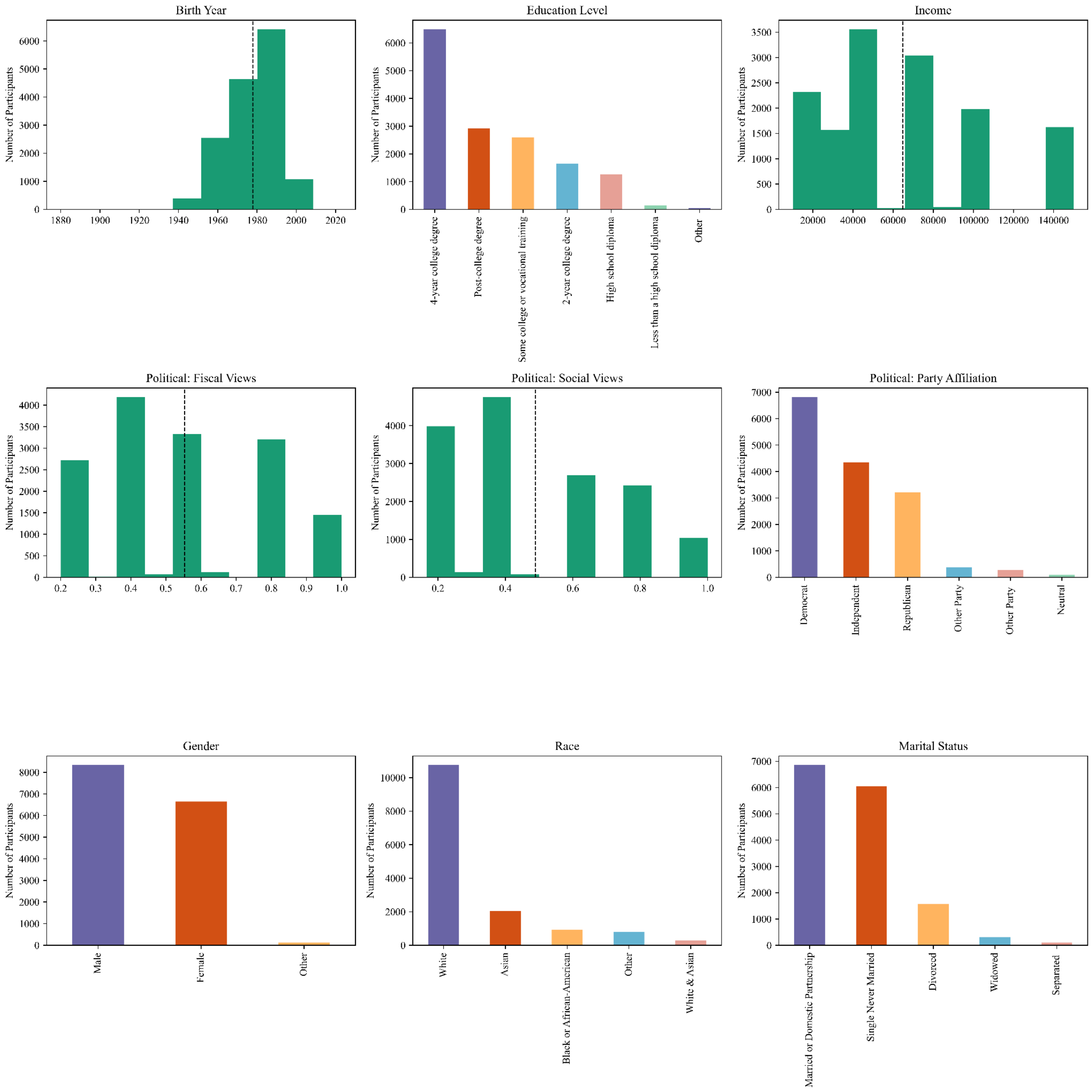


Figure 11. The distribution of demographics among participants. For numerical attributes, the dotted line indicates the mean.

Figure 11 presents our participants' demographics. On average, participants were born in 1977 (making them approximately 47 years old at the time of the experiment); they had an income of $64,771; and they had some higher education (the modal participant had a four-year college degree). Politically, our participants skew slightly left on the spectrum (with the modal participant identifying as a Democrat). Our participants are 55% male and 44% female; the vast majority (73%) identify as White. The modal participant is married or in a domestic partnership, followed closely by those who identify as single. Of course, as is clear in Figure 11, our participants occupy a wide spectrum on most of these dimensions.

F.5. Detailed Task Descriptions

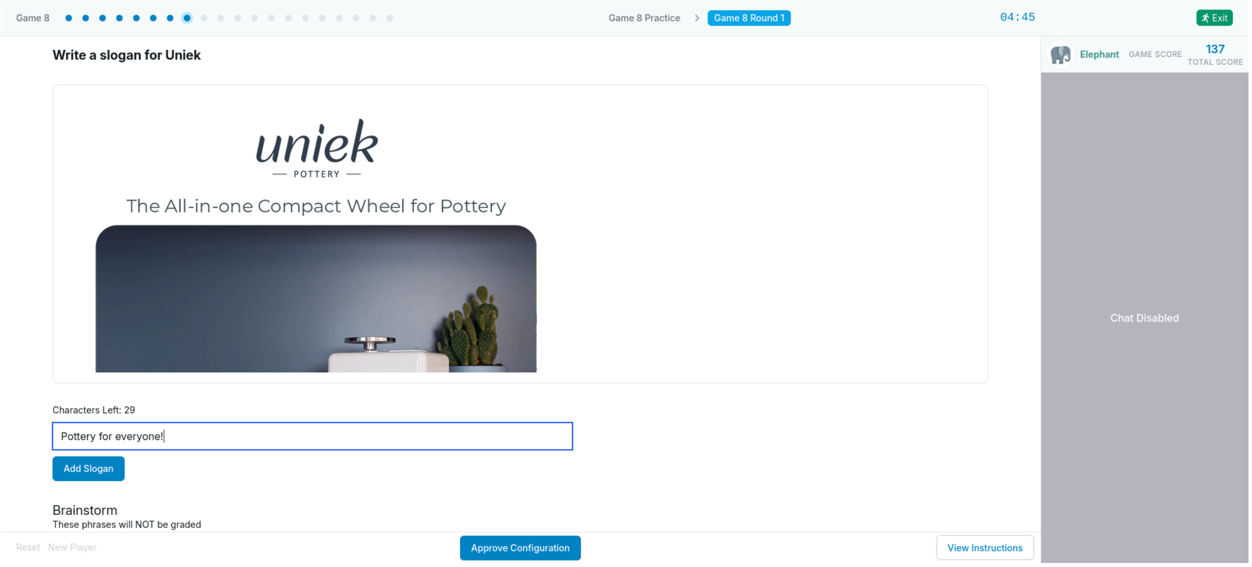
Each task was implemented as a game on Empirica. We constructed a custom wrapper that extends Empirica to make the games look and feel consistent and have similar properties and data storage for commensurability in downstream analysis. Each task was implemented so that the specific stimulus complex of a task instance was defined as parametric constants. We defined 4 sets of constants for each task: (1) a practice round, (2) a low complexity round, (3) a medium complexity round, and (4) a high complexity round. Additionally, we defined a standard protocol for consistently displaying instructions and evaluating if participants understood them including information about the gameplay, any special rules, and how scoring works for the task.

The following section lists the specifics of how each of these was determined on a per-task level and provides screenshots, instructions, and other details about each task. Full code for each task, including exact stimuli as instance constants, is available on Github at github.com/Watts-Lab/multi-task-empirica.

F.5.1 Advertisement Writing

Advertisement Writing challenges participants to create compelling advertisements for a product drawn from Kickstarter or a similar repository of relatively unknown novel product ideas with detailed descriptions. The task involves generating advertisement slogan concepts and then selecting a final candidate.

Participants had **5 minutes** to complete the task.



Write a slogan for Uniek

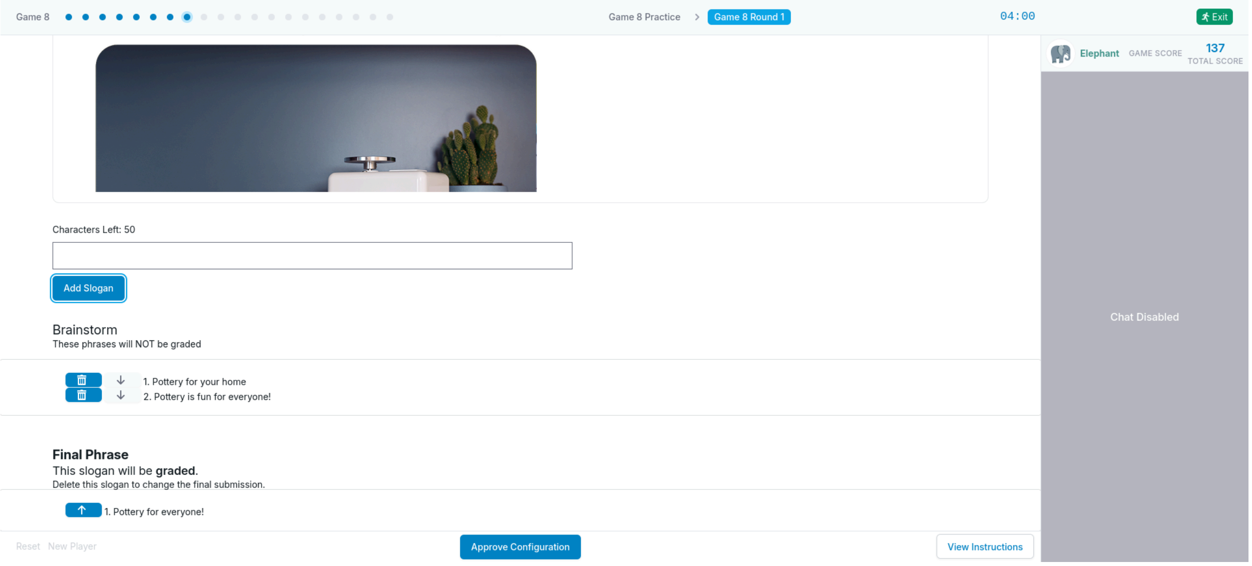
Characters Left: 29

Pottery for everyone!

Brainstorm

These phrases will NOT be graded

Figure 12. Advertisement Writing participants first write advertisement ideas in an interface that provides independent input.



Characters Left: 50

Brainstorm
These phrases will NOT be graded

1. Pottery for your home
2. Pottery is fun for everyone!

Final Phrase
This slogan will be **graded.**
Delete this slogan to change the final submission.

1. Pottery for everyone!

Figure 13. Advertisement Writing participants see contributions from everyone in their group and can collaborate to select the advertisement text they prefer for the final submission.

F.5.1.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will see a product description and then write a 30-character advertising slogan to be used in an online advertisement.

Goal: Create a slogan to advertise the product that maximizes interest in the product.

Rounds: Each round will contain a different product.

Instructions:

- Write as many slogans as you like.
- Select a single slogan to submit for scoring.
- The slogan must be 30 characters or fewer.
- You will not receive a score if a final slogan is not indicated.
- *If in a team condition:* Your team works on the same slogan and will submit one shared response. All players can click “Approve Configuration” to move on.
- You can click “Approve Configuration” to move on.

Scoring Information:
Your score is determined by the click-through rate the final slogan achieves on an online advertising platform. This score will be calculated after the experiment ends.

Comprehension Quiz:

- What is the main activity in this game?
  - To create a list of ad slogans for coffee mugs.
  - To cast actors in an ad.

    - To write a slogan for a given product. [Correct Answer]
- Will this game have a score at the end of each round?
    - Yes, it is calculated at the end of each round.
    - No, it is calculated after the experiment ends. [Correct Answer]

F.5.1.2 Instances

Instances were drawn from completed campaigns on Kickstarter.com. The research team selected instances that had a relatively understandable but novel product. Complexity levels, for instance, were established by ranking by the ratio of the amount of money a campaign requested and the amount of money it received such that campaigns that received much more than they requested were deemed easier to advertise than those that did not.

F.5.1.3 Scoring

Submissions were scored based on the likelihood of receiving a click in an online advertisement marketplace, simulated via a large language model-based pipeline. The language model was provided the advertisement with the following prompt:

“If you saw the user-provided advertisement alongside your search results, would you click on it? Please answer with either yes or no.”

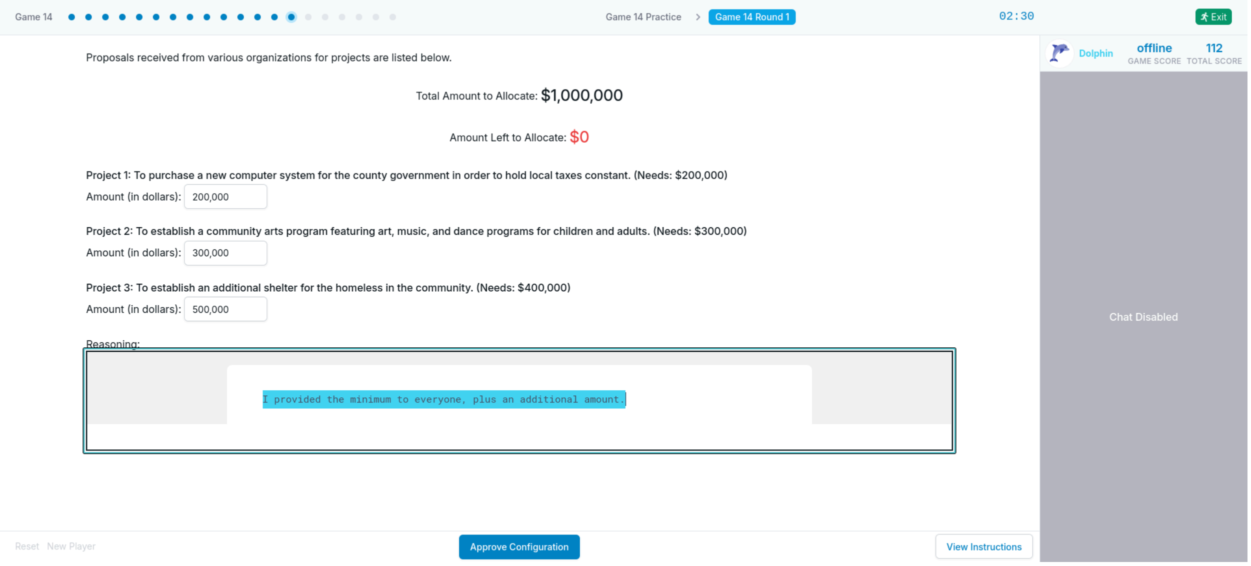
This was done using the **GPT4o chat API** with a temperature setting of 0 and configured to elicit the log probabilities of an answer of either “yes” or “no.” These were then exponentiated to acquire the probability of a submission receiving a click. This technique was validated with submissions to a real-world ad market, showing that submissions for a given campaign had a rank correlation above .73. Scores were then scaled to cover a range of 0–100 in the same ways other tasks were rescaled.

F.5.2 Allocating Resources

The Allocating Resources requires participants to distribute limited resources across various needs or projects to maximize overall benefit. This task tests players’ ability to make strategic decisions under constraints.

The high-level objective of this task is to assess decision-making skills, prioritization abilities, and understanding of trade-offs in resource management.

Participants had **3 minutes** to complete the task.



Proposals received from various organizations for projects are listed below.

Total Amount to Allocate: $1,000,000

Amount Left to Allocate: $0

Project 1: To purchase a new computer system for the county government in order to hold local taxes constant. (Needs: $200,000)
Amount (in dollars): 200,000

Project 2: To establish a community arts program featuring art, music, and dance programs for children and adults. (Needs: $300,000)
Amount (in dollars): 300,000

Project 3: To establish an additional shelter for the homeless in the community. (Needs: $400,000)
Amount (in dollars): 500,000

Reasoning:

I provided the minimum to everyone, plus an additional amount.

Figure 14. Allocating Resources participants see requests and provide a dollar amount contribution to each request. The total remaining funds are shown and calculated dynamically. Participants are also provided a collaborative writing environment to explain the reasoning behind their allocation.

F.5.2.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will evaluate competing requests for funding and decide on funding amounts for each request.

Goal: Allocate funds based on your values and describe your reasoning.

Rounds: Will have different starting funding amounts and/or different requests.

Instructions:

- *If in a team condition:* Your team works on evaluating the same requests and will submit one shared response for each request. All players must agree to move on.
- Each project is in need of money but can benefit from any contribution that you might make. The greater the contribution that you make to a particular project, the more likely it is that the project will succeed.
- Using the Amount box, type a dollar amount of funding for each request.
- Using the Reasoning box, explain exactly why you made the allocation for all requests.
- Allocate all funds in each round.
- Do not allocate more than you are given.

Scoring Information:
Your score is determined by giving complete and coherent reasons for your allocation of funds. The explanation should address all requests. This score will be calculated after the experiment ends.

Comprehension Quiz:

- What is the main activity of this game?

    - To allocate people to rooms and explain your reasons.
    - To allocate funds to requests and explain your reasons. [Correct Answer]
    - To allocate animals to owners and explain your reasons.
- Will this game have a score at the end of each round?
    - Yes, it is calculated at the end of each round.
    - No, it is calculated after the experiment ends. [Correct Answer]

F.2.2.2 Instances

Task complexity is adjusted by forcing more competition for funds by having the total amount requested across all projects exceed the total allocatable funds by increasing amounts—for example, the low complexity instance asked participants to allocate $1,000,000 across projects requesting $200,000, $300,000, and $400,000 (thus making it easy to give all the projects what they requested), while the high complexity instance asked participants to allocate $500,000 across three projects each requesting $500,000 (thus forcing the participants into a difficult decision).

The total amount available for allocation at each stage was as follows:

1. Low complexity: $1,000,000
2. Medium complexity: $500,000
3. High complexity: $500,000

The three projects used in the task were fixed, but the amount requested for each varied by complexity level. The projects and their requested amounts were as follows:

1. To purchase a new computer system for the county government in order to hold local taxes constant.
    - Instance one: $200,000
    - Instance two: $200,000
    - Instance three: $500,000
2. To establish a community arts program featuring art, music, and dance programs for children and adults.
    - Instance one: $300,000
    - Instance two: $300,000
    - Instance three: $500,000
3. To establish an additional shelter for the homeless in the community.
    - Instance one: $400,000
    - Instance two: $400,000
    - Instance three: $500,000

F.2.2.3 Scoring

Submissions were graded based on the reasoning they provided using a large language model-based pipeline, which would be given the submission justification and a submission rating in the context of the following prompt:

“You are a teacher grading an activity in which students allocate resources to projects. The total amount of money available is `total_available_money`. The projects are:

`project_title_i`, which is requesting `project_request_i` (repeated for each project).

You read the following reasoning and allocations and provide a grade between 0 and 100, based on how well it justifies the allocation of resources. If reasoning is missing, the score is 0.

Allocations:

Project i: `project_allocation_i` (repeated for each project).

Reasoning: `reasoning`

The total allocation was: `sum_allocations`

The numerical score:”

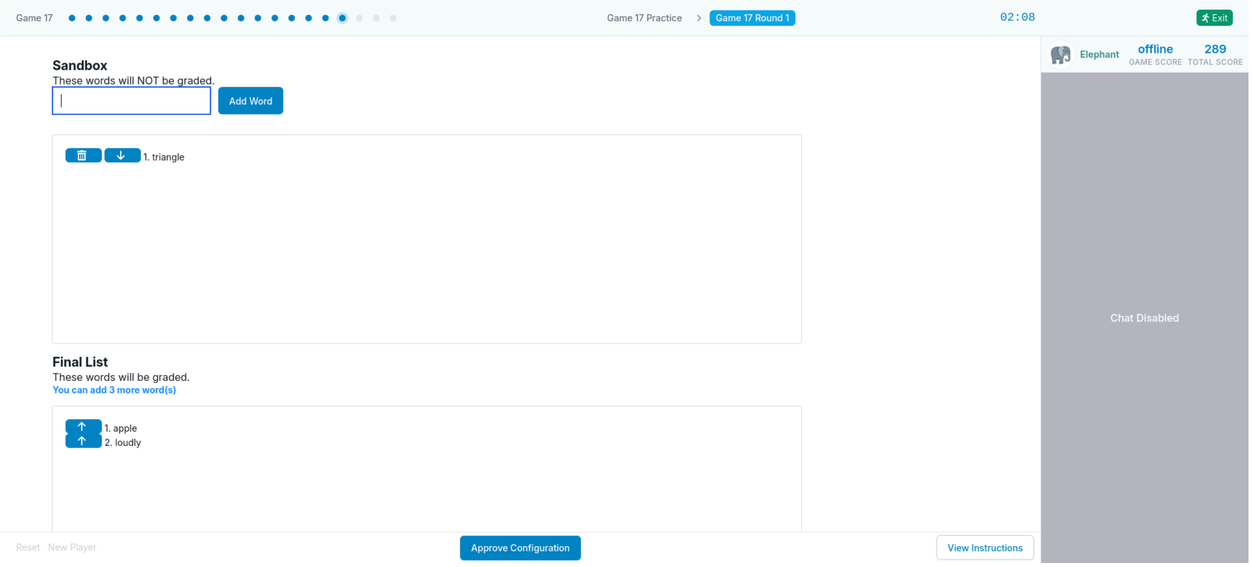
This was done using the **GPT3.5 completions API** with a temperature setting of 0 and a token limit of 1, so the model would respond with a repeatable number between 0 and 100.

F.5.3 Divergent Association

The Divergent Association requires participants to generate as many diverse and unrelated words as possible within a time limit. This task tests players’ ability to think creatively and produce various ideas.

The high-level objective of this task is to assess divergent thinking, cognitive flexibility, and the ability to access and utilize a broad range of semantic concepts.

Participants had **3 minutes** to complete the task.



Sandbox

These words will NOT be graded.

1. triangle

Final List

These words will be graded.

You can add 3 more word(s)

1. apple
2. loudly

Figure 15. The Divergent Association Task interface. Participants see a sandbox where they can enter their initial ideas, followed by a space to enter their final word list.

F.5.3.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will come up with a list of words.

Goal: Enter words that are as different from each other as possible, in all meanings and uses of the word.

Rounds: Will have different numbers of words to enter.

Instructions:

- *If in a team condition:* Your team works on the same list and will submit one shared response. All players must agree to move on.
- Entries in the list must be:
    - Single words in English
    - Only nouns (e.g. things, objects, concepts)
    - No proper nouns (e.g. no specific people or places)
    - No specialized vocabulary (e.g. no technical terms)
    - Think of the words on your own (e.g. do not just look at objects in your surroundings)
- Add words to the Sandbox List to brainstorm different possible entries [*If in a team condition:* with your team].
- Move words to the Final List before submission. [*If in a team condition:* Your team shares one final list.]
- Only words in the Final List will be graded.

Scoring Information:
Your score is determined by how different the words in the Final List are from each other, in all meanings and uses of the words. This score will be calculated after the experiment ends.

Comprehension Quiz:

- What is the main activity of this game?
    - To come up with words as different from each other as possible using the Final list. [Correct Answer]
    - To come up with the most number of words using the Sandbox list.
    - Both of the above.
- Will this game have a score at the end of each round?
    - Yes, it is calculated at the end of each round.
    - No, it is calculated after the experiment ends. [Correct Answer]

F.5.3.2 Instances

The number of words that participants were required to generate depended on the complexity level: 5 (low complexity), 10 (moderate complexity) or 15 (high complexity).

F.5.3.3 Scoring

GloVe vectors for each word were used to compute average cosine distance of all pairs of unique words in a submission. These were then scaled to a 0–100 range.

F.5.4 Guess the Correlation

Guess the Correlation presents participants with scatter plots and asks them to estimate the correlation coefficient between the two variables. This task tests players' ability to visually assess statistical relationships.

The high-level objective of this task is to assess visual-spatial skills, statistical intuition, and the ability to quickly estimate quantitative relationships from graphical data.

Participants had **2 minutes** to complete the task.

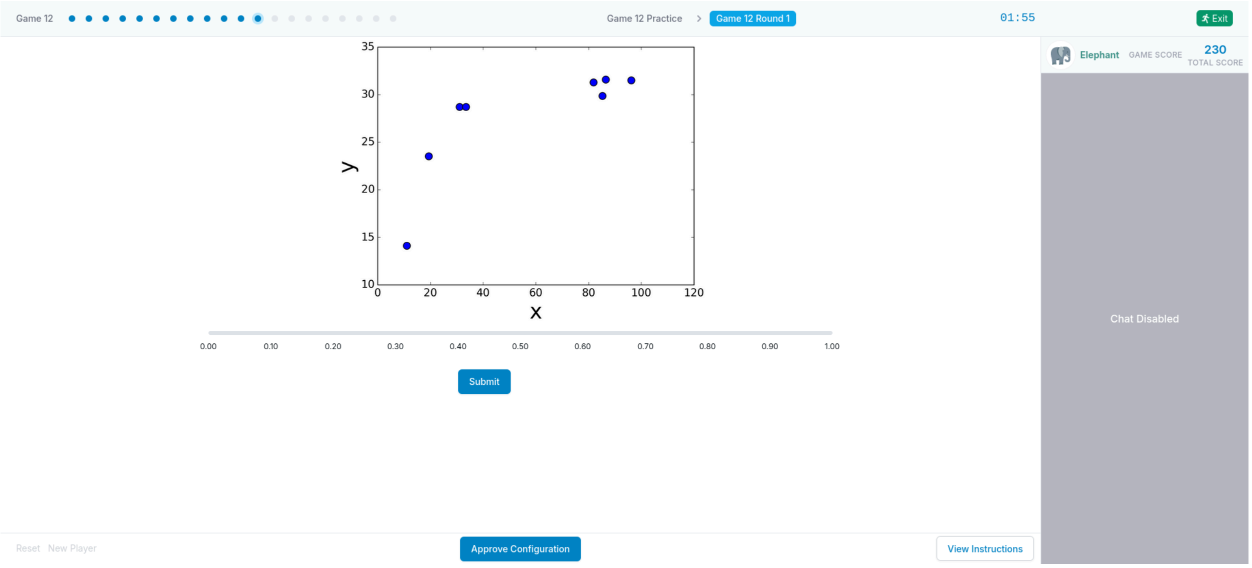


Figure 16. Starting configuration for Guessing the Correlation.

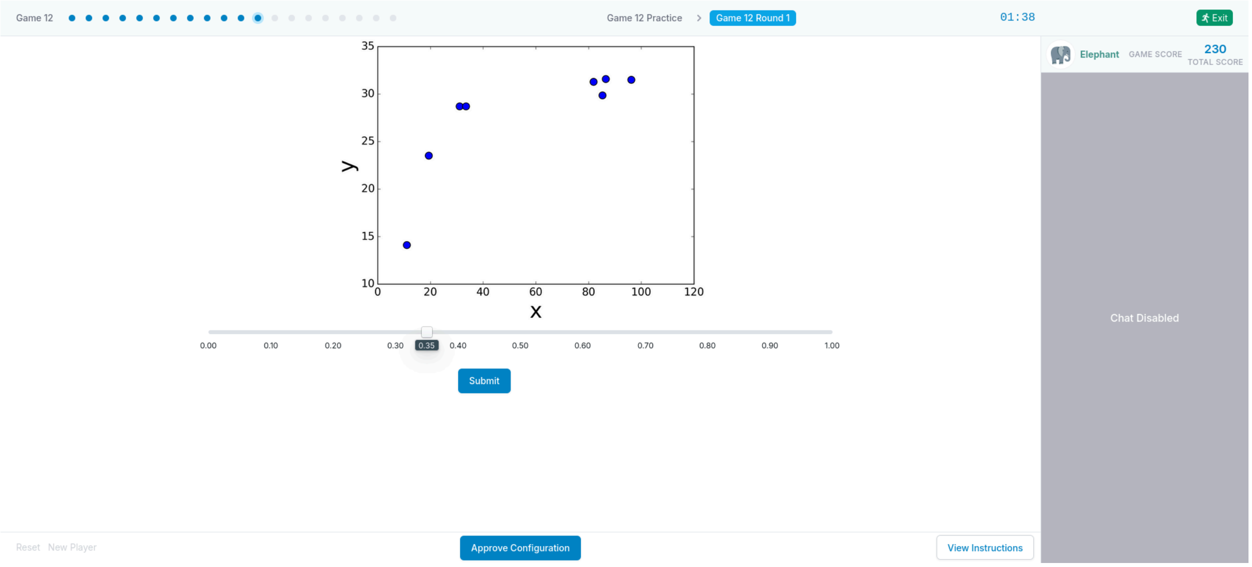
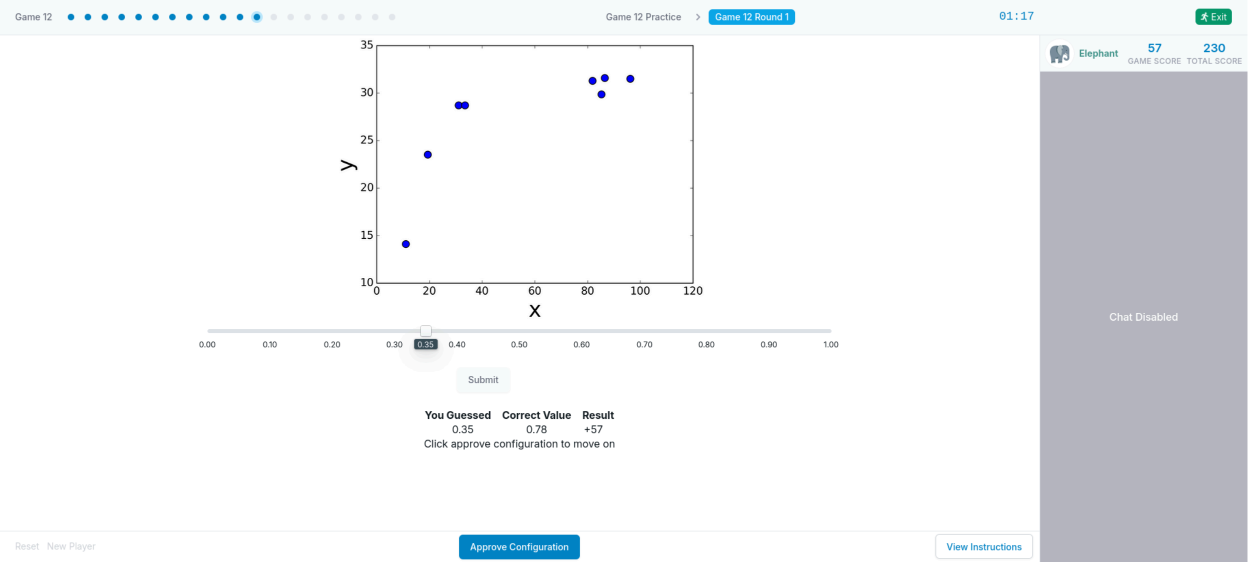


Figure 17. Guessing the Correlation interface as the participant enters their guesses.



You Guessed	Correct Value	Result
0.35	0.78	+57

Click approve configuration to move on

Figure 18. Guessing the Correlation game after showing the participant their guess, alongside the correct value and the number of points earned.

F.5.4.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will be shown graphs and asked to guess how correlated two variables are.

Goal: Make a guess as close as possible to the real correlation.

Rounds: Will have different graphs.

Instructions:

- *If in a team condition:* Your team will be shown a graph and asked to guess how correlated X and Y are. Your team works on the same guess and will submit one shared response. All players must agree to move on.
- You will be shown a graph and asked to guess how correlated X and Y are.
  - Use the slider to make a guess between zero and one.
  - Click the Submit button once your guess is set.
  - 0 is no correlation between X and Y and 1 is perfect correlation between X and Y
  - There are no negative correlations used in the game.

Scoring Information:
Your score is determined by how close your guess is to the actual correlation. The closer your guess to the actual correlation, the higher your score. This score will be calculated during the experiment.

Comprehension Quiz:

- What is the main activity in this game?
  - To guess a correlation [Correct Answer]
  - To move objects around

    - To move your cursor over dots
- Will this game have a score at the end of each round?
    - Yes, it is calculated at the end of each round. [Correct Answer]
    - No, it is calculated after the experiment ends.

F.5.4.2 Instances

Task complexity is influenced by the number of points shown and the degree of collinearity — low complexity had clear linearity with an ample number of points lying in a noisy line, medium complexity included fewer points with a less obvious correlation, and high complexity included many points with numerous outliers. Task instances were drawn directly from the task source.

F.5.4.3 Scoring

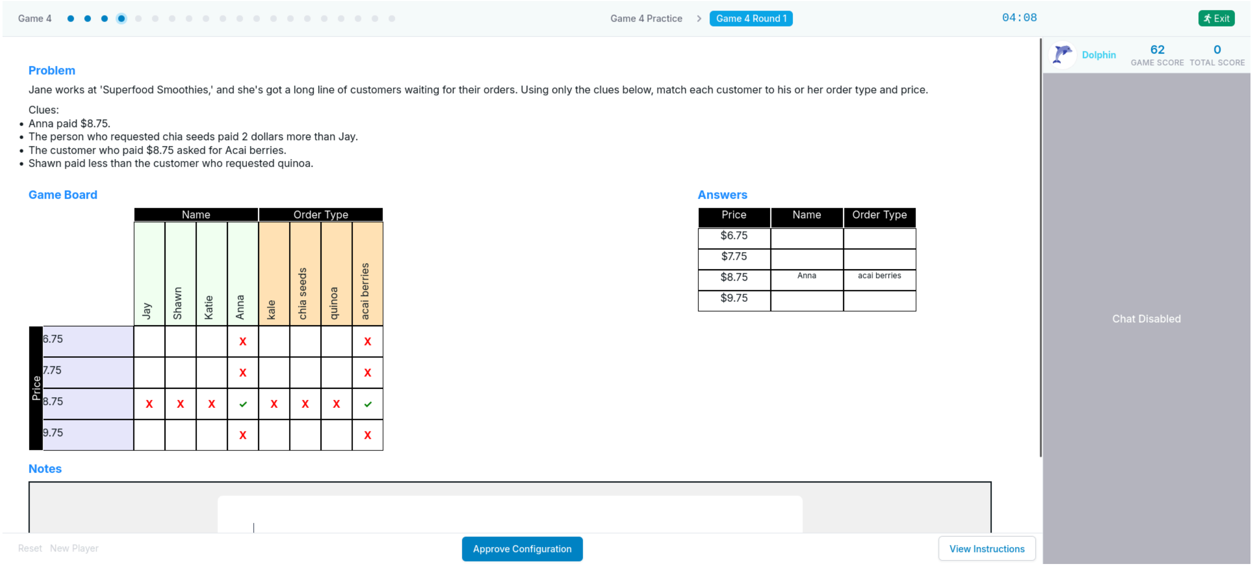
The absolute value of the difference between the guessed correlation and the actual correlation was used as the score. This was then normalized to a 0–100 range.

F.5.5 Logic Problem

The Logic Problem presents participants with a series of clues or statements from which they must deduce a correct solution. This task tests players' ability to use logical reasoning and deduction.

The high-level objective of this task is to assess analytical thinking, inference-making, and the ability to systematically eliminate possibilities.

Participants had **5 minutes** to complete the task.



Problem

Jane works at 'Superfood Smoothies,' and she's got a long line of customers waiting for their orders. Using only the clues below, match each customer to his or her order type and price.

Clues:

- Anna paid $8.75.
- The person who requested chia seeds paid 2 dollars more than Jay.
- The customer who paid $8.75 asked for Acai berries.
- Shawn paid less than the customer who requested quinoa.

Game Board

		Name				Order Type			
		Jay	Shawn	Katie	Anna	kale	chia seeds	quinoa	acai berries
Price	6.75				x				x
	7.75				x				x
	8.75	x	x	x	✓	x	x	x	✓
	9.75				x				x

Notes

Answers

Price	Name	Order Type
$6.75		
$7.75		
$8.75	Anna	acai berries
$9.75		

Figure 19. The Logic Problem interface.

F.5.5.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will use clues to find a solution to a puzzle.

Goal: Consider the clues provided and deduce the correct relationships.

Rounds: Each round will be about a different topic and have different clues.

Instructions:

- Read the situation and deduce details to fill in the logic table.
- You and your teammates are free to take notes on the interactive textbox provided while working on the solutions. The textbox will not be graded. (*If in an individual condition:* You are free to take notes on the textbox provided while working on the solutions. The textbox will not be graded.)
- Clicking on the box one time will show an "X." This means that the relationship is NOT true based on the clues.
- Clicking this box again will show an "✓." This means that the relationship is true based on the clues.
- A third click will clear the box.
- Your team will work on the same table and will submit one shared response. (*If in an individual condition:* You can click "Approve Configuration" to move on.)

Scoring Information:
You will receive points for each correct deduction, with a total of 100 points if all deductions are correct. The score will be calculated after each round is completed.

Comprehension Quiz:

- What is the main activity in this game?
  - ○ To solve different number problems
  - ○ To map relationships in a story
  - ○ To solve puzzles using clues [Correct Answer]
- Will this game have a score at the end of each round?
  - ○ Yes, it is calculated at the end of each round. [Correct Answer]
  - ○ No, it is calculated after the experiment ends.

F.5.5.2 Instances

Instances were varied in the number of clues and the amount of information in each clue. For example, the low complexity instance had clues that would, with no added consideration, answer some portion of the puzzle. The medium complexity instance had clues, all of which required considering at least one other clue to answer elements of the puzzle, and the high complexity instance required participants to consider all clues at the same time to make progress.

F.5.5.3 Scoring

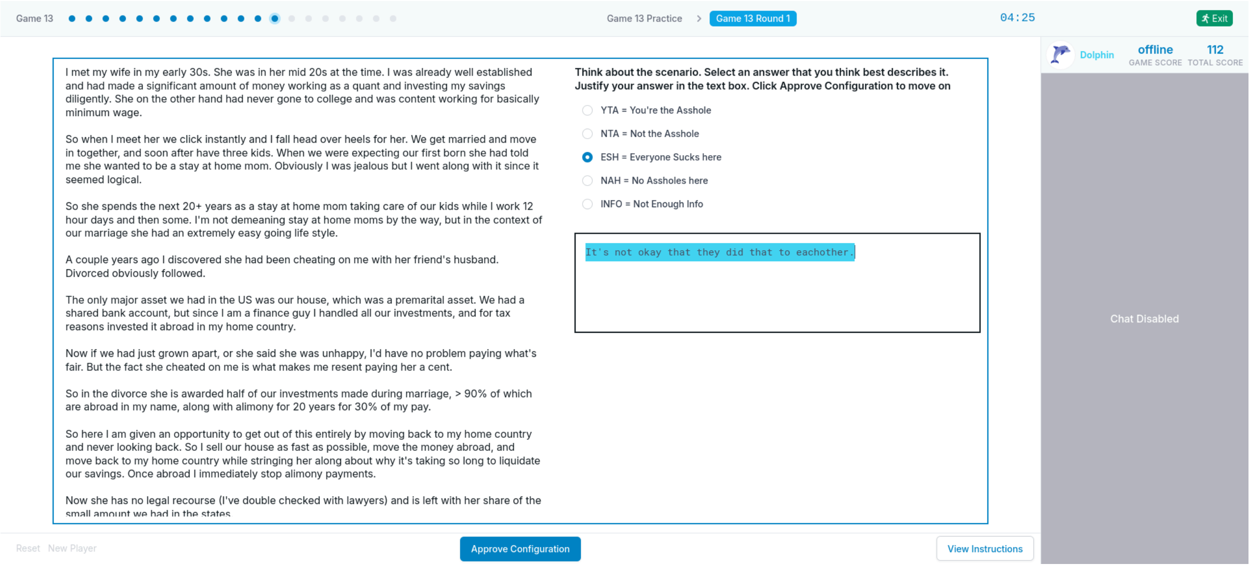
Submissions were scored by automatically comparing them to the correct answers to find the portion of answer fields that were correct. All fields were treated as equally important to the score. The scores were then scaled to a 0–100 range.

F.5.6 Moral Reasoning

Moral Reasoning presents participants with ethical dilemmas and asks them to make judgments or decisions. This task tests players' ability to analyze complex situations, consider multiple perspectives, and make value-based judgments.

The high-level objective of this task is to assess ethical decision-making, perspective-taking abilities, and the capacity to weigh competing moral principles.

Participants had **5 minutes** to complete the task.



I met my wife in my early 30s. She was in her mid 20s at the time. I was already well established and had made a significant amount of money working as a quant and investing my savings diligently. She on the other hand had never gone to college and was content working for basically minimum wage.

So when I meet her we click instantly and I fall head over heels for her. We get married and move in together, and soon after have three kids. When we were expecting our first born she had told me she wanted to be a stay at home mom. Obviously I was jealous but I went along with it since it seemed logical.

So she spends the next 20+ years as a stay at home mom taking care of our kids while I work 12 hour days and then some. I'm not demeaning stay at home moms by the way, but in the context of our marriage she had an extremely easy going life style.

A couple years ago I discovered she had been cheating on me with her friend's husband. Divorced obviously followed.

The only major asset we had in the US was our house, which was a premarital asset. We had a shared bank account, but since I am a finance guy I handled all our investments, and for tax reasons invested it abroad in my home country.

Now if we had just grown apart, or she said she was unhappy, I'd have no problem paying what's fair. But the fact she cheated on me is what makes me resent paying her a cent.

So in the divorce she is awarded half of our investments made during marriage, > 90% of which are abroad in my name, along with alimony for 20 years for 30% of my pay.

So here I am given an opportunity to get out of this entirely by moving back to my home country and never looking back. So I sell our house as fast as possible, move the money abroad, and move back to my home country while stringing her along about why it's taking so long to liquidate our savings. Once abroad I immediately stop alimony payments.

Now she has no legal recourse (I've double checked with lawyers) and is left with her share of the small amount we had in the states

Think about the scenario. Select an answer that you think best describes it. Justify your answer in the text box. Click Approve Configuration to move on

○ YTA = You're the Asshole
○ NTA = Not the Asshole
● ESH = Everyone Sucks here
○ NAH = No Assholes here
○ INFO = Not Enough Info

```
It's not okay that they did that to eachother.
```

Figure 20. Participants in Moral Reasoning read a narrative and select a rating about the fault of the narrative's author. They then see a shared writing environment.

F.5.6.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will be given controversial cases and asked to evaluate who is responsible.

Goal: Satisfy conflicting interests before coming to one final decision about who is responsible.

Rounds: Will have different scenarios.

Instructions:

- *If in a team condition:* Your team works on the same scenario and will submit one shared response. All players must agree to move on.
- Consider various perspectives and satisfy conflicting interests before selecting an answer.
- In the text box, describe your thinking about each of the options.
- Since this is open-ended, there are no right or wrong answers.

Scoring Information:
Your score is determined by the degree to which you consider the balance of competing perspectives on the problem. This balance should be clearly stated in your response. This score will be calculated after the experiment ends.

Comprehension Quiz:

- What is the main activity of this game?
    - To evaluate who is responsible. [Correct Answer]
    - To write a response to a prompt.
    - To allocate funds to requests.
- Will this game have a score at the end of each round?
    - Yes, it is calculated at the end of each round. [Correct Answer]
    - No, it is calculated after the experiment ends.

F.5.6.2 Instances

Task complexity was modified by selecting situations from a corpus used in a prior experiment. A small pilot of games with selected situations was used to establish complexity rank — the situation leading to the most consistent responses across several individuals was considered the low complexity one, and the one leading to the least consistent responses was considered the high complexity one.

F.5.6.3 Scoring

Submissions were graded based on the reasoning they provided using a large language model-based pipeline, which would be given the submission justification and a submission rating in the context of the following prompt:

“You are a teacher grading a moral reasoning activity in which students rate who is responsible in a situation and explain their reasoning. The rating has the following options:

- YTA: You're the Asshole,
- NTA: Not the Asshole,
- ESH: Everyone Sucks Here,
- NAH: No Assholes Here,
- INFO: Not Enough Info.

You read the following situation and provide a grade between 0 and 100 based on how well the reasoning justifies the rating in detail. Reasoning that provides an analysis of alternative ratings is the best. If the reasoning is missing, the score is 0. If the reasoning is short, the score is 50.

Situation: {the task stimulus}

Reasoning: {the reasoning the participants provided}

Rating: {the rating the participants provided}

The numerical score:”

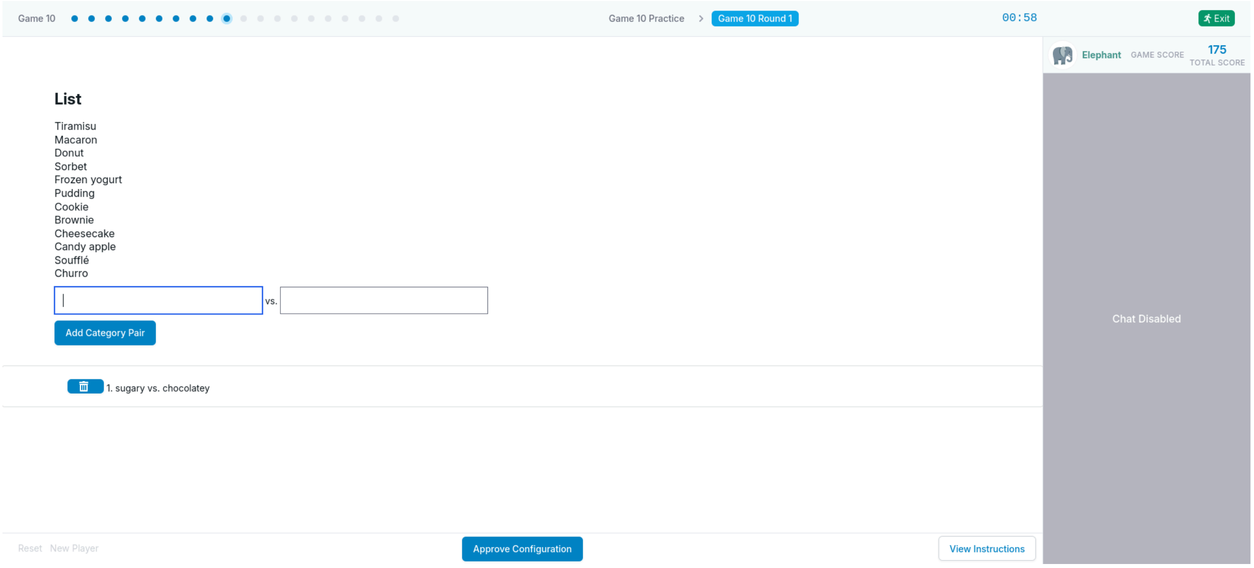
This was done using the **GPT3.5 completions API** with a temperature setting of 0 and a token limit of 1, so the model would respond with a repeatable number between 0 and 100.

F.5.7 Putting Food Into Categories

Putting Food Into Categories requires participants to sort various food items into appropriate categories. This task tests players' ability to classify objects based on common attributes and understand hierarchical relationships.

The high-level objective of this task is to assess categorical thinking, knowledge organization, and the ability to recognize patterns and relationships among different items.

Participants had **2 minutes** to complete the task.



List

Tiramisu
Macaron
Donut
Sorbet
Frozen yogurt
Pudding
Cookie
Brownie
Cheesecake
Candy apple
Soufflé
Churro

1. sugary vs. chocolatey

Figure 21. The Putting Food Into Categories interface. Participants see a list of food, alongside two text boxes to enter potential category names. The food items must be sorted into one of the two proposed categories (e.g., "Healthy" versus "Unhealthy").

F.5.7.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will create classifications for a list of foods.

Goal: Generate as many different classifications as possible.

Rounds: Each round will contain different types of food items.

Instructions:

- Your team will work on the same list and will submit one shared response. (*If in an individual condition:* You can click "Approve Configuration" to move on.)
- Enter as many unique classifications as possible within the time limit.
- A classification contains two characteristics that divide the items in the list.
- All items should fall into one of the two characteristics.

Scoring Information:
The score is determined by the number of unique classifications. More classifications score more points. Classifications that are repeated will not count. The score will be calculated after the game ends.

Comprehension Quiz:

- What is the main activity of this game?
    - To describe every food item in a list.
    - To categorize food items in a list to divide the items. [Correct Answer]
    - To paint every food item in a list in their natural color.
- Will this game have a score at the end of each round?
    - Yes, it is calculated at the end of each round.
    - No, it is calculated after the experiment ends. [Correct Answer]

F.5.7.2 Instances

Instances were drawn from prior literature in which this task had been used, and complexity rank was established in a pre-test evaluating qualitative difficulty for each of the three groups of foods, in order of increasing complexity: fruits, vegetables, and desserts. 11 items of each type were used as stimuli for each instance.

F.5.7.3 Scoring

Submissions were scored based on the validity of each category, measured by a large language model-based pipeline. The language model was provided with the list of items and the names of each of the pairs of categories the participants had produced. The model was then asked to assign items to each category using the following prompt architecture:

System prompt: Respond with JSON, where each item the user provided is placed in one of two possible categories. The categories are *category 1* or *category 2*.

User prompt: All items: *item 1, item 2, ...*.

category 1: Items from all items that would be in the category *category 1* as opposed to *category 2*.

category 2: Items from all items that would be in the category *category 2* as opposed to *category 1*.

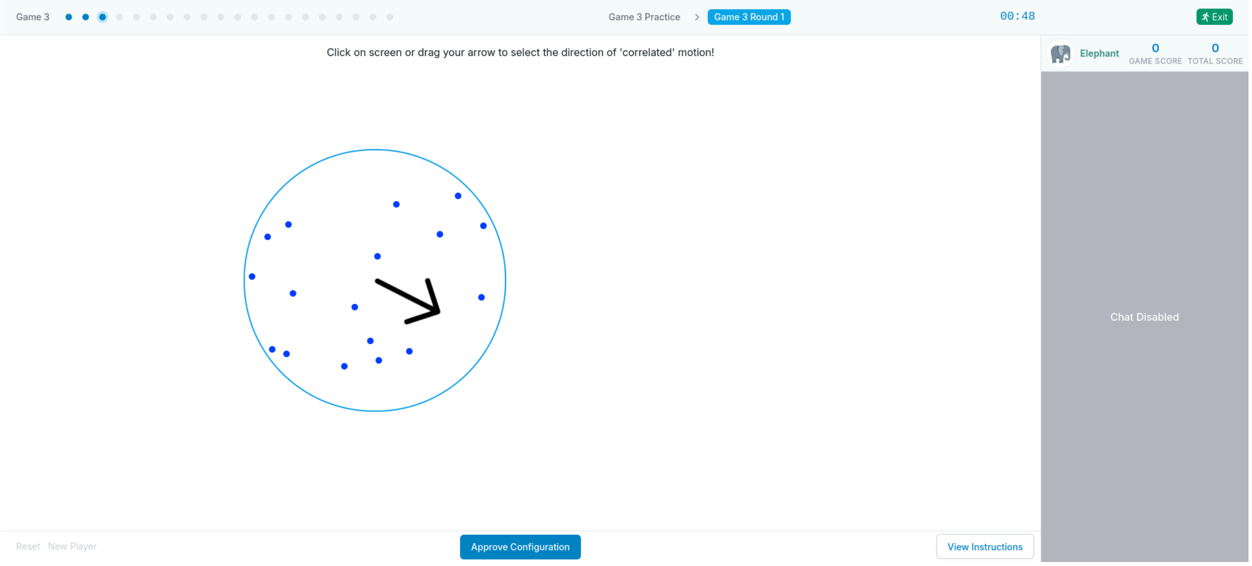
This was done using the **GPT4o chat API** with a temperature setting of 0 and configuration to receive a valid JSON Object as the response. The score was then computed by checking how many of the pairs of unique categories had at least one item in each category. If either category had no members, the category pair was not given a point. If any item could not be assigned to one of the two categories, the category pair was not given a point. Points were scaled to scores on a 0–100 range.

F.5.8 Random Dot Motion

Random Dot Motion requires participants to determine the overall direction of motion in a field of moving dots. This task tests players' ability to perceive coherent motion amidst visual noise.

The high-level objective of this task is to assess perceptual decision-making, the ability to integrate noisy sensory information, and visual attention.

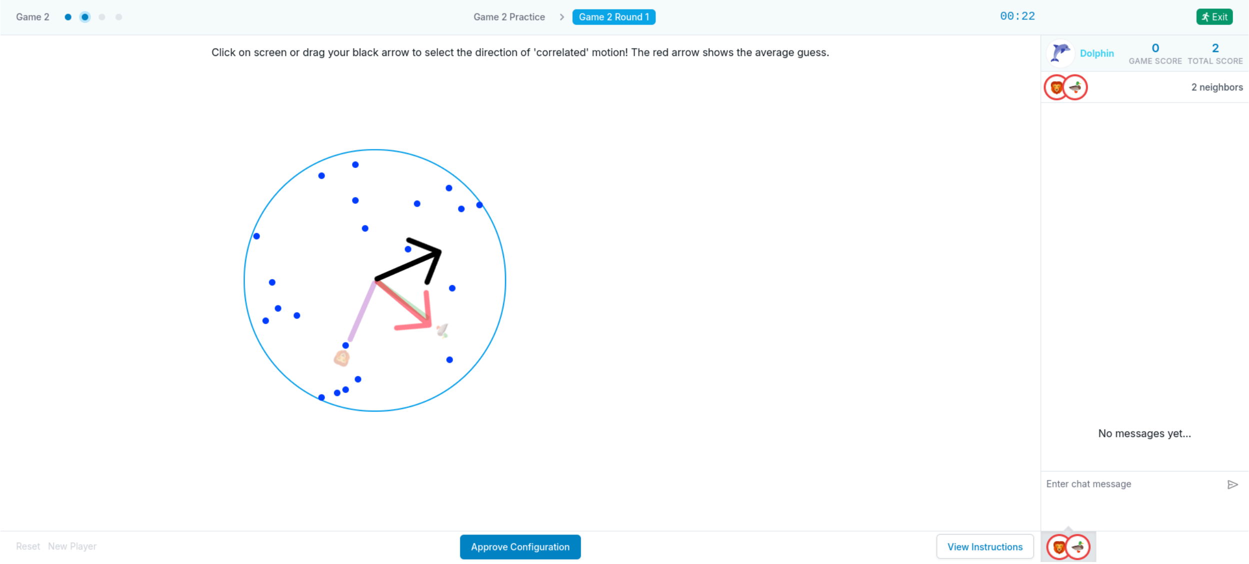
Participants had **1 minute** to complete the task.



Click on screen or drag your arrow to select the direction of 'correlated' motion!

Chat Disabled

Figure 22. Playing Random Dot Motion as an individual.



Click on screen or drag your black arrow to select the direction of 'correlated' motion! The red arrow shows the average guess.

No messages yet...

Figure 23. Playing Random Dot Motion as a group. Groups can see the direction of one anothers' guesses and adjust on the basis of social information.

F.5.8.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will determine the direction of the motion of the dots.

Goal: Accurately select the direction of the majority of dots, using the arrow.

Rounds: In each round, dots will move in different directions.

Instructions:

- Drag your arrow in the same direction, or click in the same direction, as to the motion of the majority of dots.
- Each member of the team has their own arrow. The blue arrow points towards the average of all teammates' arrows. The direction of the blue arrow is the only arrow that is scored. (*If in an individual condition:* You can click “Approve Configuration” to move on.)

Scoring Information:
Your score is determined by how accurately the arrow points in the direction that the majority of dots move. The closer your arrow is to the correct direction, the higher the score. The score will be calculated after the end of each round.

Comprehension Quiz:

- What is the main activity of this game?
    - To draw correlated dots.
    - To direct every dot in the board to the same motion.
    - To determine the motion of the majority of dots. [Correct Answer]
- Will this game have a score at the end of each round?
    - Yes, it is calculated at the end of each round. [Correct Answer]
    - No, it is calculated after the experiment ends.

F.5.8.2 Instances

Instances were created by parameterizing the movement of dots to adjust the direction of the dots and the level of correlation of the dots moving the dominant direction. Low complexity used a correlation of .5, medium complexity used a correlation of .3, and high complexity used a correlation of .05.

F.5.8.3 Scoring

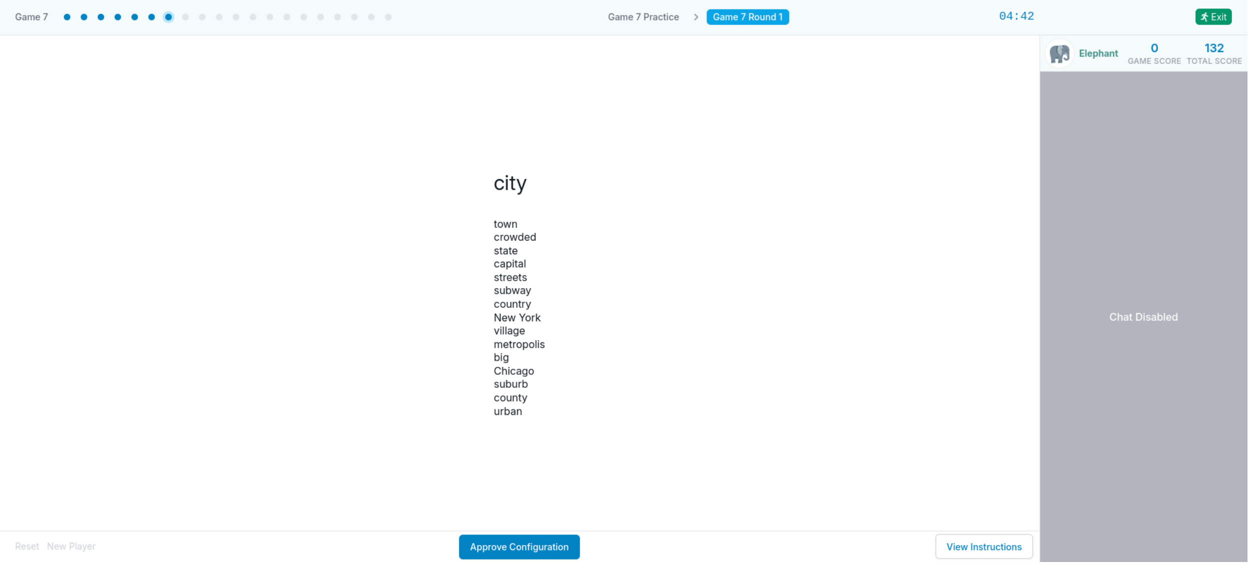
Submissions were scored by averaging the participants’ responses and computing the absolute difference between the known direction of the dots and the direction indicated by the response. The 0-180° range of possible scores was rescaled to 0–100.

F.5.9 Recall Association

Recall Association is designed to measure participants' ability to form and recall associations between paired items. Players are presented with groups of words, which they must memorize and then recall after a distractor task.

The high-level objective of this task is to assess cognitive abilities related to associative memory and resistance to interference.

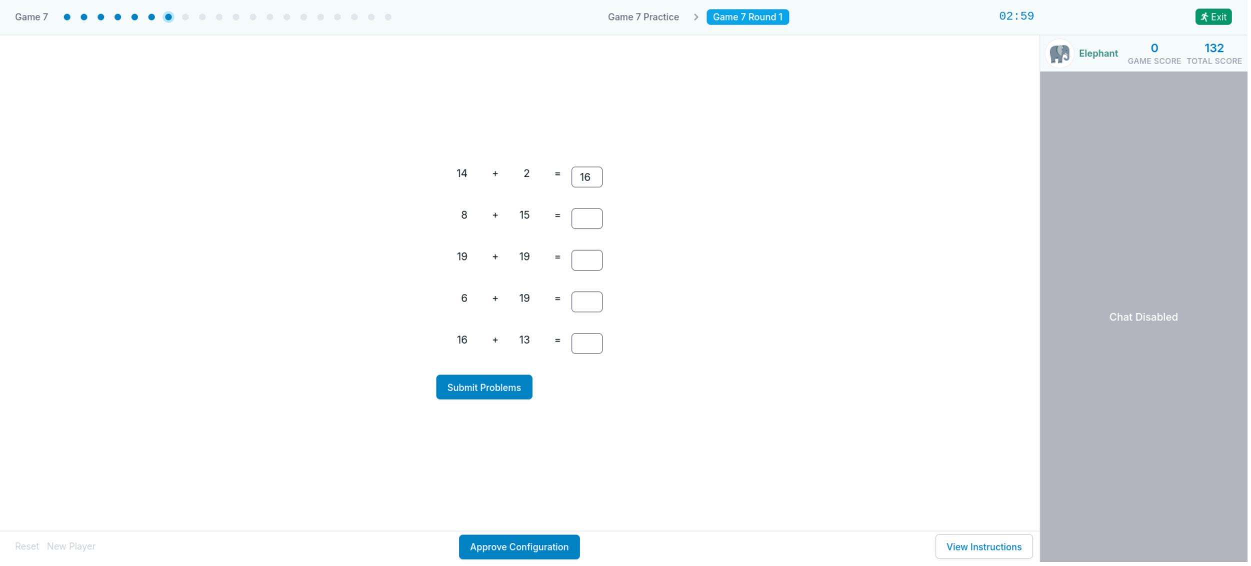
Participants had **5 minutes** to complete the task.



city

town
crowded
state
capital
streets
subway
country
New York
village
metropolis
big
Chicago
suburb
county
urban

Figure 24. Starting list of words for Recall Association.



14 + 2 = 16

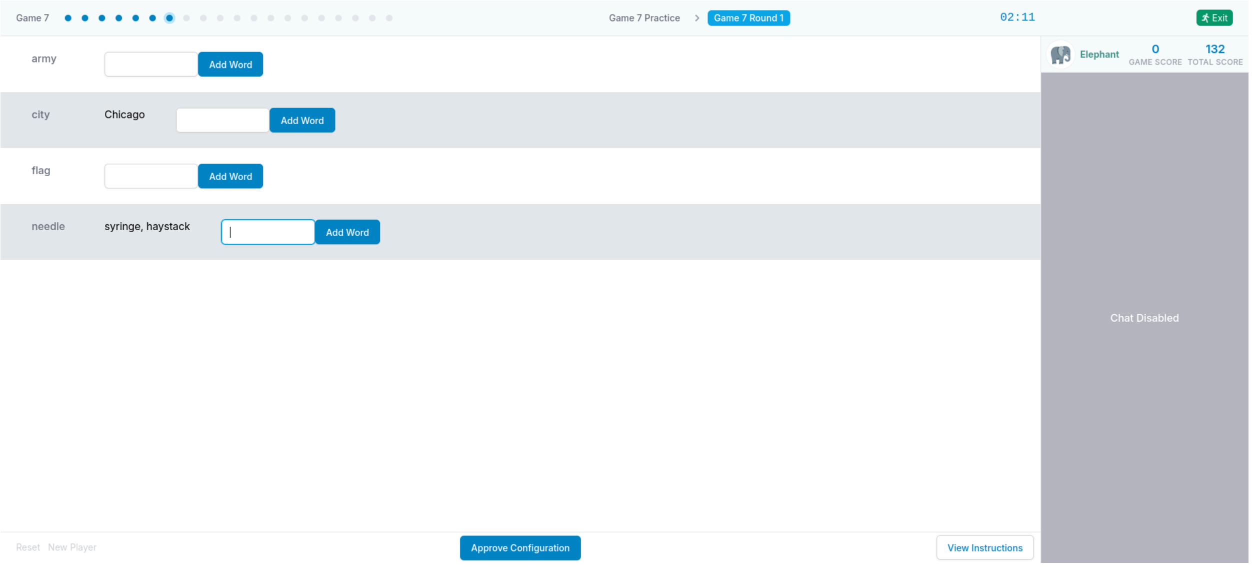
8 + 15 =

19 + 19 =

6 + 19 =

16 + 13 =

Figure 25. Distractor task for Recall Association.



army

city Chicago

flag

needle syringe, haystack

Figure 26. Interface for entering the recalled words.

F.5.9.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will be shown several lists of words that are grouped by a theme, then recall as many words as possible.

Goal: Accurately recall as many words as possible with minimal error.

Rounds: Each round will have different lists of words.

Instructions:

- Read each list of words so you can remember them. Each list will only be displayed once for a short time.
- After you see all the lists, you will have several math problems to complete.
- After completing the problems, recall as many words as possible and enter them under the correct theme.
- You may not type on your keyboard or switch tabs while the lists are being displayed. Doing so will lead to suspension from that round of the game, however, you may participate in the following rounds.
- Your team works on the same lists and will submit one shared response. (*If in an individual condition:* You can click "Approve Configuration" to move on.)

Scoring Information:
Your score is determined by the total number of words you recall correctly. A max score requires you to recall all words correctly. Your score will be calculated at the end of each round.

Comprehension Quiz:

- What is the main activity of this game?

    - To recall displayed words. [Correct Answer]
    - To recall synonyms.
    - To write things in a given category.
- Will this game have a score at the end of each round?
    - No, it is calculated after the experiment ends.
    - Yes, it is calculated at the end of each round. [Correct Answer]

F.5.9.2 Instances

Instances were drawn from existing literature, and complexity was associated with the number of lists participants would work through. Participants were shown 2, 4, and 6 lists, respectively, for low, medium, and high complexity levels.

F.5.9.3 Scoring

Submissions received a point for each word they correctly recalled for the appropriate list using a regular expression to handle differences in writing, such as capitalization or spaces before or after words. Scores reflect the percentage of the total possible points a submission earned.

F.5.10 Recall Word Lists

Recall Word Lists is designed to measure participants' short-term memory capacity. Players are presented with a list of words, which they must memorize and then recall after a brief delay.

The high-level objective of this task is to assess cognitive abilities related to memory retention and recall under time pressure.

Participants had **5 minutes** to complete the task.

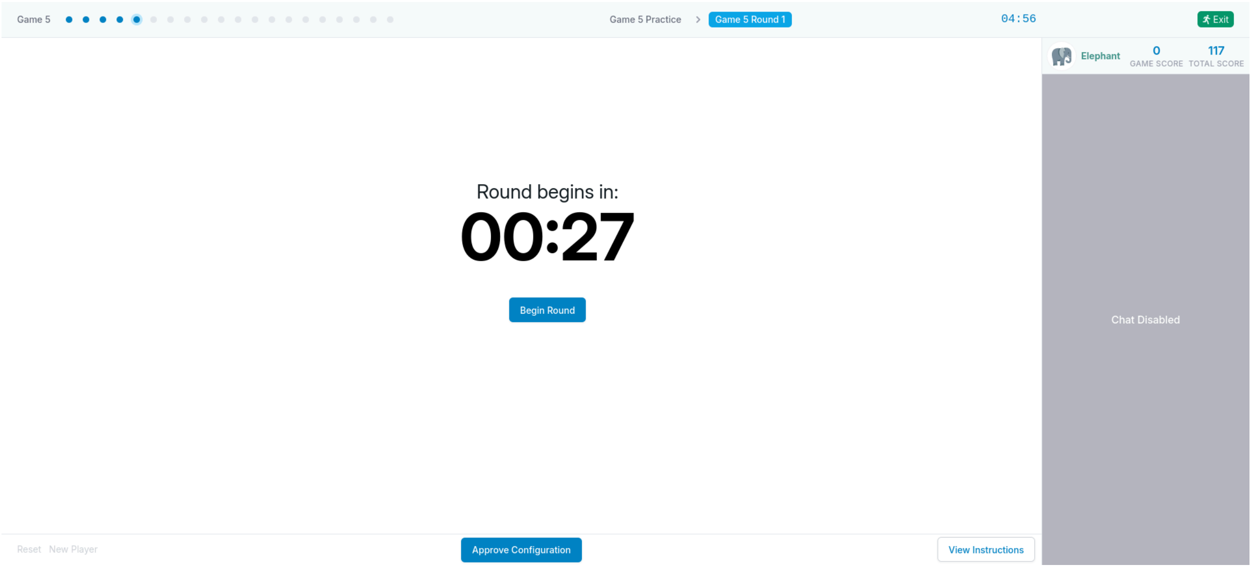
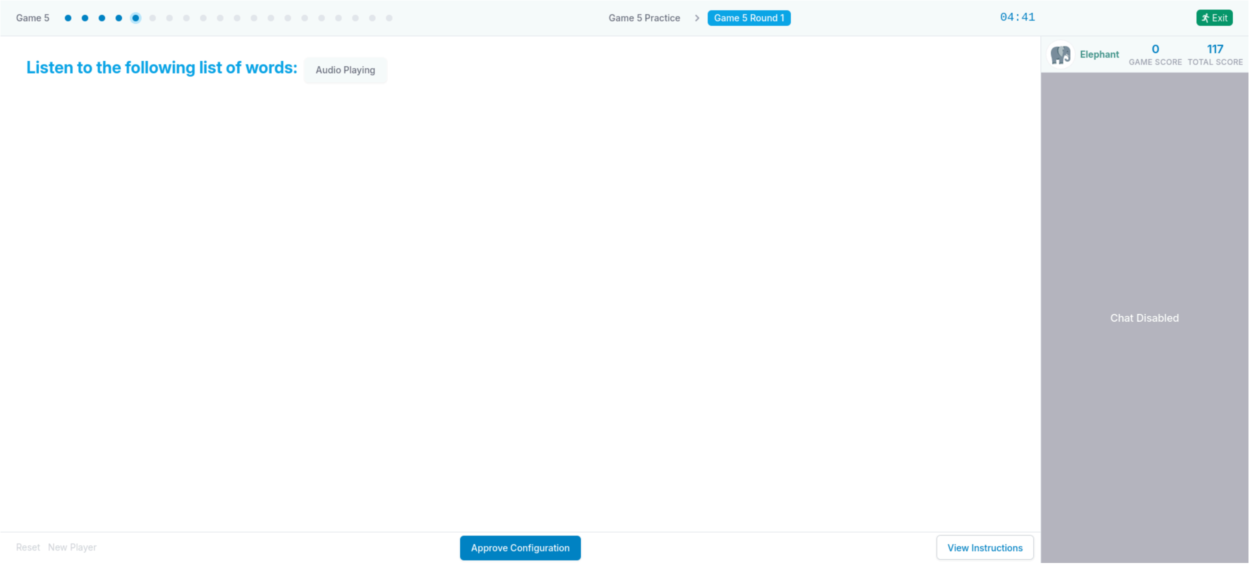
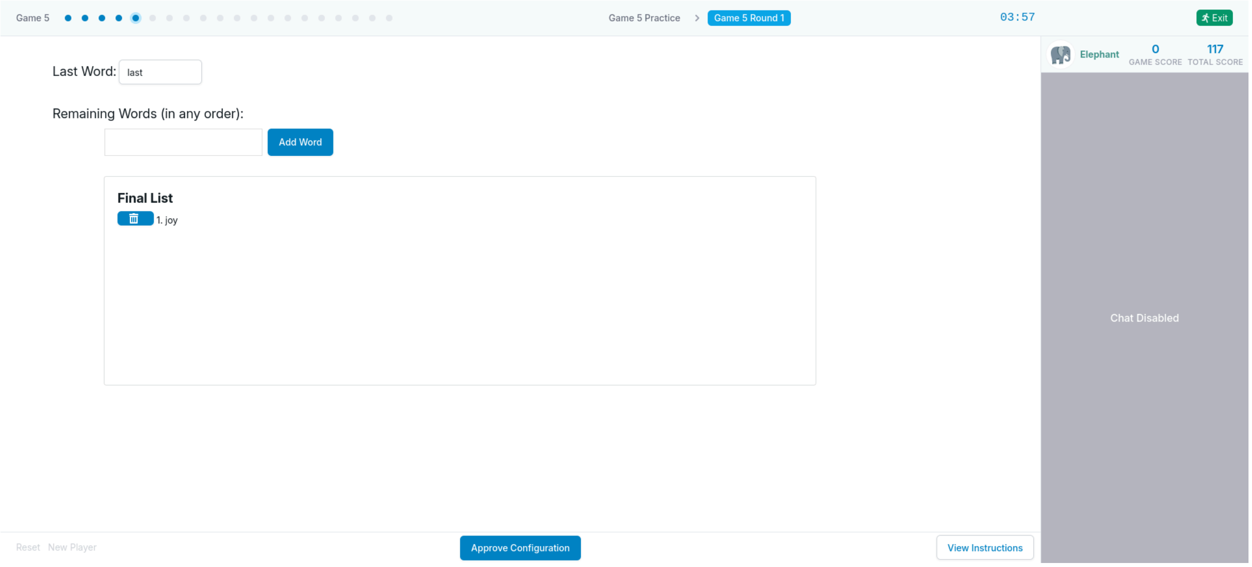


Figure 27. Starting interface for Recall Word Lists.



Listen to the following list of words: Audio Playing

Figure 28. Interface for Recall Word Lists while the participant is listening to the words.



Last Word: last

Remaining Words (in any order):

Final List

1. joy

Figure 29. Interface for Recall Word Lists while the participant is entering the final list of words.

F.5.10.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will hear a list of words and will then be asked to recall as many of them as possible.

Goal: After hearing the list, recall as many of the words as possible.

Rounds: Will contain different words.

Instructions:

- *If in a team condition:* Your team works on the same response and will submit one shared response. All players can agree to move on.
- The audio will play shortly after the round starts. Please make sure your volume is turned up.
- If you are typing while the audio is playing, **you will not be able to participate in the round**.
- If you change your tab while the audio is playing, **you will not be able to participate in the round**.
- Type the last word shared in the audio in the box labelled “Last Word.” This word MUST be correct to score any points.
- Enter the other words in any order in the list.

Scoring Information:
Your score is determined by the number of words that are correctly typed. Typed words must be spelled correctly in order to earn any points. Any incorrect words will cause points to be deducted from your total score. The “last word” must be correct or you will score 0 points regardless of how many other words you recalled correctly. This score will be calculated and displayed to you during the experiment.

Comprehension Quiz:

- What is the main activity in this game?
    - To create a list of any words.
    - To create a list of animals.
    - To replicate the list of words stated in the audio [Correct Answer]
- Will this game have a score at the end of each round?
    - Yes, it is calculated at the end of each round. [Correct Answer]
    - No, it is calculated after the experiment ends.

F.5.10.2 Instances

Instances were selected based on word lists used in existing literature, and complexity levels were established based on the number of words to be recalled: 5, 7, and 10, respectively for low, medium and high complexity levels.

F.5.10.3 Scoring

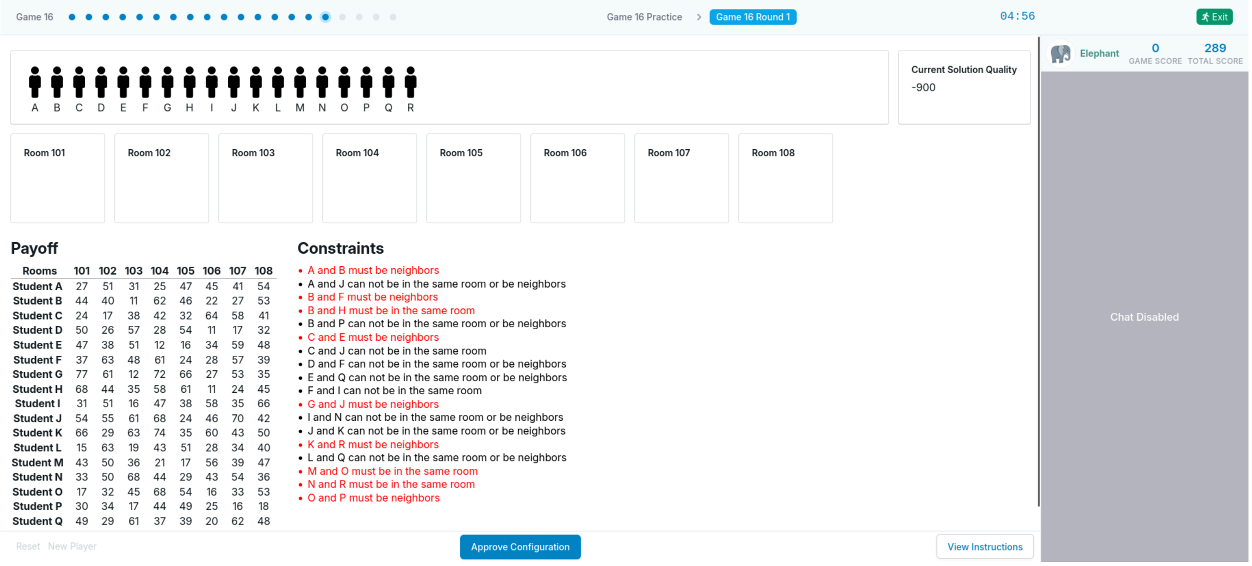
Submissions were scored on the basis of the percentage of the list of words that were correctly recalled. Participants were also asked to write in the last word in a text input field and if this was incorrect their score was 0, independent of the other words they recalled. Regular expressions were used to account for differences in input style, e.g. spacing or capitalization.

F.5.11 Room Assignment

Room Assignment is a constraint satisfaction problem where participants must assign people to rooms based on various preferences and constraints. This task tests players’ ability to optimize solutions while adhering to multiple rules.

The high-level objective of this task is to assess complex problem-solving skills, attention to detail, and the ability to balance multiple competing objectives.

Participants had **5 minutes** to complete the task.



Current Solution Quality
-900

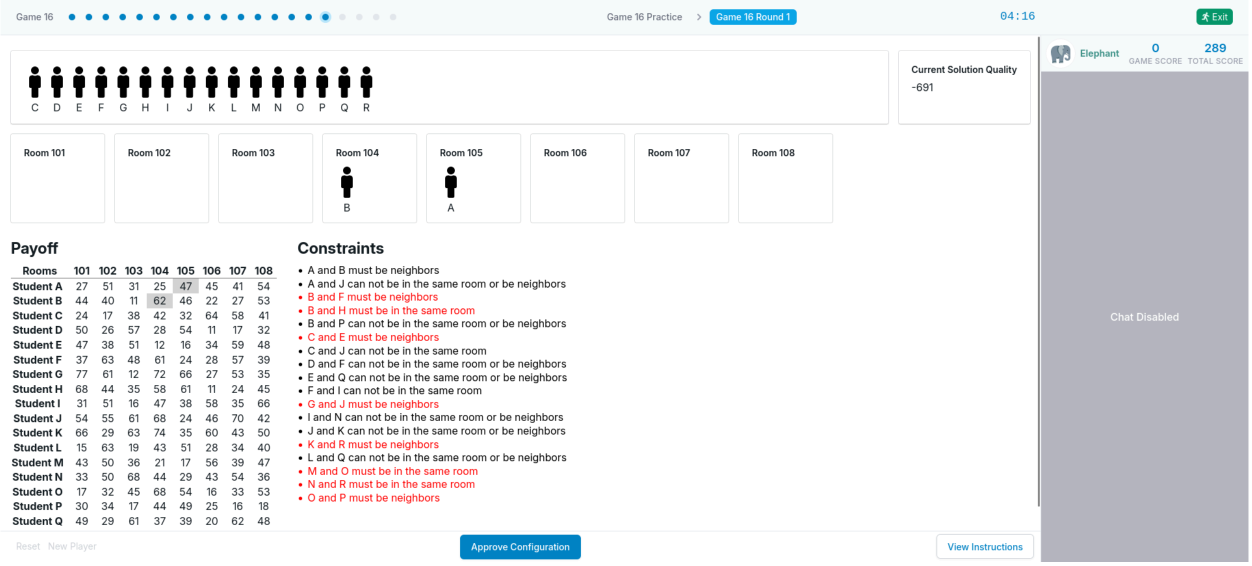
Payoff

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

Constraints

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

Figure 30. Starting interface for Room Assignment Task. Participants see a row of students to be allocated at the top of the screen, a row of rooms where the students can be allocated, a payoff chart for students and rooms (e.g., Student A gets a payoff of 51 from being in Room 102 but only 27 from being in Room 101), and a list of constraints (e.g., A and B must be neighbors).



Current Solution Quality
-691

Payoff

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

Constraints

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

Figure 31. The Room Assignment interface during gameplay.

F.5.11.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will create a room assignment plan by assigning a group of students into dorm rooms.

Goal: Create the room assignment plan that maximizes overall satisfaction for the group.

Rounds: Will have different numbers of students, rooms, or payoff calculations.

Instructions:

- *If in a team condition:* Your team works on the same room assignment plan and will submit one shared response. All players must agree to move on.
- Assign students to rooms by dragging and dropping them into the rooms.
- Note that for each student and room, there is a different payoff.
- The specific constraints of each round are listed once the round begins.
- Keep in mind some general constraints: some students can't live together or be neighbors, some students must live together or be neighbors.
- Rooms can be left empty.
- All students should be placed in a room.

Scoring Information:
Your score is determined by the payoff from the final room assignment plan. It will be normalized with the maximum payoff possible (final score = your payoff / max payoff). This score will be calculated during the experiment.

Comprehension Quiz:

- What is the main activity in this game?
    - To allocate people to rooms. [Correct Answer]
    - To allocate funds to requests.
    - To allocate animals to owners.
- Will this game have a score at the end of each round?
    - Yes, it is calculated at the end of each round. [Correct Answer]
    - No, it is calculated after the experiment ends.

F.5.11.2 Instances

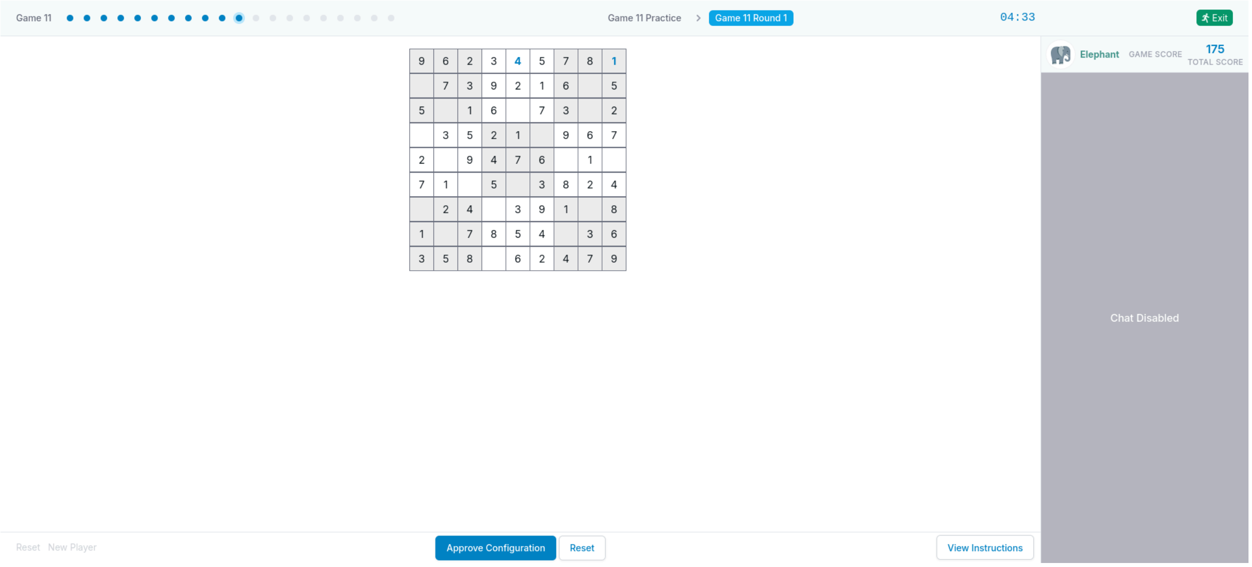
Task complexity is modified by adjusting the number of rooms, students, and constraints—the low complexity version of the task involved 6 students, 4 rooms and 2 constraints, while the medium involved 8 students, 5 rooms and 4 constraints and the high complexity version involved 18 students, 8 rooms and 18 constraints.

F.5.11.3 Scoring

The score comprises the sum of student payoffs for the rooms they were assigned and a penalty of 100 points for each constraint that was not satisfied.

F.5.12 Sudoku

Participants solved the classic 9x9 grid puzzle and had **5 minutes** to complete the task.



9	6	2	3	4	5	7	8	1
	7	3	9	2	1	6		5
5		1	6		7	3		2
	3	5	2	1		9	6	7
2		9	4	7	6		1	
7	1		5		3	8	2	4
	2	4		3	9	1		8
1		7	8	5	4		3	6
3	5	8		6	2	4	7	9

Figure 32. The Sudoku game interface.

F.5.12.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will complete puzzles by typing in numbers within the empty boxes.

Goal: Input the correct numbers, such that all constraints are met and there are no empty boxes remaining.

Rounds: Will contain different starting configurations and number of empty boxes.

Instructions::

- *If in a team condition:* Your team works on the same board and will submit one shared response. All players can agree to move on.
- The puzzle board is a 9x9 grid of boxes, where each row and column has 9 boxes. Additionally, the board is split up into nine 3x3 mini-grids. Each mini-grid is indicated by a shared grey or white background color.
- Fill all blank cells with a number to complete the puzzle. Each board should be completed so that:
  - Every row contains all the digits 1 through 9 only once.
  - Every column contains all the digits 1 through 9 only once.
  - Every 3x3 mini-grid contains all the digits 1 through 9 only once.

Scoring Information:
Your score is determined by the number of correct entries and wrong entries that you input in the board. Your score will be deducted if you input wrong numbers. This score will be calculated during the experiment.

Comprehension Quiz:

- What is the main activity of this game?

    - To allocate people to rooms.
    - To enter letters into boxes.
    - To enter numbers into boxes. [Correct Answer]
- Will this game have a score at the end of each round?
    - Yes, it is calculated at the end of each round. [Correct Answer]
    - No, it is calculated after the experiment ends.

F.5.12.2 Instances

Task complexity was influenced by the number of blank cells—20, 40, and 50 blanks for low, medium and high complexity levels respectively.

F.5.12.3 Scoring

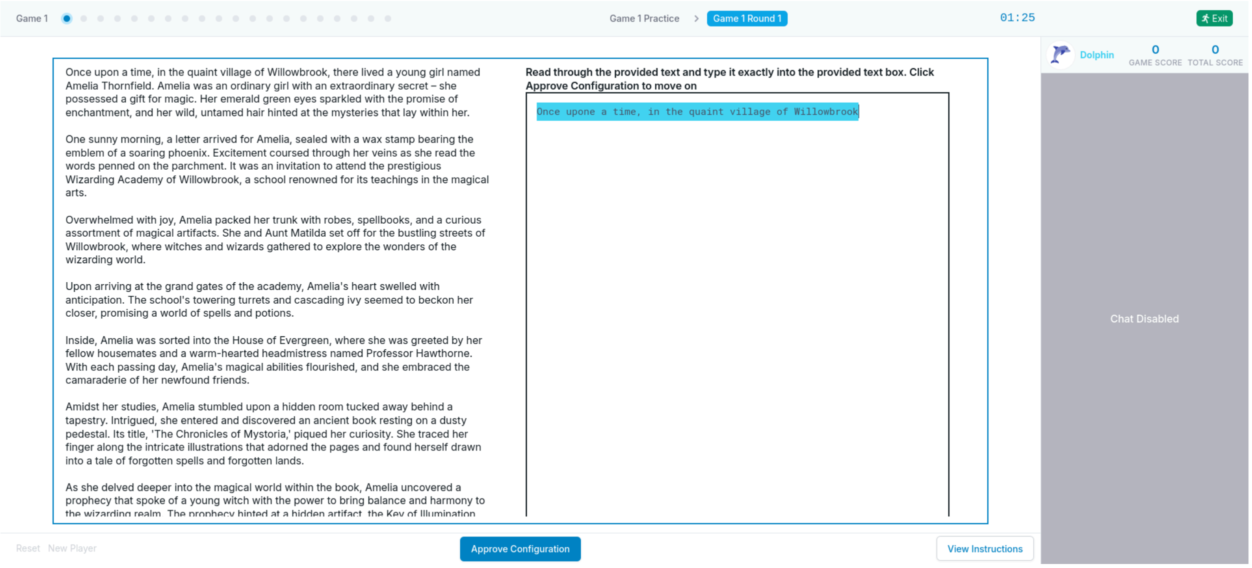
Participants received points depending on the number of blanks on the sudoku board. They received positive points for every correct number added and negative points for every incorrect number added. The default score is 50, and the value of all the blanks on the board is 50. So the lowest achievable score is 0 if every number they add is incorrect, and the highest achievable score is 100 in the case where every number is correct.

F.5.13 Typing Game

The typing game is designed to measure participants' typing speed and accuracy. Players are presented with a text passage and asked to type it as quickly and accurately as possible.

The high-level objective of this task is to assess basic motor skills and attention to detail in a simple, well-defined task.

Participants had **3 minutes** to complete the task.



Once upon a time, in the quaint village of Willowbrook, there lived a young girl named Amelia Thornfield. Amelia was an ordinary girl with an extraordinary secret – she possessed a gift for magic. Her emerald green eyes sparkled with the promise of enchantment, and her wild, untamed hair hinted at the mysteries that lay within her.

One sunny morning, a letter arrived for Amelia, sealed with a wax stamp bearing the emblem of a soaring phoenix. Excitement coursed through her veins as she read the words penned on the parchment. It was an invitation to attend the prestigious Wizarding Academy of Willowbrook, a school renowned for its teachings in the magical arts.

Overwhelmed with joy, Amelia packed her trunk with robes, spellbooks, and a curious assortment of magical artifacts. She and Aunt Matilda set off for the bustling streets of Willowbrook, where witches and wizards gathered to explore the wonders of the wizarding world.

Upon arriving at the grand gates of the academy, Amelia's heart swelled with anticipation. The school's towering turrets and cascading ivy seemed to beckon her closer, promising a world of spells and potions.

Inside, Amelia was sorted into the House of Evergreen, where she was greeted by her fellow housemates and a warm-hearted headmistress named Professor Hawthorne. With each passing day, Amelia's magical abilities flourished, and she embraced the camaraderie of her newfound friends.

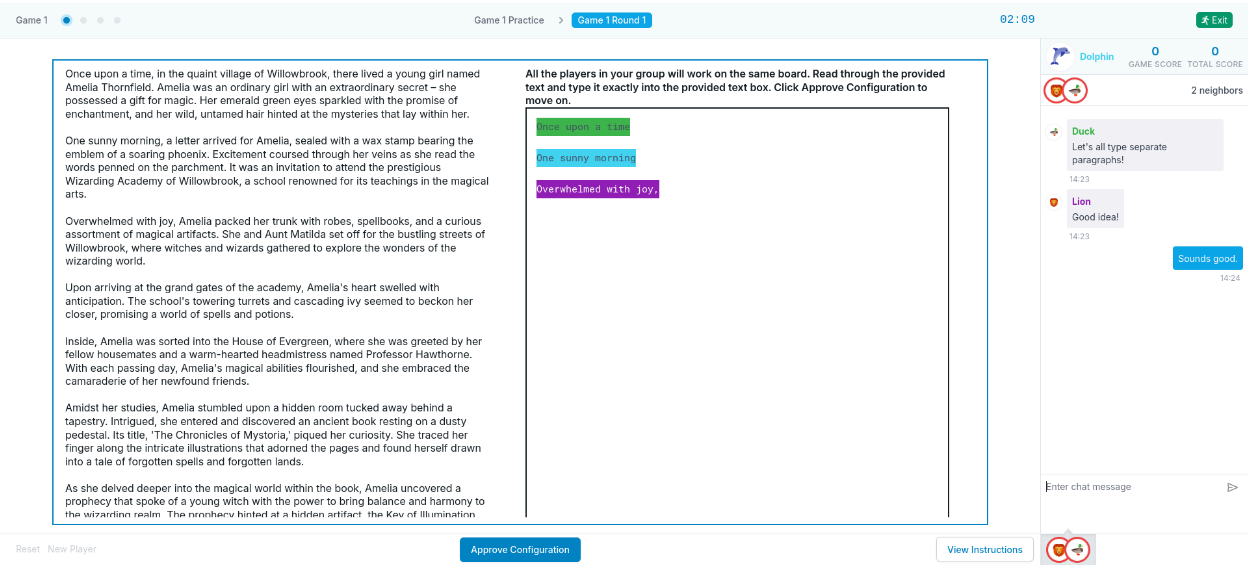
Amidst her studies, Amelia stumbled upon a hidden room tucked away behind a tapestry. Intrigued, she entered and discovered an ancient book resting on a dusty pedestal. Its title, 'The Chronicles of Mystoria,' piqued her curiosity. She traced her finger along the intricate illustrations that adorned the pages and found herself drawn into a tale of forgotten spells and forgotten lands.

As she delved deeper into the magical world within the book, Amelia uncovered a prophecy that spoke of a young witch with the power to bring balance and harmony to the wizarding realm. The prophecy hinted at a hidden artifact, the Key of Illumination

Read through the provided text and type it exactly into the provided text box. Click Approve Configuration to move on

```
Once upone a time, in the quaint village of Willowbrook
```

Figure 33. The Typing Game interface for individual players.



Once upon a time, in the quaint village of Willowbrook, there lived a young girl named Amelia Thornfield. Amelia was an ordinary girl with an extraordinary secret – she possessed a gift for magic. Her emerald green eyes sparkled with the promise of enchantment, and her wild, untamed hair hinted at the mysteries that lay within her.

One sunny morning, a letter arrived for Amelia, sealed with a wax stamp bearing the emblem of a soaring phoenix. Excitement coursed through her veins as she read the words penned on the parchment. It was an invitation to attend the prestigious Wizarding Academy of Willowbrook, a school renowned for its teachings in the magical arts.

Overwhelmed with joy, Amelia packed her trunk with robes, spellbooks, and a curious assortment of magical artifacts. She and Aunt Matilda set off for the bustling streets of Willowbrook, where witches and wizards gathered to explore the wonders of the wizarding world.

Upon arriving at the grand gates of the academy, Amelia's heart swelled with anticipation. The school's towering turrets and cascading ivy seemed to beckon her closer, promising a world of spells and potions.

Inside, Amelia was sorted into the House of Evergreen, where she was greeted by her fellow housemates and a warm-hearted headmistress named Professor Hawthorne. With each passing day, Amelia's magical abilities flourished, and she embraced the camaraderie of her newfound friends.

Amidst her studies, Amelia stumbled upon a hidden room tucked away behind a tapestry. Intrigued, she entered and discovered an ancient book resting on a dusty pedestal. Its title, 'The Chronicles of Mystoria,' piqued her curiosity. She traced her finger along the intricate illustrations that adorned the pages and found herself drawn into a tale of forgotten spells and forgotten lands.

As she delved deeper into the magical world within the book, Amelia uncovered a prophecy that spoke of a young witch with the power to bring balance and harmony to the wizarding realm. The prophecy hinted at a hidden artifact, the Key of Illumination

All the players in your group will work on the same board. Read through the provided text and type it exactly into the provided text box. Click Approve Configuration to move on.

```
Once upon a time

One sunny morning

Overwhelmed with joy,
```

Duck
Let's all type separate paragraphs!
14:23

Lion
Good idea!
14:23

Sounds good.
14:24

Figure 34. The Typing Game interface for groups. Different group members' contributions are highlighted in different colors.

F.5.13.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will retype a text.

Goal: The goal is to accurately reproduce the provided text by typing it in the text box.

Rounds: will have different bodies of text

Instructions:

- Be sure to avoid spelling mistakes and typos.
- Including extra spacing between words or paragraphs will not lower your score.
- Punctuations must be placed at precisely the same position as in the original text.
- Copy and paste have been disabled.
- *If in a team condition:* Your team works on the same document and will submit one shared response. All players can click "Approve Configuration" to move on.
- You can click "Approve Configuration" to move on.

Scoring Information:
Your score is determined by the accuracy with which you've managed to reproduce the text.

For each word correctly replicated, a point will be awarded; for every typo or missing word, a point will be deducted. It is, therefore, possible to receive a score of 0, even if part of your submission is correct.

Your score will be updated after the end of each round.

Comprehension Quiz:

- What is the main activity of this game?

    - To write an ending for the story in the provided text.
    - To replicate the provided text by typing. [Correct Answer]
    - To summarise the provided text.
- Will this game have a score at the end of each round?
    - Yes, it is calculated at the end of each round. [Correct Answer]
    - No, it is calculated after the experiment ends.

F.5.13.2 Instances

Instances were produced by asking a language model to write a story of about 500 words in length. Complexity assigned by pre-testing writing accuracy on other participants --- the lower the score, the higher the complexity level.

F.5.13.3 Scoring

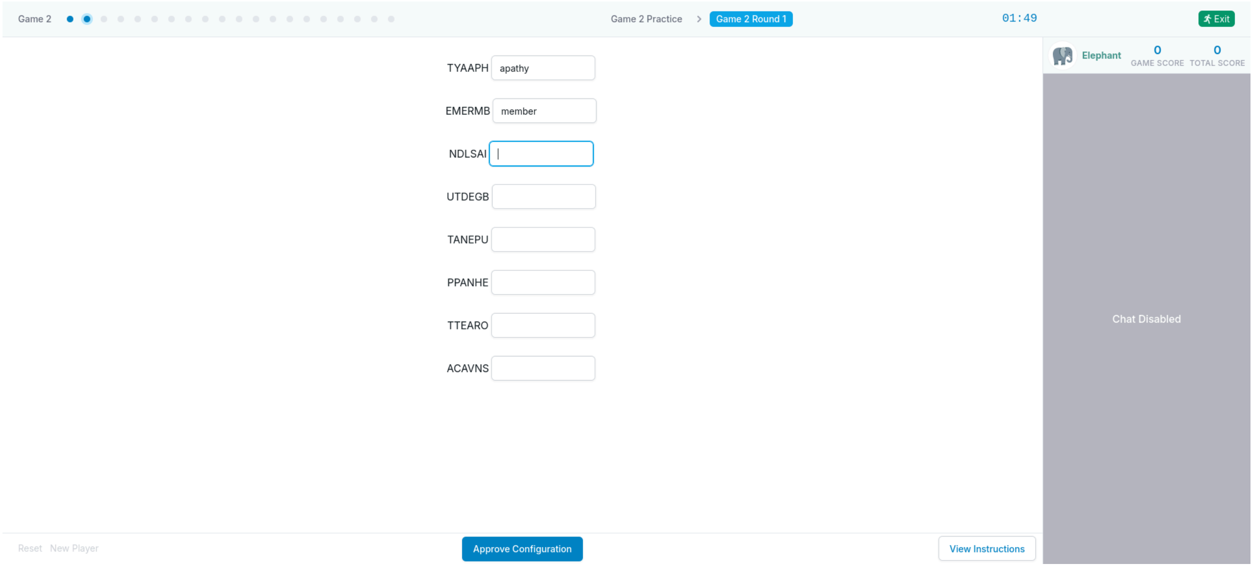
Submissions were scored by computing the edit difference between the submission and the true text using a standard difference algorithm which was scaled to a percentage.

F.5.14 Unscramble Words

Unscramble Words presents participants with scrambled letters and asks them to form valid words. This task tests players' vocabulary and ability to recognize word patterns.

The high-level objective of this task is to assess language processing skills, pattern recognition in linguistic contexts, and vocabulary depth.

Participants had **2 minutes** to complete the task.



TYAAPH apathy

EMERMB member

NDLSAI

UTDEGB

TANEPU

PPANHE

TTEARO

ACAVNS

Figure 35. The Unscramble Words interface. Participants see a list of anagrams, with a text box for each anagram where participants can enter the unscrambled word.

F.5.14.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will unscramble groups of letters and make words out of them.

Goal: Unscramble as many words from the list as possible.

Rounds: will contain different words and the difficulty of these words will vary per round.

Instructions:

- Each group of letters will spell one and only one English word.
- Use all letters in each group of letters to spell a word.
- Proper names (of people and places) are not used.
- Use the text box near the group of letters to type in the correct word.
- Answers are not case sensitive, so you can type in both lowercase and uppercase letters.
- *If in a team condition:* Your team will work on the same list and will submit one shared response. All players can click "Approve Configuration" to move on.
- You can click "Approve Configuration" to move on.

Scoring Information:
Your score is determined by the number of words you unscramble. You will be graded out of a 100 with a 100 being awarded if you unscramble all the words presented. Your score will be calculated and displayed at the end of each round.

Comprehension Quiz:

- What is the main activity of this game?
  - To enter letters in their correct order into boxes. [Correct Answer]
  - To enter number patterns into boxes.
  - To paint boxes with distinct colors.
- Will this game have a score at the end of each round?
  - Yes, it is calculated at the end of each round. [Correct Answer]
  - No, it is calculated after the experiment ends.

F.5.14.2 Instances

Task instances differed in the words' level of complexity. Low-complexity words included "LUNCH" and "DRESS;" Medium-complexity words included "APATHY" and "BUDGET," and high-complexity words included "PARADOX" and "CONFINE."

We present the full list of words for each level of complexity, with the scrambled version presented to participants shown in parentheses.

Low complexity word list

- ROUND (ONDRU)
- LUNCH (HCUNL)
- FAULT (AUFLT)
- DRESS (SERSD)
- PROUD (UPDRO)
- LEAVE (VEALE)
- HONOR (HORON)

- MOUTH (OUTMH)

Medium complexity word list

- APATHY (TYAAPH)
- MEMBER (EMERMB)
- ISLAND (NDLSAI)
- BUDGET (UTDEGB)
- PEANUT (TANEPU)
- HAPPEN (PPANHE)
- ROTATE (TTEARO)
- CANVAS (ACAVNS)

High complexity word list

- FREEDOM (EDMOFRE)
- PUDDING (UPGNIDD)
- PARADOX (DPXORAA)
- ABSENCE (BACENSE)
- CONFINE (ONFEINC)
- EXTREME (ETEXERM)
- SCIENCE (ESICNCE)
- CONVICT (TONCIVC)

F.5.14.3 Scoring

Scores were computed based on the number of words that participants were able to correctly unscramble. Participants received one point for every word from the answer list that appeared in their final submission, with the maximum answer being the length of the word list (that is, participants successfully unscrambled every word).

F.5.15 Whac-A-Mole

Whac-A-Mole is a reaction time game where participants must quickly respond to visual stimuli by “hitting” targets as they appear. This task tests players’ hand-eye coordination, attention, and processing speed.

The high-level objective of this task is to assess motor response speed, visual attention, and the ability to inhibit responses to non-target stimuli.

Participants had **2 minutes** to complete the task.

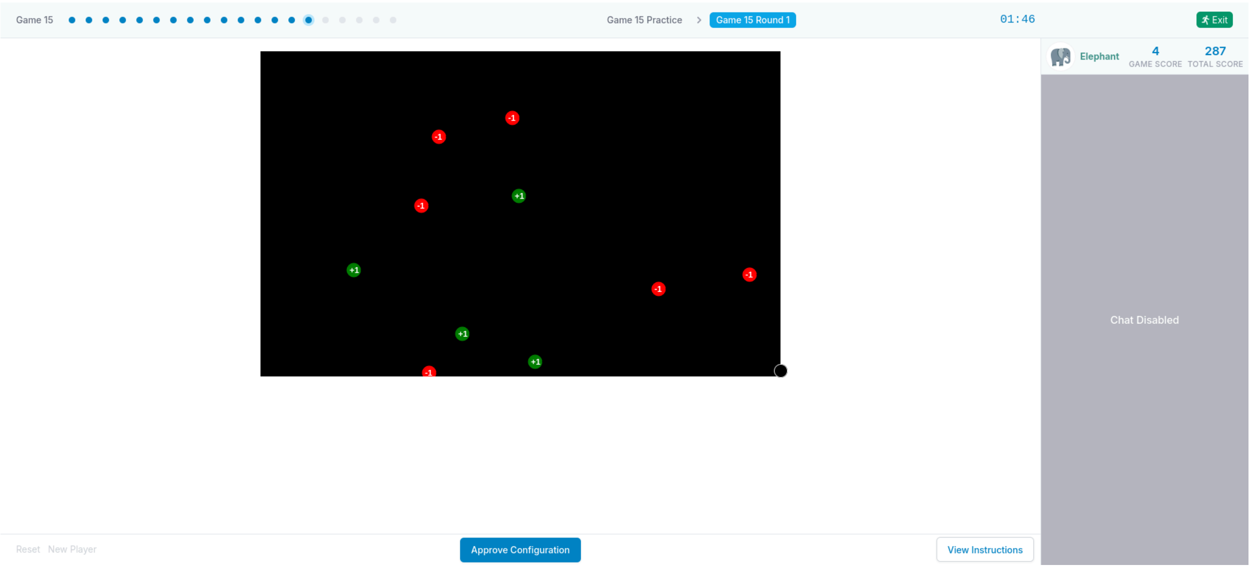


Figure 36. The Whac-A-Mole interface. Participants mouse over the moving dots on the screen to “whack” them. Different colored dots have either positive (green), negative (red), or neutral (blue; not pictured) point values.

F.5.15.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will move your cursor over dots that briefly appear and then disappear on the screen.

Goal: Get the highest score by moving your cursor over the dots (clicking is NOT required).

Rounds: Will have different types of dots or dots may appear and disappear faster/slower.

Instructions:

- *If in a team condition:* Your team works on the same board and will contribute to a shared score. All players can agree to move on.
- Your cursor is a circle on the board.
- Different color dots will appear on your screen for a short amount of time.
- Using your mouse, move your cursor over dots before they disappear.
- Note: Clicking the dots is NOT required, you simply have to move your cursor over the dots.
- Some dots add to your score, some dots subtract from it, and some dots have no effect. Dots will have a score inside (e.g., “+1” or “0” or “-1”).

Scoring Information:
Your score is determined by the dots you move your cursor over: green dots will increase your score, red dots will decrease your score, and blue dots will leave your score unchanged. It will be normalized with the maximum score possible (final score = your score / max score). This score will be calculated during the experiment.

Comprehension Quiz:

- What is the activity in this game?

    - ○ To guess a correlation
    - ○ To move objects around
    - ○ To move your cursor over dots [Correct Answer]
- Will this game have a score at the end of each round?
    - ○ Yes, it is calculated at the end of each round. [Correct Answer]
    - ○ No, it is calculated after the experiment ends.

F.5.15.2 Instances

Participants must click on (‘whack’) a series of green dot targets (‘moles’) that move through the screen. Clicking on each green dot target provided 1 point. Clicking on a red dot removed 1 point. Task complexity was adjusted by the types of dots included: low complexity had only green dots, medium complexity had green and red dots, high complexity had green and red as well as distractor blue dots (which provided 0 points). A random seed was used to generate consistent placement and randomly generated dots. The rate of dot appearance was also increased with task complexity — from nominally 1 per second, to nominally 2 per second, and more than 4 per second respectively.

F.5.15.3 Scoring

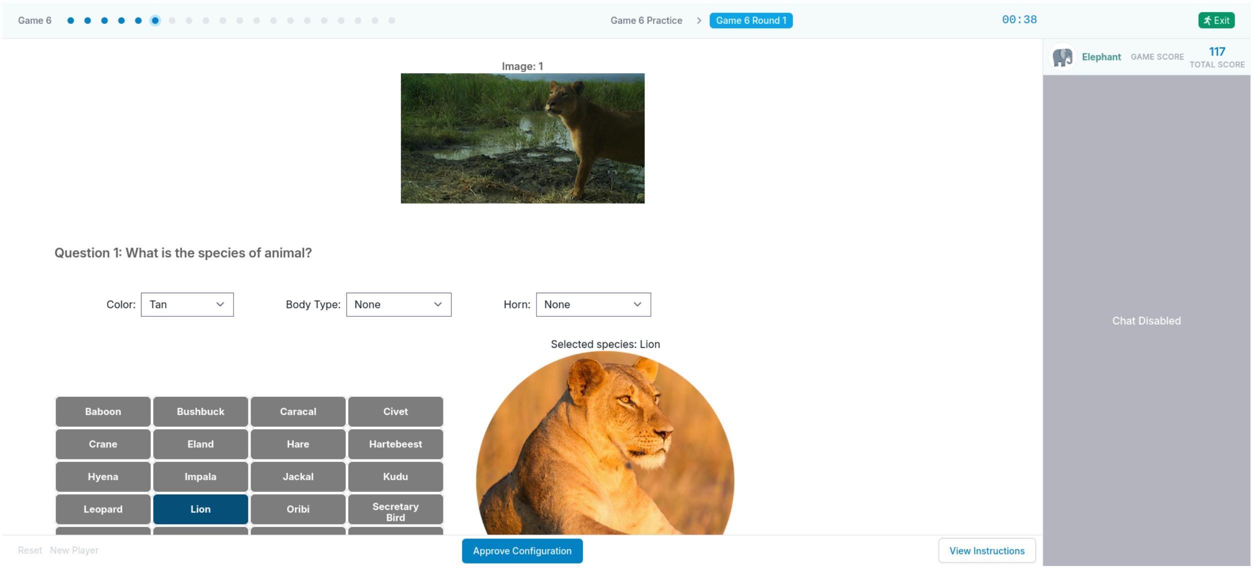
The ratio of the points earned and the maximum possible points per round, rescaled to a 0–100 range, were used as the score.

F.5.16 WildCam

WildCam involves identifying and classifying animals in wildlife camera trap images. This task tests players’ visual perception, pattern recognition, and knowledge of animal species.

The high-level objective of this task is to assess visual discrimination abilities, attention to detail, and the application of biological knowledge in real-world contexts.

Participants had **1 minute** to complete the task.



Game 6 Practice > Game 6 Round 1

00:38

Elephant GAME SCORE 117 TOTAL SCORE

Image: 1

Question 1: What is the species of animal?

Color: Tan

Body Type: None

Horn: None

Selected species: Lion

Figure 37. The WildCam interface.

F.5.16.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will be shown photos of wildlife and asked to annotate them by answering a series of questions about them.

Goal: Correctly answer all questions regarding the pictures.

Rounds: Each round will have different pictures, 1 picture in the practice round and 5 in all others.

Instructions:

- *If in a team condition:* Your team works on the same images and any player may answer any question. Your team will submit one shared response. All players must agree to move on.
- Take a good look at the picture before answering the questions.
- Once you answer all questions for an image, it will automatically progress to the next page. You cannot go back.

Scoring Information:
Your score is determined by the percentage of questions you answer correctly and will be computed at the end of each round.

Comprehension Quiz:

- What is the main activity of this game?
  - To annotate images of wild life. [Correct Answer]
  - To rate image quality.
  - To describe landscapes in images.
- Will this game have a score at the end of each round?
  - No, it is calculated after the experiment ends.
  - Yes, it is calculated at the end of each round. [Correct Answer]

F.5.16.2 Instances

Low, Medium, and High complexity task instances differed in the type, number, and behavior of animals observed. Below, we list the specific types of animals included in each level of complexity.

Low Complexity

- **Lion** (Number of animals: 1; Behavior: Standing; Young: No)
- **Baboon** (Number of animals: 2; Behavior: Moving; Young: No)
- **Buffalo** (Number of animals: 1; Behavior: Moving; Young: No)
- **Ground Hornbill** (Number of animals: 1; Behavior: Moving; Young: No)
- **Wild Dog** (Number of animals: 5; Behavior: Standing; Young: No)

Medium Complexity

- **Hyena** (Number of animals: 2; Behavior: Resting; Young: No)
- **Elephant** (Number of animals: 2; Behavior: Moving; Young: Yes)

- **Vulture** (Number of animals: 11-50; Behavior: Interacting; Young: Yes)
- **Warthog** (Number of animals: 4; Behavior: Moving; Young: Yes)
- **Baboon** (Number of animals: 2; Behavior: Interacting; Young: No)

High Complexity

- **Wildcat** (Number of animals: 1; Behavior: Standing; Young: No)
- **Hippopotamus** (Number of animals: 1; Behavior: Moving; Young: No)
- **Mongoose** (Number of animals: 2; Behavior: Moving; Young: No)
- **Lion** (Number of animals: 6; Behavior: Moving; Young: Yes)
- **Hyena** (Number of animals: 1; Behavior: Standing; Young: No)

F.5.16.3 Scoring

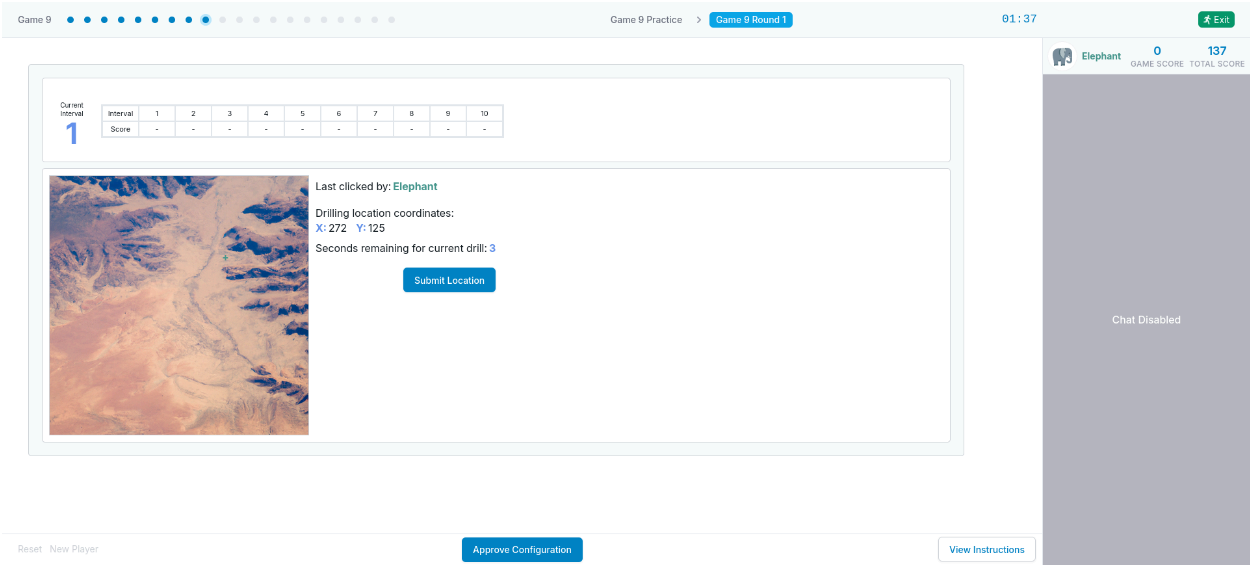
Participants were scored based on the number of correct classifications made, relative to the defined ground-truth information about the animals in each image. The maximum possible score involved getting all questions correct.

F.5.17 Wildcat Wells

Wildcat Wells simulates oil exploration, where participants must decide where to drill based on limited geological information. This task tests players' decision-making under uncertainty and risk assessment abilities.

The high-level objective of this task is to assess strategic decision-making, risk management, and the ability to interpret and act on probabilistic information.

Participants had **1 minute and 45 seconds** to complete the task.



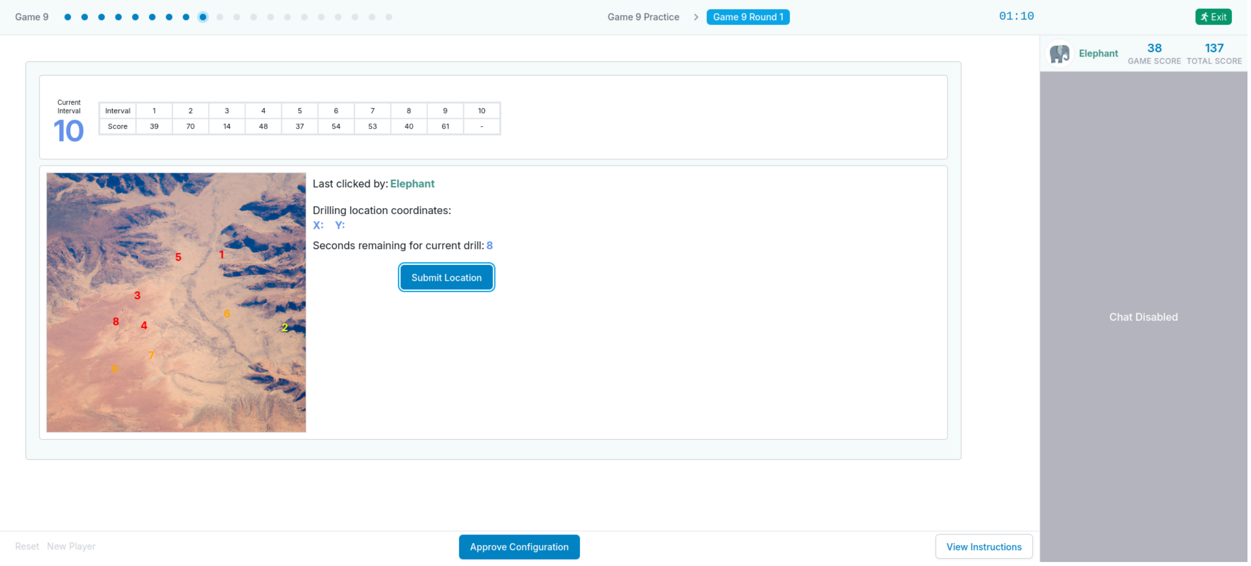
Interval	1	2	3	4	5	6	7	8	9	10
Score	-	-	-	-	-	-	-	-	-	-

Last clicked by: Elephant

Drilling location coordinates:
X: 272 Y: 125

Seconds remaining for current drill: 3

Figure 38. The starting interface for Wildcat Wells.



Interval	1	2	3	4	5	6	7	8	9	10
Score	39	70	14	48	37	54	53	40	61	-

Last clicked by: Elephant

Drilling location coordinates:
X: Y:

Seconds remaining for current drill: 8

Figure 39. The ending interface for Wildcat Wells. Participants see the locations where they have previously drilled, alongside their score totals.

F.5.17.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will be presented with a terrain map.

Goal: Obtain as much oil as possible by drilling at strategic locations on the map.

Rounds: will have different maps. Each round is divided into a series of ten 10-second intervals.

Instructions:

- During each 10-second interval, choose an oil well by selecting a location coordinate on the map. This point will be submitted at the end of the interval.
- Players must select a new point during every interval.
- *If in a team condition:* All members of your team will work together to select a coordinate for the oil well. This means that the point selected on the map was chosen by you or one of your teammates.
- Your team works on the same response and will submit one shared response. (*If in an individual condition:* You can click "Approve Configuration" to move on.)

Scoring Information:
Your score is determined by the amount of oil you discover. Thus, the goal is to try to find the "wells" with the most oil before time runs out. This score will be calculated at the end of each round.

Comprehension Quiz:

- What is the main activity in this game?
    - ○ Determine the shortest distance between points on a map.
    - ○ Select points on a map that represent wells with the most oil. [Correct Answer]

    - Complete problems to increase the amount of oil in the wells.
- Will this game have a score at the end of each round?
    - Yes, it is calculated at the end of each round. [Correct Answer]
    - No, it is calculated after the experiment ends.

F.5.17.2 Instances

Task complexity varied by manipulating the "roughness" of the payoffs in the oil field landscape. In the Low-complexity task, there is a single peak in the center; in the Medium-complexity task, there are a small number of varying sharp peaks that players have to discover; and in the High-complexity task, there are a greater number of sharp peaks and deep troughs, giving players a more complex search task in the landscape.

F.5.17.3 Scoring

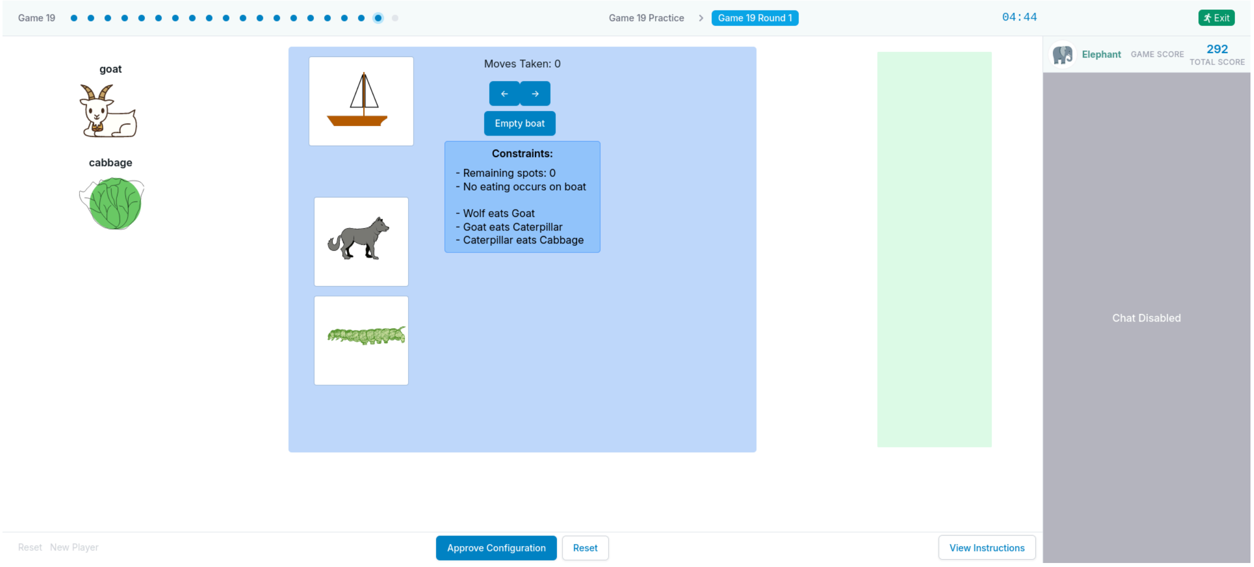
Games are scored based on the sum of the payoffs of the locations where participants chose to "drill" for oil.

F.5.18 Wolf, Goat, and Cabbage Problem

The Wolf, Goat, and Cabbage Problem is a classic river-crossing puzzle that tests participants' problem-solving and logical reasoning skills. Players must figure out how to transport a wolf, a goat, and a cabbage across a river without leaving incompatible pairs alone.

The high-level objective of this task is to assess strategic thinking, planning, and the ability to consider multiple constraints simultaneously. An important realization ("Eureka moment") crucial to success on this task is realizing that it is possible to bring things *back* across the river, rather than only bringing things in one direction.

Participants had **5 minutes** to complete the task.



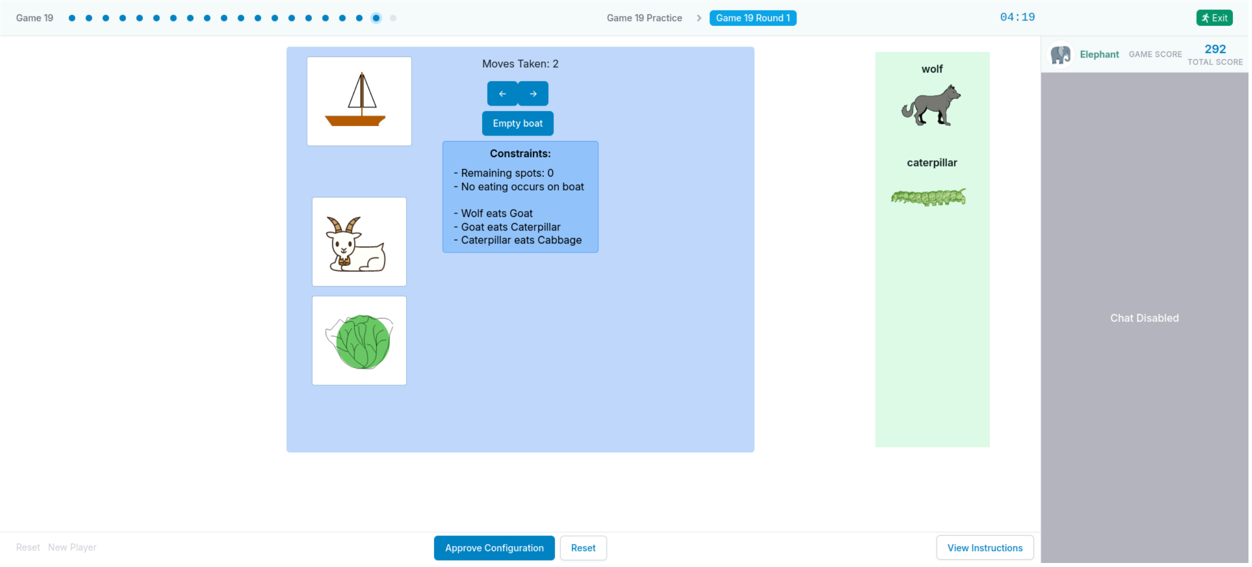
Moves Taken: 0

Constraints:
- Remaining spots: 0
- No eating occurs on boat

- Wolf eats Goat
- Goat eats Caterpillar
- Caterpillar eats Cabbage

Figure 40. The starting screen for the Wolf, Goat, and Cabbage Transfer problem. Participants are shown the animals that they need to transport across the river, alongside a list of constraints (e.g., "Wolf Eats Goat," "Goat Eats Cabbage"). Note that, to add

complexity, the boat had different capacity in different versions of the task, and there were different numbers of items to transport; here, the participant also had to transport 4 items: the Wolf, Goat, Cabbage, and Caterpillar.

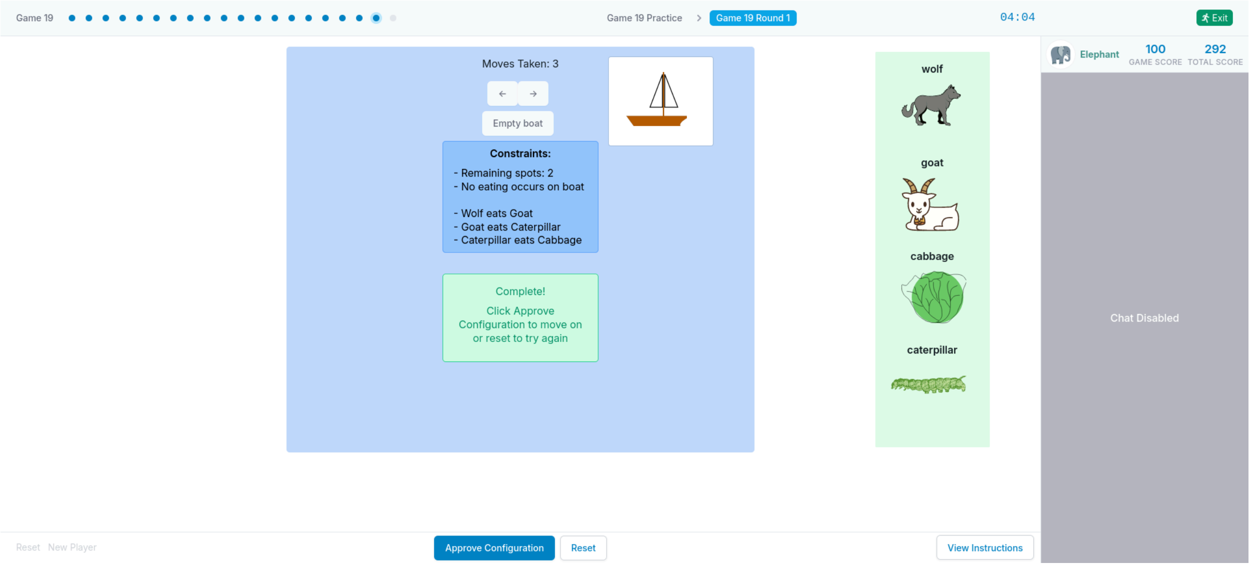


Moves Taken: 2

Constraints:
- Remaining spots: 0
- No eating occurs on boat

- Wolf eats Goat
- Goat eats Caterpillar
- Caterpillar eats Cabbage

Figure 41. The interface of the Wolf, Goat, and Cabbage Transfer problem during gameplay. Two of the items (Wolf and Caterpillar) have already been transported. The Goat and Cabbage are in the boat.



Moves Taken: 3

Constraints:
- Remaining spots: 2
- No eating occurs on boat

- Wolf eats Goat
- Goat eats Caterpillar
- Caterpillar eats Cabbage

Complete!
Click Approve Configuration to move on or reset to try again

Figure 42. The ending interface of the Wolf, Goat, and Cabbage Transfer problem. All four items are now on the other side of the river.

F.5.18.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will strategically move characters across a river, using a boat with a limited number of spots.

Goal: Move all characters from the left side to the right side in the fewest number of moves while respecting the constraints to avoid anyone eating anyone else.

Rounds: Will have different characters, constraints, or number of spots on the boat.

Instructions:

- *If in a team condition:* Your team works on the same board and will contribute to shared moves. All players can agree to move on.
- Place characters into the boat by dragging and dropping them over the boat.
- A move is counted only when you shift the boat across the river bank (e.g., from left to right or from right to left).
- Leaving two or more characters unattended when the constraints specify that one character will eat another will result in a failed attempt. This will require you to reset the board.
- Characters are considered "unattended" if the boat is on the opposite side of the river from them.
- You [*If in a team condition:* and your team] have an unlimited number of attempts and may reset the board at any point in the game.
- The constraints of each round are listed once the round begins. For example, a constraint may be "The Triceratops will eat the Grass." You will fail if you leave the Triceratops on the same side as the Grass.

Note: No eating ever occurs on the boat.

Scoring Information:
Your score is determined by the number of moves it takes you to get all characters to the right side. You must get all characters to the right to score any points. The more moves you take, the less points you will score. We will only keep your best solution each round. This score will be calculated during the experiment.

Comprehension Quiz:

- What is the main activity in this game?
  - To guess a correlation
  - To move characters from one side to another [Correct Answer]
  - To move your cursor over dots
- Will this game have a score at the end of each round?
  - Yes, it is calculated at the end of each round. [Correct Answer]
  - No, it is calculated after the experiment ends.

F.5.18.2 Instances

Task complexity was modified by adding items and adjusting the number of spaces on the boat — low complexity involved 4 items and a boat with one space for items; medium complexity involved 4 items and a boat with 2 spaces; and high complexity involved 5 items and a boat with 2 spaces.

F.5.18.3 Scoring

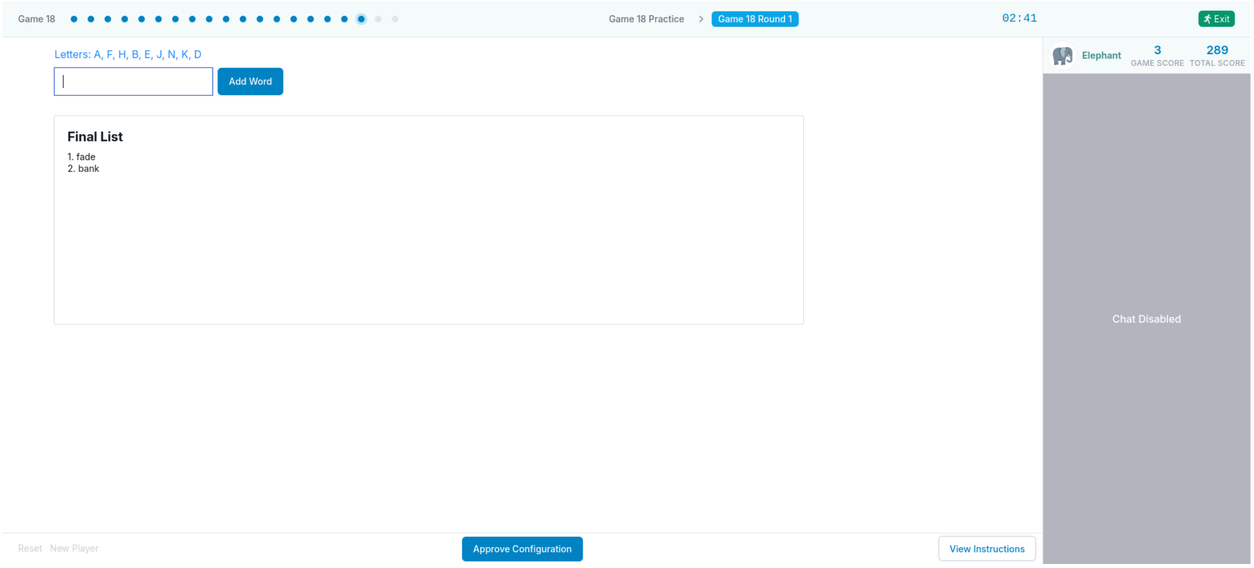
If the players didn't manage to move all the animals to the other shore, they would receive a score of 0. If they managed to reach the other side, their score would relate to the inverse of the number of steps they took, so the best score would be the smallest possible number of steps for a given puzzle. If players made more than one successful attempt, they received the best possible score. The score was scaled to span 0–100.

F.5.19 Word Construction

Word Construction requires participants to create words using a given set of letters. This task tests players’ vocabulary and word-forming abilities.

The high-level objective of this task is to assess linguistic creativity, vocabulary size, and the ability to manipulate letter combinations to form words.

Participants had **3 minutes** to complete the task.



Letters: A, F, H, B, E, J, N, K, D

Final List

1. fade
2. bank

Figure 43. The Word Construction task interface. Participants see a collection of letters (A, F, H, B, E, J, N, K, D) and a space to enter words formed using the letters.

F.5.19.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will come up with a list of words from a given set of letters.

Goal: Generate as many valid 4-letter English words as possible from the provided letters.

Rounds: Will have different sets of letters.

Instructions:

- *If in a team condition:* Your team works on the same list and will submit one shared response. All players must agree to move on.
- Each word:
  - Must be 4-letters in English
  - Cannot be repeated from earlier words

    - Note: Letters can be used more than once in a word

Scoring Information:
Your score is determined by the number of words you generate divided by the total number of words possible. It will be normalized with the maximum number of words possible (final score = your number / max number possible). This score will be calculated during the experiment.

Comprehension Quiz:

- What is the main activity in this game?
    - To create a list of words of any size.
    - To create a list of animals.
    - To create a list of words that are 4 letters. [Correct Answer]
- Will this game have a score at the end of each round?
    - Yes, it is calculated at the end of each round. [Correct Answer]
    - No, it is calculated after the experiment ends.

F.5.19.2 Instances

The size of the set of theoretically possible words varied across levels of complexity; the number of valid words that could be formed by each set decreased as complexity increased. The letters could be arranged to form 128 (low), 114 (moderate), and 82 (high) words, respectively.

F.5.19.3 Scoring

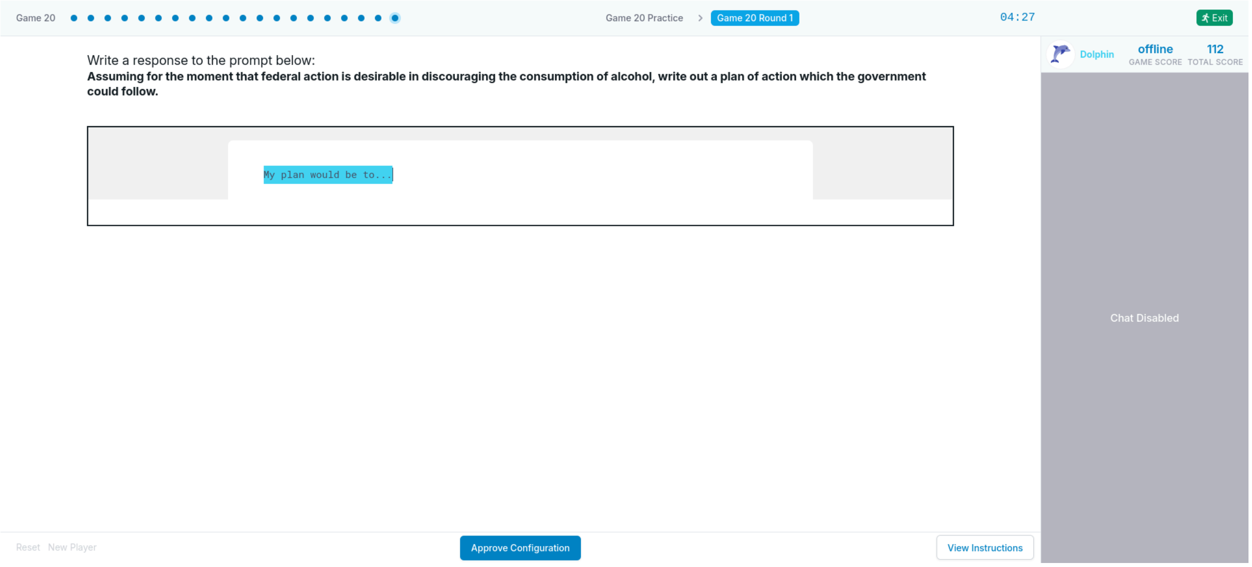
For each set of letters, the number of possible words were computed from the NLTK word list. The ratio of the number of unique submissions and the pre-computed number of possible words was used as a percentage score. For example, if there are 100 possible words that could be formed from the given set of words, and the participant generated 50 of them, then their raw (pre-scaled) score would be 50%.

F.5.20 Writing Story

Writing Story challenges participants to create a coherent narrative based on given prompts or elements. This task tests players' creative writing skills and ability to construct engaging stories.

The high-level objective of this task is to assess narrative creativity, language fluency, and the ability to structure coherent and engaging stories.

Participants had **5 minutes** to complete the task.



Write a response to the prompt below:
Assuming for the moment that federal action is desirable in discouraging the consumption of alcohol, write out a plan of action which the government could follow.

My plan would be to...

Figure 44. The Writing Story task interface. Participants see a prompt at the top of the screen and a space to write their story.

F.5.20.1 Instructions

Participants were presented with the following instructions:

Game Overview:
In this game, you will respond to a series of writing prompts.

Goal: Address the prompt with a thoughtful written response.

Rounds: Will have different prompts.

Instructions:

- *If in a team condition:* Your team works on the same response and will submit one shared response.
- Use the textbox to write your response.
- Since this is open-ended, there are no right or wrong answers.

Scoring Information:
Your score is determined by the quality of your response (including length, originality, quality of presentation, adequacy, and creativity). This score will be calculated after the experiment ends.

Comprehension Quiz:

- What is the main activity in this game?
  - To write a response to a prompt. [Correct Answer]
  - To make a list of words that are as different as possible.
  - To brainstorm as many words as possible.
- Will this game have a score at the end of each round?
  - Yes, it is calculated at the end of each round.
  - No, it is calculated after the experiment ends. [Correct Answer]

F.5.20.2 Instances

Task complexity is influenced by the topics in the prompt, for example, the low complexity prompt involved discouraging consuming alcohol, while the medium complexity prompt involved impacts of consumption on the environment, and the high complexity prompt involved impacts of technology on society. Pilot trials of the task with individuals were used to rank the prompts by average score.

The exact text of each prompt was as follows:

- Assuming for the moment that federal action is desirable in discouraging the consumption of alcohol, write out a plan of action which the government could follow.
- How does the growth in population affect the environment? Write a paragraph or two summarizing your opinions on this question.
- Does technology make us more alone/isolated? Write a paragraph or two summarizing your opinions on this question.

F.5.20.3 Scoring

Stories were graded by a large language model-based pipeline supplied with the story prompt and the submitted story. The following prompt was used:

“You are a teacher grading a writing activity. You read the following prompt and response and give it a grade between 0 and 100, based on how well it responds to the prompt and how good the writing is overall.

Prompt: {story prompt}

Response: {submitted response}

The numerical score:”

This was done using the **GPT3.5 completions API** with a temperature setting of 0 and a token limit of 1, so the model would respond with a repeatable number between 0 and 100.

Appendix G: A Large-Scale Integrative Experiment of Group Advantage – Analysis Details

G.1: Formal Definitions of Group Advantage

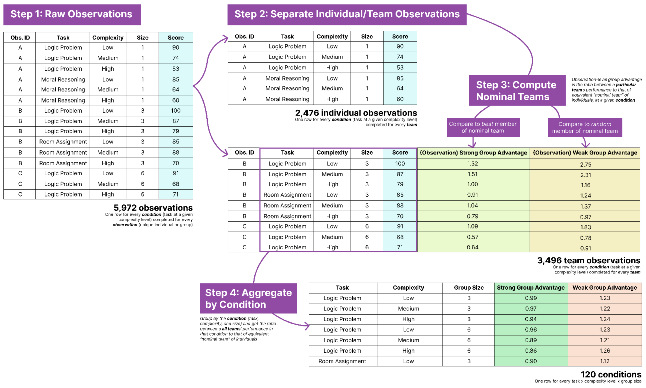
We define *group advantage* as the ratio of the performance of a team of size *n* to the aggregate performance of *n* individuals working alone (often known as a *nominal group*).

In our study, we operationalize a nominal group as a collection of individuals who had been assigned to the independent working condition, who are then randomly sampled (without replacement) to have their performance aggregated as if they had completed the task as a "group." Nominal groups are identical in size to interacting groups, consisting of three individuals (for small groups) or six individuals (for large groups). This comparison allows us to attribute any performance gains to group interaction rather than mere resource differences.

Furthermore, we compute two levels of group advantage—*weak* and *strong*—defining weak group advantage as an interacting group outperforming a randomly selected member of the corresponding nominal group, and strong group advantage as outperforming the best-performing member of the corresponding nominal group.

Finally, we analyze group advantage at two levels of analysis. Our primary level of analysis is that of a *condition* (*n* = 120, as we have 20 tasks × 3 complexity levels × 2 group sizes). Here, we examine the general extent to which all teams in a given condition (task × complexity × group size) tend to outperform equivalent nominal teams. Our secondary level of analysis is that of an *observation* (*n* = 3,496). Here, we examine the extent to which a specific team outperforms nominal teams.

In summary, we compute **strong group advantage** and **weak group advantage**, at the **observation-level** and the **condition-level** (Figure 45). We begin with 5,972 total observations, in which an observation is a single unit (individual or team) working to complete the assigned tasks. These observations are then divided into 2,476 individual observations and 3,496 team observations. The individual observations are used to create a baseline of "nominal teams" against which the team observations are compared. This comparison can take place at either the level of analysis of the observation (that is, did *this particular team* outperform an equivalent nominal team?) or at the level of analysis of a condition (that is, did *all teams* assigned to a particular condition outperform an equivalent nominal team?).



Step 1: Raw Observations

Obs. ID	Task	Complexity	Size	Score
A	Logic Problem	Low	1	90
A	Logic Problem	Medium	1	74
A	Logic Problem	High	1	53
A	Moral Reasoning	Low	1	85
A	Moral Reasoning	Medium	1	64
A	Moral Reasoning	High	1	60
B	Logic Problem	Low	3	100
B	Logic Problem	Medium	3	87
B	Logic Problem	High	3	79
B	Room Assignment	Low	3	85
B	Room Assignment	Medium	3	88
B	Room Assignment	High	3	70
C	Logic Problem	Low	6	91
C	Logic Problem	Medium	6	68
C	Logic Problem	High	6	71

5,972 observations
One row for every ***condition*** (task at a given complexity level) completed for every ***observation*** (unique individual or group)

Step 2: Separate Individual/Team Observations

Obs. ID	Task	Complexity	Size	Score
A	Logic Problem	Low	1	90
A	Logic Problem	Medium	1	74
A	Logic Problem	High	1	53
A	Moral Reasoning	Low	1	85
A	Moral Reasoning	Medium	1	64
A	Moral Reasoning	High	1	60

2,476 individual observations
One row for every ***condition*** (task at a given complexity level) completed for every ***team***

Step 3: Compute Nominal Teams

Observation-level group advantage is the ratio between a ***particular team****'s performance to that of equivalent "nominal team" of individuals, at a given* ***condition***

Compare to *best* member of nominal team

Compare to *random* member of nominal team

Obs. ID	Task	Complexity	Size	Score	(Observation) Strong Group Advantage	(Observation) Weak Group Advantage
B	Logic Problem	Low	3	100	1.52	2.75
B	Logic Problem	Medium	3	87	1.51	2.31
B	Logic Problem	High	3	79	1.00	1.16
B	Room Assignment	Low	3	85	0.91	1.24
B	Room Assignment	Medium	3	88	1.04	1.37
B	Room Assignment	High	3	70	0.79	0.97
C	Logic Problem	Low	6	91	1.09	1.83
C	Logic Problem	Medium	6	68	0.57	0.78
C	Logic Problem	High	6	71	0.64	0.91

3,496 team observations
One row for every ***condition*** (task at a given complexity level) completed for every ***team***

Step 4: Aggregate by Condition

Group by the ***condition*** *(task, complexity, and size) and get the ratio between a* ***all teams****' performance in that condition to that of equivalent "nominal team" of individuals*

Task	Complexity	Group Size	Strong Group Advantage	Weak Group Advantage
Logic Problem	Low	3	0.99	1.23
Logic Problem	Medium	3	0.97	1.22
Logic Problem	High	3	0.94	1.24
Logic Problem	Low	6	0.96	1.23
Logic Problem	Medium	6	0.89	1.21
Logic Problem	High	6	0.86	1.26
Room Assignment	Low	3	0.90	1.12

120 conditions
One row for every task x complexity level x group size

Figure 45. A summary of the analytical pipeline. We begin with 5,972 total observations, in which an observation is a single unit (individual or team) working to complete the assigned tasks. These observations are then divided into 2,476 individual observations and 3,496 team observations. The individual observations are used to create a baseline of "nominal teams" against which the team observations are compared. This comparison can take place at either the level of analysis of the observation (that is, did *this particular team* outperform an equivalent nominal team?) or at the level of analysis of a condition (that is, did *all teams* assigned to a particular condition outperform an equivalent nominal team?).

G.1.1. Strong Group Advantage

Let us define three parameters: the task t {Room Assignment, Writing Story, Divergent Association, …}, the level of complexity c {low, medium, high}, and the group size s {3, 6}. A condition is a tuple of (t, c, s)—a specific task at a specific level of complexity, completed by groups of a specific size.

Let there be N independent workers who completed task t at the level of complexity c, in which each worker earns a particular score $w_{\mathrm{i},□,\mathrm{x}}$ for $i = 1, \ldots, N$. The scores for the independent workers can then be aggregated into nominal groups g of an equivalent size s by permuting the different ways of selecting s worker scores from the pool of N. We can then define the average best member performance (ABMP) as follows:

$$ABMP_{\{t,c,s\}} = \frac{1}{\binom{N}{s}} \times \sum_{g=1}^{\binom{N}{s}} max(w_{\{i,\, t,\, c\}})_g$$

Then, let there be M interacting groups of size s that completed the task t at the level of complexity c. Each interacting group earns a score for the condition (t, c, s), which we represent as $T_{\{j,t,c,s\}}$ (for $j = 1, \ldots, M$). We can then compute the observation-level strong group advantage of a group of size s by taking the ratio between the group's score $T_{\{j,t,c,s\}}$ and the average best member performance (ABMP):

$$Strong\ Group\ Advantage\ (observation)_{\{t,\, c,\, s\}} = \frac{T_{\{j,\, t,\, c,\, s\}}}{ABMP_{\{t,c,s\}}}$$

Finally, to compute condition-level strong group advantage, we take the ratio between the average performance of all groups in a particular condition and the ABMP:

$$Strong\ Group\ Advantage\ (condition)_{\{t,\, c,\, s\}} = \frac{\overline{T_{\{t,c,s\}}}}{ABMP_{\{t,c,s\}}}$$

G.1.1.1 Computational Modification to ABMP

In practice, however, we perform a computational modification that is equivalent to the expression above, but affords computational efficiency. Specifically, rather than sampling $\binom{N}{s}$ nominal teams (the number of all teams of size s that can be created from the pool of N independent workers), let S_i be the score achieved by each independent worker $i = 1, ..., N$, and let B_i be the number of times each score S_i is the highest within a given nominal group. Assuming that it is possible to order the scores from highest to lowest, B_i can be computed by:

$$B_{i,\{t,\, c,\, s\}} = \binom{N-i}{s-1}$$

Accordingly, the ABMP can be computed by:

$$ABMP_{\{t,c,s\}} = \frac{\sum_i B_{i,\{t,c,s\}} S_{i,\{t,c,s\}}}{N}$$

G.1.2. Weak Group Advantage

Weak group advantage is computed similarly to strong group advantage; however, rather than comparing the average best member performance, we define the quantity average random member performance (ARMP) by taking the average of a randomly-selected score from each of the permuted nominal groups:

$$ARMP_{\{t,c,s\}} = \frac{1}{\binom{N}{s}} \times \sum_{g=1}^{\binom{N}{s}} random(w_{\{i,\,t,\,c\}})_{g}$$

Then, we can define observation-level weak group advantage as the ratio of the group's score $T_{\{j,t,c,s\}}$ to the ARMP:

$$Weak\ Group\ Advantage\ (observation)_{\{t,\,c,\,s\}} = \frac{T_{\{j,\,t,\,c,\,s\}}}{ARMP_{\{t,c,s\}}}$$

Finally, we can define condition-level weak group advantage as the ratio between the average performance of all groups in a particular condition and the ARMP:

$$Weak\ Group\ Advantage\ (condition)_{\{t,\,c,\,s\}} = \frac{\overline{T_{\{t,c,s\}}}}{ARMP_{\{t,c,s\}}}$$

G.1.2.1. Computational Modification to ARMP

Similar to the computational modification for ABMP, we perform a computational modification to efficiently calculate the ARMP without sampling $\binom{N}{s}$ nominal teams (the number of all teams of size *s* that can be created from the pool of *N* independent workers). Specifically, we observe that, since we exhaustively sample a random individual from all possible combinations of nominal teams, this is mathematically equivalent to simply taking the average score of all individual workers in a condition:

$$ARMP_{\{t,c,s\}} = \frac{\sum_{i} S_{i,\{t,c,s\}}}{N} = \overline{S_{i,\,\{t,c,s\}}}$$

G.1.3. Standard Error Estimation

We calculated the standard errors for both strong and weak group advantage using the delta method for propagating uncertainty in ratios. The standard error calculations account for the variability in both team and individual performance and are based on the following formulas:

$$SE_{strong} = Strong\ Group\ Advantage \times \sqrt{\left(\frac{SD_{Team}^{2}}{Team^{2} \times n}\right) + \left(\frac{SD_{Best\ Individual}^{2}}{Best\ Individual^{2} \times \binom{N}{s}}\right)}$$

$$SE_{weak} = Weak\ Group\ Advantage \times \sqrt{\left(\frac{SD_{Team}^{2}}{Team^{2} \times n}\right) + \left(\frac{SD_{Individual}^{2}}{Individual^{2} \times \binom{N}{s}}\right)}$$

These calculations assume that team and individual performances are independent and that the performance scores are approximately normally distributed, and that all nominal teams are equally likely.

G.1.4. Assumptions and Exceptions

Importantly, our operationalization of group advantage assumes that aggregating the scores of *s* independent workers is comparable to the joint score achieved by a team of *s* members. This is a sensible baseline for additive tasks, but may not make sense for tasks that involve social learning. For example, in an estimation task, having each member make their own independent judgments is qualitatively distinct from allowing members to discuss their judgments before submitting their final answer; in the latter case, members have the opportunity to learn from each others' guesses.

This assumption is particularly salient for one task — **Random Dot Motion**, an estimation task that taps into social learning. In the group version of the task, members of the group are able to view others' guesses of the direction of the dots' motion, and the group is evaluated on how close the *average direction of the group's guesses* is to the true direction of the dots. In this case, our estimation of the ARMP, the average score, $\overline{S_{i,\{t,c,s\}}}$, is an *underestimate*, as it would only account for the mean individual performance without the benefit of social information. To correct for this bias, we implement an exception to our calculation of the ARMP specifically for the Random Dot Motion task: **instead of averaging the final scores of all independent workers, we average their raw submissions first, which better approximates the pooled information available to group members**.

G.2. Construction of McGrath Categorical Baseline

In this section, we describe the process of transforming the Task Space into the eight (in our data, seven, since we do not include Mixed-Motive) "Types" of tasks proposed by Joseph McGrath (1984). Using the columns in the Task Space that originated from McGrath's typology, we apply the following rules encapsulating the relationships proposed in the original framework:

- Type 3 ("Intellective") and Type 4 ("Decision-Making") are combined within the same dimension; these dimensions are first separated by operationalizing Type 4 as the inverse of Type 3; that is, the more that a task is like Type 3 (Intellective tasks with a correct answer), the less that it is like Type 4 (Decision-Making tasks that are open-ended).
- Tasks are then assigned to a naive "Type" based on the dimension with the largest value. For example, if the dimension associated with "Type 5" has the highest value among all of the McGrath-proposed dimensions, the task is initially assigned to Type 5.
- If a task is assigned to Type 4 ("Decision-Making"), check that the column with the second largest value is not Type 2; this indicates that the task is, in fact, a "Generate" (Type 2) task. Because Generate tasks often involve brainstorming and creative generation, they are also open-ended like Type 4 tasks; this rule helps to disambiguate the two types.
- If a task is classified as Type 8 ("Performance"), ensure that it is also high on the "Conceptual-Behavioral" dimension (that is, it is a psychomotor task). McGrath defined Type 8 tasks as psychomotor, or physically-oriented, tasks that have an all-or-nothing outcome. For tasks that are high on the Type 8 dimensions, but that are not psychomotor, reclassify them based on the dimension with the next highest value.
- If the dimension with the highest value is Conceptual-Behavioral, the task is psychomotor, and is therefore Type 8.
- If the task has both a correct answer and a psychomotor component, reclassify it as Type 8.

The code used to implement these rules is presented below (in Python):

```
for i in range(len(df)):
    task_vec_mcgrath = df[mcgrath_colnames].iloc[i][1:]
    # the last one is conceptual-behavioral
    conceptual_behavioral = task_vec_mcgrath[-1]

    # separately break out Type 4 as the inverse of Type 3
    task_vec_mcgrath["Type 4 (Decision-Making)"] = 1-task_vec_mcgrath["Type
3 and Type 4 (Objective Correctness)"]

    # get the naive max task type
    task_name = df.iloc[i]["task_name"]
    task_type = task_vec_mcgrath.idxmax()
    type_val = task_vec_mcgrath[task_type]

    # Type 4 must be assigned ONLY if it's not a Generate Task
    if task_type == "Type 4 (Decision-Making)" and
task_vec_mcgrath[:-1].idxmax() == "Type 2 (Generate)":
        task_type = "Type 2 (Generate)"

    # Type 8 needs to be psychomotor, take next biggest category if it
isn't
    if task_type == "Type 8 (Performance)" and conceptual_behavioral < 0.5:
        task_vec_mcgrath = task_vec_mcgrath.drop(columns=["Type 8
(Performance)"])
        task_type = task_vec_mcgrath.idxmax()

    # Max type is Conceptual, reclassify as Type 8
    if task_type == "Conceptual-Behavioral":
        task_type = "Type 8 (Performance)"

    # Correct answer and psychomotor component, reclassify as Type 8
    if task_type == "Type 3 and Type 4 (Objective Correctness)":
        if type_val >= 0.5 and conceptual_behavioral >= 0.5:
            task_type = "Type 8 (Performance)"
        elif type_val >= 0.5 and conceptual_behavioral < 0.5:
            task_type = "Type 3 (Intellective)"

    mcgrath_categorical_buckets[task_name] = task_type

mcgrath_df = pd.DataFrame({
    "task_name": mcgrath_categorical_buckets.keys(),
    "mcgrath_category": mcgrath_categorical_buckets.values()
})

mcgrath_df_categorical = pd.concat([mcgrath_df["task_name"],
    pd.get_dummies(mcgrath_df["mcgrath_category"],
dtype=int).add_suffix('_cat')], axis=1)

mcgrath_categorical = list(mcgrath_df_categorical.columns)
mcgrath_categorical.remove("task_name")
```

```
# after calculating the categories, return the dataframe
return (df.merge(mcgrath_df_categorical, on="task_name"),
mcgrath_categorical)
```

Finally, the results of applying this algorithm on our data (Table 8) demonstrate that the algorithm has yielded sensible results; “Intellective” tasks include the Logic Problem, Sudoku, Wolf Goat Cabbage, and Room Assignment; “Generate” tasks include Writing Story, Advertisement Writing, Word Construction, Divergent Association, and Putting Food Into Categories; the sole “Decision-Making” Task is Moral Reasoning, and the sole “Cognitive Conflict” task is Allocating Resources; finally, Typing and Whac-a-Mole are Performance tasks.

We also observe that some tasks have puzzling, or perhaps unclear, categorizations. Our two recall tasks (Recall Association and Recall Word Lists) are both classified as Type 3 (Intellective), likely because the tasks both involve recalling a “correct” answer. However, these tasks intuitively feel different from the other “Intellective” tasks, as they don't involve a problem-solving component—simply rote memorization. However, there is no category in McGrath’s typology for rote memorization tasks, and therefore the recall tasks are relegated to Type 3. Similarly, “Random Dot Motion” and “Guess the Correlation” both involve making educated estimations about some underlying truth, rather than problem-solving. But because there is no category for estimation tasks, these tasks are also relegated to Type 3. This phenomenon reinforces our argument from the main text that categories of tasks are often not sufficiently expressive—some tasks may not belong neatly to any one category, causing loss of within-category variation.

Table 8: Categorical Assignment of Tasks to McGrath (1984)’s Types

Task Name	**Type 2 (Generate)**	**Type 3 (Intellective)**	**Type 4 (Decision-Making)**	**Type 5 (Cognitive Conflict)**	**Type 8 (Performance)**
Advertisement Writing	**1**	0	0	0	0
Allocating Resources	0	0	0	**1**	0
Divergent Association	**1**	0	0	0	0
Guessing the Correlation	0	**1**	0	0	0
Logic Problem	0	**1**	0	0	0
Moral Reasoning	0	0	**1**	0	0
Putting Food Into Categories	**1**	0	0	0	0
Random Dot Motion	0	**1**	0	0	0

Table 8: Categorical Assignment of Tasks to McGrath (1984)'s Types

Task Name	**Type 2 (Generate)**	**Type 3 (Intellective)**	**Type 4 (Decision-Making)**	**Type 5 (Cognitive Conflict)**	**Type 8 (Performance)**
Recall Association	0	**1**	0	0	0
Recall Word Lists	0	**1**	0	0	0
Room Assignment	0	**1**	0	0	0
Sudoku	0	**1**	0	0	0
Typing Game	0	0	0	0	**1**
Unscramble Words	0	**1**	0	0	0
Whac-A-Mole	0	0	0	0	**1**
WildCam	0	**1**	0	0	0
Wildcat Wells	0	**1**	0	0	0
Wolf, Goat, and Cabbage	0	**1**	0	0	0
Word Construction	**1**	0	0	0	0
Writing Story	**1**	0	0	0	0

Table 8. Assignment of the 20 tasks in our data to "types" from McGrath (1984).

G.3. Within Task "Type" Variance

In the main text, we write that there is significant heterogeneity in group advantage, both between different tasks and within a canonical "type" of tasks, such as Generate or Intellective. To statistically test this claim, we first identify the three McGrath task categories that contain more than one task in our data — "Generate," "Intellective," and "Performance." (The other two categories, "Decision-Making" and "Cognitive Conflict," have only a single task in our data, making the analysis of within-type variation meaningless.)

Using condition-level data, we examine the strong and weak group advantage ratio for each of the conditions (recall that a condition is task × complexity × group size, which means that, for a single task, there are 6 condition-level datapoints, representing 3 levels of complexity × 2 group sizes). We then statistically test the null hypothesis that the variance in group advantage for a given task category is equal to the overall variance in group advantage in the data (all 120 group conditions).

G.3.1 Descriptive Statistics

In Table 9, we present descriptive statistics of the strong and weak group advantage; we observe that, for the three task types containing more than one task, the mean variance in group advantage is comparable to the overall variance in the data; Type 2 (Generate) has greater variance than in the overall data, Type 3 (Intellective) has slightly less variance than in the overall data, and Type 8 (Performance) has greater variance for strong group advantage and slightly less variance for weak group advantage than in the overall data.

Category	Strong Group Advantage			Weak Group Advantage		
	Mean	**Std**	**Count**	**Mean**	**Std**	**Count**
Type 2 (Generate)	1.04	0.33	30	1.73	0.81	30
Type 3 (Intellective)	0.95	0.23	66	1.54	0.59	66
Type 4 (Decision-Making)	0.69	0.06	6	1.02	0.12	6
Type 5 (Cognitive Conflict)	0.67	0.10	6	1.21	0.12	6
Type 8 (Performance)	0.84	0.37	12	1.52	0.61	12
Overall	**0.94**	**0.28**	**120**	**1.55**	**0.64**	**120**

Table 9: Descriptive statistics for strong and weak outcomes for task types (from McGrath's typology) and the overall data.

G.3.2 Levene Test for Variance Across McGrath "Types"

Next, we perform a Levene Test for Equality of Variances to test whether the variances in the subsample for a particular task type is significantly different from the variance in the overall dataset.

The computed Levene test statistic (*W*) and associated *p*-values obtained are presented in Table 10. In all cases, we are unable to reject the null hypothesis that the variances are equal (all $p > 0.05$).

Category	W_{Strong}	p-value$_{Strong}$	W_{Weak}	p-value$_{Weak}$	dF.1	dF.2
Type 2 (Generate)	1.01	0.32	2.14	0.15	1	148
Type 3 (Intellective)	3.38	0.07	0.43	0.51	1	184
Type 8 (Performance)	3.61	0.06	0.07	0.79	1	130

Table 10: Levene's test statistics for categorical conditions.

G.4. Linear Models Showing Heterogeneous Interaction Effects

In the main text, we observe that interaction effects between moderators such as group complexity are also heterogeneous by the task type. To statistically test this claim, we fit a series of linear mixed-effects models on the *observation*-level data (in which a single datapoint is one team's performance within a given condition). Specifically, we fit a regression of:

```
group advantage ~ task complexity + group size
```

with a random effect for the group identifier (since the same group appears repeatedly in the data).

Since we fit one model for each task, the coefficients for `task complexity` and `group size` can be interpreted as interaction terms between a given task and each of `task complexity` and `group size`. Thus, a positive and significant coefficient on `task complexity` suggests that groups are likely to have more advantage as the task increases in complexity; a negative and significant coefficient on `task complexity` suggests that groups are likely to have less advantage as the task increases in complexity. Similarly, a positive and significant coefficient on `group size` suggests that groups are likely to have more advantage for the task when they are larger; a negative and significant coefficient on `group size` suggests that groups are likely to have more advantage for the task when they are smaller.

Full results from these 40 regressions (2 outcomes × 20 tasks) is presented below.

Advertisement Writing
Strong Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.894	0.187	4.795	0.000
task complexity	0.072	0.054	1.341	0.180
group size	-0.098	0.033	-2.966	0.003
group identifier	0.033	0.057		

Weak Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.217	0.435	2.800	0.005
task complexity	0.347	0.129	2.697	0.007
group size	-0.120	0.076	-1.570	0.116
group identifier	0.120	0.125		

Allocating Resources
Strong Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.850	0.154	5.511	0.000
task complexity	0.039	0.038	1.030	0.303
group size	-0.057	0.029	-1.989	0.047
group identifier	0.046	0.059		

Weak Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.483	0.284	5.225	0.000
task complexity	-0.007	0.070	-0.095	0.924
group size	-0.057	0.053	-1.089	0.276
group identifier	0.157	0.110		

Divergent Association

Strong Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.945	0.039	24.099	0.000
task complexity	0.035	0.012	2.941	0.003
group size	-0.008	0.007	-1.274	0.203
group identifier	0.000	0.010		

Weak Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.133	0.049	23.153	0.000
task complexity	0.024	0.015	1.650	0.099
group size	-0.000	0.008	-0.057	0.955
group identifier	0.001	0.013		

Guess the Correlation

Strong Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.974	0.082	11.893	0.000
task complexity	-0.032	0.025	-1.274	0.202
group size	-0.012	0.014	-0.896	0.370
group identifier	0.001	0.022		

Weak Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.253	0.119	10.522	0.000
task complexity	-0.054	0.036	-1.472	0.141
group size	0.024	0.020	1.169	0.242
group identifier	0.002	0.031		

Logic Problem
Strong Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.094	0.040	27.109	0.000
task complexity	-0.032	0.012	-2.701	0.007
group size	-0.020	0.007	-2.902	0.004
group identifier	0.001	0.012		

Weak Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.208	0.053	22.855	0.000
task complexity	0.009	0.015	0.587	0.557
group size	0.001	0.009	0.097	0.923
group identifier	0.002	0.016		

Moral Reasoning
Strong Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.666	0.128	5.190	0.000
task complexity	-0.035	0.026	-1.345	0.179
group size	0.021	0.025	0.822	0.411
group identifier	0.057	0.069		

Weak Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.843	0.187	4.498	0.000
task complexity	-0.038	0.038	-0.997	0.319
group size	0.057	0.037	1.561	0.119
group identifier	0.119	0.098		

Putting Food Into Categories

Strong Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.594	0.199	2.986	0.003
task complexity	-0.038	0.030	-1.255	0.209
group size	0.180	0.041	4.394	0.000
group identifier	0.193	0.155		

Weak Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.324	0.326	0.992	0.321
task complexity	0.014	0.051	0.282	0.778
group size	0.417	0.067	6.223	0.000
group identifier	0.508	0.246		

Random Dot Motion

Strong Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.152	0.109	10.600	0.000
task complexity	-0.089	0.034	-2.657	0.008
group size	-0.052	0.018	-2.856	0.004
group identifier	0.000	0.051		

Weak Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.286	0.169	7.612	0.000
task complexity	-0.014	0.052	-0.269	0.788
group size	-0.058	0.028	-2.021	0.043
group identifier	0.000	0.046		

Recall Association

Strong Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.247	0.179	6.968	0.000
task complexity	0.093	0.029	3.202	0.001
group size	-0.007	0.037	-0.181	0.856
group identifier	0.148	0.127		

Weak Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.645	0.325	5.062	0.000
task complexity	0.136	0.053	2.576	0.010
group size	0.178	0.066	2.686	0.007
group identifier	0.487	0.231		

Recall Word Lists

Strong Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.039	0.148	7.009	0.000
task complexity	-0.049	0.041	-1.174	0.241
group size	-0.008	0.026	-0.317	0.751
group identifier	0.022	0.047		

Weak Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.099	0.299	3.672	0.000
task complexity	0.148	0.083	1.775	0.076
group size	0.068	0.053	1.265	0.206
group identifier	0.095	0.096		

Room Assignment

Strong Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.910	0.086	10.545	0.000
task complexity	-0.012	0.012	-1.008	0.313
group size	0.004	0.018	0.210	0.834
group identifier	0.038	0.077		

Weak Group Advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.895	0.120	7.465	0.000
task complexity	0.114	0.017	6.519	0.000
group size	0.028	0.025	1.113	0.266

Sudoku

Strong group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.070	0.072	14.820	0.000
task complexity	-0.062	0.018	-3.473	0.001
group size	-0.007	0.013	-0.499	0.618
group identifier	0.010	0.028		

Weak group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.031	0.091	11.330	0.000
task complexity	0.001	0.023	0.065	0.949
group size	0.016	0.017	0.943	0.345
group identifier	0.015	0.034		

Typing game

Strong group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.869	0.245	3.545	0.000
task complexity	0.012	0.029	0.423	0.672
group size	0.056	0.051	1.094	0.274
group identifier	0.280	0.248		

Weak group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.951	0.433	2.193	0.028
task complexity	0.070	0.054	1.289	0.197
group size	0.203	0.090	2.259	0.024
group identifier	0.859	0.415		

Unscramble Words

Strong group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.202	0.101	11.920	0.000
task complexity	0.083	0.031	2.674	0.008
group size	-0.023	0.017	-1.363	0.173
group identifier	0.000	0.039		

Weak group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.318	0.237	5.559	0.000
task complexity	0.291	0.073	3.973	0.000
group size	0.094	0.040	2.336	0.019
group identifier	0.002			

Whac-A-Mole
Strong group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.351	0.142	9.493	0.000
task complexity	-0.107	0.035	-3.046	0.002
group size	-0.131	0.028	-4.767	0.000
group identifier	0.052	0.060		

Weak group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	2.231	0.268	8.334	0.000
task complexity	-0.176	0.066	-2.652	0.008
group size	-0.181	0.052	-3.494	0.000
group identifier	0.182	0.111		

WildCam
Strong group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.024	0.163	6.298	0.000
task complexity	-0.012	0.041	-0.283	0.777
group size	-0.024	0.030	-0.780	0.435
group identifier	0.058	0.063		

Weak group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.197	0.267	4.491	0.000
task complexity	0.043	0.069	0.627	0.531
group size	0.044	0.049	0.897	0.370
group identifier	0.137	0.097		

Wildcat Wells
Strong group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.816	0.109	7.504	0.000
task complexity	0.018	0.026	0.699	0.485
group size	-0.023	0.021	-1.130	0.258
group identifier	0.031	0.046		

Weak group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.992	0.170	5.821	0.000
task complexity	0.034	0.040	0.853	0.394
group size	0.028	0.033	0.859	0.391
group identifier	0.078	0.073		

Wolf, Goat, and Cabbage Problem
Strong group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.671	0.122	5.516	0.000
task complexity	0.051	0.029	1.740	0.082
group size	0.031	0.023	1.321	0.186
group identifier	0.041	0.052		

Weak group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.461	0.239	6.121	0.000
task complexity	-0.183	0.062	-2.940	0.003
group size	0.148	0.045	3.324	0.001
group identifier	0.117	0.088		

Word Construction

Strong group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.884	0.193	4.588	0.000
task complexity	0.147	0.031	4.755	0.000
group size	0.043	0.039	1.107	0.268
group identifier	0.153	0.140		

Weak group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.027	0.384	2.671	0.008
task complexity	0.190	0.077	2.471	0.013
group size	0.306	0.075	4.061	0.000
group identifier	0.481	0.209		

Writing Story

Strong group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	0.900	0.043	20.745	0.000
task complexity	0.015	0.011	1.358	0.174
group size	0.002	0.008	0.236	0.814
group identifier	0.004	0.017		

Weak group advantage

Variable	Coef.	Std.Err.	z	P>|z|
Intercept	1.141	0.050	22.685	0.000
task complexity	-0.049	0.013	-3.794	0.000
group size	0.008	0.009	0.877	0.381
group identifier	0.005	0.019		

G.5. Primary Model Details

Our models focus on the *condition* level of analysis ($n = 120$) and examine 3 *covariate sets* (Task Space, McGrath Categorical, McGrath Subspace) × 2 *dependent variables* (Strong Group Advantage and Weak Group Advantage), resulting in 6 primary models.

These 6 primary models are hyperparameter-tuned ElasticNets, evaluated using root mean-squared error (RMSE). Further evaluation details can be found in Appendix G.7 (Model Evaluation), and our code can be found on GitHub.

G.5.1 ElasticNet Model Details

We use scikit-learn's ElasticNetCV to conduct a hyperparameter search for the ElasticNet models using 5-fold cross-validation, and fit the final model using the optimal hyperparameters using scikit-learn's ElasticNet. All ElasticNet models include task dimensions, group size, and task complexity, as well as all polynomial combinations of degree 2. We use PolynomialFeatures from scikit-learn to calculate the higher-order interactions:

```
def add_interactions(X):
    poly = PolynomialFeatures(degree=2, interaction_only=True, include_bias=False)
    return poly.fit_transform(X)
```

The ElasticNet objective function is as follows:

```
 1 / (2 * n_samples) * ||y - Xw||^2_2
 + alpha * l1_ratio * ||w||_1
 + 0.5 * alpha * (1 - l1_ratio) * ||w||^2_2
```

We first perform a grid search across 50 values of `alpha` and the following values of `l1_ratio`: `[0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 1]`. For every combination of `alpha` and `l1_ratio`, `ElasticNetCV` performs 5-fold cross-validation (training on 80% of the training data and testing on 20% of the training data). Once the optimal hyperparameters are identified, we fit a single ElasticNet model on the *full* training data. This model is then evaluated on the held-out test data.

```
def train_wave_a_enet(wave_a_data, dv_type, ivs, random_state=19104):
    assert dv_type in ["strong", "weak"]
    X = wave_a_data[ivs]
    X_interactions = add_interactions(X)
    y = wave_a_data[dv_type]

    # Define a range of values for alpha and l1_ratio
    alphas = np.logspace(-4, 1, 50)
```

```
   l1_ratio = [0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 1]

   # Initialize and fit ElasticNetCV
   elastic_net_cv = ElasticNetCV(cv=5, alphas=alphas, l1_ratio=l1_ratio,
random_state=random_state)
   elastic_net_cv.fit(X_interactions, y)

   # Re-fit the model on the full training dataset using the best hyperparameters
   elastic_net = ElasticNet(alpha=elastic_net_cv.alpha_, l1_ratio=elastic_net_cv.l1_ratio_)
   elastic_net.fit(X_interactions, y)

   return elastic_net

def test_wave_b_enet(wave_a_data, wave_b_data, model, dv_type, ivs):
   assert dv_type in ["strong", "weak"]

   X = wave_b_data[ivs]
   X_interactions = add_interactions(X)
   y_actual = wave_b_data[dv_type]
   y_pred = model.predict(X_interactions)

   r2 = custom_r2(y_pred, y_actual, wave_a_data, wave_b_data, dv_type)

   return r2, y_actual, y_pred
```

G.5.2 Distribution of Pairwise Differences

Figure 5 in the main text presents the performance of the ElasticNet models on the held-out test data (Wave 3), with error bars representing marginal 95% bootstrap intervals for each model's individual RMSE performance. Notably, these error bars are not tests of the *differences* in model performance. The pairwise differences in performance are the relevant estimand for evaluating whether the Task Space is more effective at predicting group advantage than the McGrath-inspired baselines. In Figure 46 below, we plot the empirical distribution of the paired differences in RMSE with a reference line at zero. This figure makes clear that, when evaluated on the same bootstrapped resample, the Task Space significantly outperforms both McGrath Categorical and McGrath Subspace.

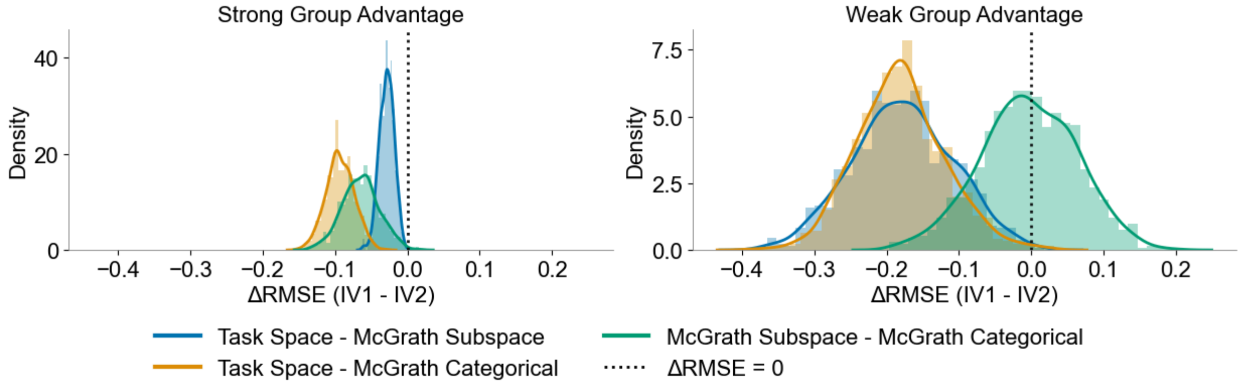


Figure 46. Empirical difference distributions in model performance for the primary ElasticNet models. Significance in our main text is assessed on paired bootstrap differences computed on the same resamples. We use the empirical difference distribution to report the mean difference in RMSE, 95% CI and a two-sided *p*-value for each pair of models. Thus, in the main text, error bars for the individual models in Figure 5 may overlap even when the paired difference is consistently nonzero. Because both models are evaluated on the same resamples, their estimates move together, and much of the variability that widens each bar cancels out in the difference.

G.6. Supplementary Models and Robustness Checks

Next, we test the robustness of our main findings against two additional model types: Ordinary Least Squares (OLS) and neural networks (NN). We also evaluate the Task Space models against two additional baselines based on the Laughlin and Ellis (1986) and Steiner (1972) taxonomies.

G.6.1 Ordinary Least Squares (OLS) Model Details

Our OLS robustness check models are implemented using Statsmodels' OLS function. Our train/test pipeline is as follows; note that the "custom R^2" function referenced in the code is an out-of-sample R^2 measure detailed further in Appendix G.7 (Model Evaluation). Furthermore, note that, in our code, the variable `ivs` contains both the focal task variables and the exogenous variables (group size and task complexity).

```
def train_wave_a(wave_a_data, dv_type, ivs):
   assert dv_type in ["strong", "weak"]
   X = wave_a_data[ivs]
   X = sm.add_constant(X, has_constant='add')
   y = wave_a_data[dv_type]
   return sm.OLS(y, X).fit()

def test_wave_b(wave_a_data, wave_b_data, model, dv_type, ivs):
   assert dv_type in ["strong", "weak"]

   X = wave_b_data[ivs]
   X = sm.add_constant(X, has_constant='add')
   y_actual = wave_b_data[dv_type]
   y_pred = model.predict(X)
   r2 = custom_r2(y_pred, y_actual, wave_a_data, wave_b_data, dv_type)

   return r2, y_actual, y_pred
```

Table 11 shows the OLS results. Using two-sided p-values derived directly from the empirical distribution of the differences across 1,000 bootstrapped resamples of Wave 3, we find that the Task Space significantly outperforms McGrath Categorical for strong group advantage (mean ΔRMSE = -0.072, 95% $\text{CI}_{\Delta\text{RMSE}}$[-0.093, -0.044], $p = 0.002$). The Task Space also directionally outperforms the McGrath Subspace, but this difference is not statistically significant (mean ΔRMSE = -0.025, 95% $\text{CI}_{\Delta\text{RMSE}}$[-0.070, 0.017], $p = 0.300$). For weak group advantage, we find that the Task Space directionally outperforms McGrath Categorical, but the result is not statistically significant (mean ΔRMSE = -0.119, 95% $\text{CI}_{\Delta\text{RMSE}}$[-0.239, 0.021], $p = 0.092$). However, the Task Space significantly outperforms the McGrath Subspace (mean ΔRMSE = -0.122, 95% $\text{CI}_{\Delta\text{RMSE}}$[-0.186, -0.042], $p = 0.002$). Finally, the McGrath Subspace is not significantly different from McGrath Categorical for either strong group advantage (mean ΔRMSE = -0.046, 95% $\text{CI}_{\Delta\text{RMSE}}$[-0.103, 0.013], $p = 0.140$) or weak group advantage (mean ΔRMSE = 0.003, 95% $\text{CI}_{\Delta\text{RMSE}}$[-0.115, 0.130], $p = 0.965$).

Table 11: Supplementary OLS Models

No.	DV Type	Covariate Set	R^2 *Higher is better;* *Perfect Prediction = 1*	RMSE *Lower is better;* *Perfect Prediction = 0*
1	Strong	Task Space	0.33	0.30
2	Weak	Task Space	0.35	0.61
3	Strong	McGrath Categorical	-0.03	0.38
4	Weak	McGrath Categorical	0.06	0.73
5	Strong	McGrath Subspace	0.21	0.33
6	Weak	McGrath Subspace	0.06	0.73

Table 11. Replication of the main findings using Ordinary Least Squares (OLS) models.

G.6.2 Alternative Task Frameworks (Steiner, Laughlin)

We repeat our primary models using the subspace of task variables inspired by the Steiner (1972) and the Laughlin and Ellis (1986) taxonomies. (We note that, in Steiner's case, the variables are not comprehensive; some of Steiner's proposed dimensions do not pertain to the task class, so they are not present in the Task Space.) We follow the same methods as in Appendix G.5.

The independent variables used for the Steiner models are the following:

```
'steiner_continuous': ['Divisible-Unitary',
                       'Maximizing',
                       'Optimizing',
                       'playerCount',
                       'Low',
                       'Medium',
                       'High']
```

The independent variables used for the Laughlin and Ellis models are the following:

```
'laughlin_continuous': ['Decision Verifiability',
                        'Shared Knowledge',
                        'Within-System Solution',
                        'Answer Recognizability',
                        'Time Solvability',
                        'Intellective-Judgmental',
                        'Eureka Question',
                        'playerCount',
                        'Low',
                        'Medium',
                        'High']
```

The results of these models are presented in Table 12 below. Using two-sided *p*-values derived directly from the empirical distribution of the differences across 1,000 bootstrapped resamples of the Wave 3 data, we replicate our analysis using both ElasticNet and OLS models.

The ElasticNet results reinforce the main findings (that the Task Space outperforms a subspace inspired by a single taxonomy). For strong group advantage, the Task Space significantly outperforms the Steiner Subspace (mean ΔRMSE = -0.061, 95% $\mathrm{CI}_{\Delta\mathrm{RMSE}}$[-0.082, -0.038], *p* = 0.002) but directionally outperforms the Laughlin Subspace without statistical significance (mean ΔRMSE = -0.033, 95% $\mathrm{CI}_{\Delta\mathrm{RMSE}}$[-0.081, 0.013], *p* = 0.192). For weak group advantage, the Task Space significantly outperforms both the Steiner Subspace (mean ΔRMSE = -0.084, 95% $\mathrm{CI}_{\Delta\mathrm{RMSE}}$[-0.130, -0.038], *p* = 0.002) and the Laughlin Subspace (mean ΔRMSE = -0.158, 95% $\mathrm{CI}_{\Delta\mathrm{RMSE}}$[-0.279, -0.038], *p* = 0.012). Performance differences between the Steiner and Laughlin Subspaces are not statistically significant for either strong group advantage (mean ΔRMSE = 0.028, 95% $\mathrm{CI}_{\Delta\mathrm{RMSE}}$[-0.006, 0.063], *p* = 0.136) or weak group advantage (mean ΔRMSE = -0.074, 95% $\mathrm{CI}_{\Delta\mathrm{RMSE}}$[-0.171, 0.038], *p* = 0.166).

The OLS models show similar, though weaker, patterns. For strong group advantage, the Task Space directionally outperforms the Steiner Subspace, but this difference is not statistically significant (mean ΔRMSE = -0.045, 95% $\mathrm{CI}_{\Delta\mathrm{RMSE}}$[-0.102, 0.013], *p* = 0.128); similarly, the Task Space directionally outperforms the Laughlin Subspace (mean ΔRMSE = -0.038, 95% $\mathrm{CI}_{\Delta\mathrm{RMSE}}$[-0.137, 0.060], *p* = 0.450), but the difference is nonsignificant. For weak group advantage, the Task Space significantly outperforms the Steiner Subspace (mean ΔRMSE = -0.080, 95% $\mathrm{CI}_{\Delta\mathrm{RMSE}}$[-0.106, -0.053], *p* = 0.002) and marginally outperforms the Laughlin Subspace (mean ΔRMSE = -0.209, 95% $\mathrm{CI}_{\Delta\mathrm{RMSE}}$[-0.387, -0.003], *p* = 0.050). As with the ElasticNet models, differences between Steiner and Laughlin subspaces are not statistically significant for either strong group advantage (mean ΔRMSE = 0.007, 95% $\mathrm{CI}_{\Delta\mathrm{RMSE}}$[-0.040, 0.062], *p* = 0.827) or weak group advantage (mean ΔRMSE = -0.129, 95% $\mathrm{CI}_{\Delta\mathrm{RMSE}}$[-0.314, 0.079], *p* = 0.214).

Table 12: Supplementary Models with Steiner and Laughlin Frameworks

No.	**DV Type**	**Covariate Set**	**Model Type**	**R^2** *Higher is better;* *Perfect Prediction = 1*	**RMSE** *Lower is better;* *Perfect Prediction = 0*
1	Strong	Steiner Subspace	ElasticNet	0.15	0.34
2	Weak	Steiner Subspace	ElasticNet	0.22	0.67
3	Strong	Laughlin Subspace	ElasticNet	0.28	0.31
4	Weak	Laughlin Subspace	ElasticNet	0.03	0.74
5	Strong	Steiner Subspace	OLS	0.10	0.35
6	Weak	Steiner Subspace	OLS	0.16	0.69
7	Strong	Laughlin Subspace	OLS	0.12	0.35
8	Weak	Laughlin Subspace	OLS	-0.20	0.82

Table 12. Results (R^2 and RMSE) of our robustness check models using the Steiner and the Laughlin and Ellis taxonomies. We replicate the results using both OLS and ElasticNet models.

G.6.3 Neural Network Model Details

Our neural network robustness check models are implemented in Tensorflow, using Adam as the optimizer and ReLu as the activation function. When optimizing the neural network models, we take a leave-one-out approach, separately training a model on 14 of the 15 tasks, then evaluating the loss on the single held out task. The total loss in each iteration is then computed as the sum of all losses across the 15 cross-validation folds. This approach seeks to reduce the risk of overfitting by penalizing the model for poor performance on an unseen task.

We applied this leave-one-task-out cross-validation to tune the following hyperparameters (using the Bayesian optimization package, Optuna, and 100 iterations):

```
learning_rate = trial.suggest_loguniform("learning_rate", 1e-6, 1e-1)
   n_units = trial.suggest_int("n_units", 32, 512)
   n_layers = trial.suggest_categorical("n_layers", [1, 2, 3, 4, 5])
   batch_size = trial.suggest_categorical("batch_size", [32, 64, 128, 256])
   dropout_rate = trial.suggest_uniform("dropout_rate", 0.0, 0.3)
   activation = trial.suggest_categorical("activation", ['relu'])
```

Once tuned, we built the final model as follows:

```
def create_best_model(input_shape, best_params):
   learning_rate = best_params['learning_rate']
   n_units = best_params['n_units']
   n_layers = best_params['n_layers']
   batch_size = best_params['batch_size']
   dropout_rate = best_params['dropout_rate']
   activation = best_params['activation']

   # Build model
   model = Sequential()
   model.add(Dense(n_units, activation=activation, input_shape=input_shape))

   for _ in range(1, n_layers):
       model.add(Dense(n_units, activation=activation))
       model.add(Dropout(dropout_rate))

   model.add(Dense(1))
   model.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=learning_rate),
                 loss='mean_squared_error')

   return model, batch_size
```

We present the code from our complete train/test pipeline below. Similar to our approach with ElasticNet, we first use cross-validation to identify the optimal hyperparameters (i.e., 100 iterations × 15 leave-one-task-out cross-validations). We then used the optimized hyperparameters to fit a single neural network on the training data, before evaluating the model on the test data.

```
def train_wave_a_nn(wave_a_data, ivs, dv_type):
   assert dv_type in ["strong", "weak"]

   # Random seed for reproducibility
   set_random_seeds()

   # Objective function for Optuna
   def objective(trial):
       return leave_one_task_out(wave_a_data, lambda shape: create_model(trial, shape), ivs,
dv_type)

   # Hyperparameter optimization
   study = create_study(direction="minimize")
   study.optimize(objective, n_trials=100)

   best_params = study.best_trial.params

   input_shape = (wave_a_data[ivs].shape[1],)
   model, batch_size = create_best_model(input_shape, best_params)

   X_train = wave_a_data[ivs]
   y_train = wave_a_data[dv_type]

   model.fit(X_train, y_train, batch_size=batch_size, epochs=100, verbose=0)

   print(f"Best hyperparameters: {best_params}")

   return model

def test_wave_b_nn(wave_a_data, wave_b_data, model, ivs, dv_type):
   X_new = wave_b_data[ivs]
   y_actual = wave_b_data[dv_type]

   # Ensure consistent shape
   if len(X_new.shape) == 1:
       X_new = X_new.reshape(-1, 1)

   y_pred = model.predict(X_new).flatten()

   r2 = custom_r2(y_pred, y_actual, wave_a_data, wave_b_data, dv_type)

   return r2, y_actual, y_pred
```

The results of our neural network models are summarized in Table 13. Using two-sided p-values derived from the empirical distribution of the differences across 1,000 bootstrapped resamples of Wave 3, we again replicate our main results. For strong group advantage, the Task Space significantly outperforms McGrath Categorical (mean ΔRMSE = -0.084, 95% $\text{CI}_{\Delta\text{RMSE}}$[-0.135, -0.033], $p = 0.002$). However, the Task Space is not significantly different from the McGrath Subspace (mean ΔRMSE = -0.009, 95% $\text{CI}_{\Delta\text{RMSE}}$[-0.050, 0.039], $p = 0.695$). For weak group advantage, the Task Space significantly outperforms both McGrath Categorical (mean ΔRMSE = -0.183, 95% $\text{CI}_{\Delta\text{RMSE}}$[-0.296, -0.052], $p = 0.012$) and McGrath Subspace (mean ΔRMSE = -0.190, 95% $\text{CI}_{\Delta\text{RMSE}}$[-0.293, -0.078], $p = 0.002$). Among the McGrath-inspired baselines, McGrath Subspace significantly outperforms McGrath Categorical for strong group advantage (mean ΔRMSE = -0.076, 95% $\text{CI}_{\Delta\text{RMSE}}$[-0.114, -0.036], $p = 0.002$), not for weak group advantage (mean ΔRMSE = 0.006, 95% $\text{CI}_{\Delta\text{RMSE}}$[-0.124, 0.133], $p = 0.907$).

Table 13: Supplementary Neural Network Models

No.	**DV Type**	**Covariate Set**	**R^2** *Higher is better;* *Perfect Prediction = 1*	**RMSE** *Lower is better;* *Perfect Prediction = 0*
1	Strong	Task Space	0.37	0.29
2	Weak	Task Space	0.32	0.62

Table 13: Supplementary Neural Network Models				
No.	**DV Type**	**Covariate Set**	**R²** *Higher is better;* *Perfect Prediction = 1*	**RMSE** *Lower is better;* *Perfect Prediction = 0*
3	Strong	McGrath Categorical	-0.05	0.38
4	Weak	McGrath Categorical	-0.15	0.81
5	Strong	McGrath Subspace	0.33	0.30
6	Weak	McGrath Subspace	-0.16	0.81

Table 13. Replication of the main findings using Neural Networks.

G.7. Model Evaluation

We evaluate all models using root mean squared error (RMSE), implemented as follows:

```
def get_rmse(y_true, y_pred):
   return np.sqrt(np.mean((y_true - y_pred) ** 2))
```

Lower values of RMSE are better; a model making perfect predictions would have zero error (and thus RMSE = 0). RMSE has no upper bound, as models can have arbitrarily poor performance.

As a secondary evaluation criterion, we also examine *out-of-sample R^2*, which we define as

$$R^2_{test} = 1 - \frac{\sum_{i=1}^{N}\left(\widehat{y_i} - y_i\right)^2}{\sum_{i=1}^{N}\left(\bar{x} - y_i\right)^2}$$

$\widehat{y}_i$ is the predicted advantage ratio for the ith group in the test data, y_i is the observed value for that group, and $\bar{x}$ is the average ratio over all instances of groups of the same size performing all the tasks in the training set at the same complexity level. Thus, R^2_{test} quantifies the proportion of-out-of-sample variance (squared error) explained by our model compared to a naïve baseline model that always predicts the mean of the training data. An $R^2_{test} = 1$ would correspond to perfect prediction for every test instance; $R^2_{test} = 0$ would correspond to performance equivalent to the naïve baseline; and negative values of R^2_{test} would correspond to predictions that are worse than the naïve baseline.

We emphasize that R^2_{test} represents a "true" out of sample measure of predictive accuracy compared to standard R^2 metrics, which may inflate the measurement of variance explained by training and testing on the same data. In contrast, this criterion evaluates the extent to which our models generalize from one wave to another, despite collecting the data at an entirely different time period and on an entirely different set of tasks.

The out-of-sample R^2 function is defined in Python as follows:

```
def custom_r2(y_pred, y_actual, wave_a_data, wave_b_data, dv_type):
    # Compute R^2 on the test set, using the training set as a baseline
    naive_prediction_errs = []
    for i, row in wave_b_data.iterrows():

        # get all instances of groups of the same size performing
        # all the tasks in the training set at the same complexity level.
        playerCount = row["playerCount"]
        wave_a_subset = wave_a_data[(wave_a_data["playerCount"] == playerCount) &
                                    (wave_a_data["Low"] == row["Low"]) &
                                    (wave_a_data["Medium"] == row["Medium"]) &
                                    (wave_a_data["High"] == row["High"])]
        y_training = np.mean(wave_a_subset[dv_type])
        # predict the value of the DV (in wave_b) using the mean of the training data (from
wave_a)
        y_actual_i = row[dv_type]
        fold_err = (y_actual_i - y_training)**2
        naive_prediction_errs.append(fold_err)

    r2 = 1 - np.sum((y_pred - y_actual)**2) / np.sum(naive_prediction_errs)

    return r2
```

Appendix H: Review of Meta-Analyses of Teamwork

In the main paper, we claim that, unlike in other social science domains such as personality, there has been relatively little attention paid to the various attributes of tasks across studies, and that the ability to empirically verify the importance of different task attributes across contexts has been limited as a result. To support this claim, we conducted a review of meta-analyses related to the subject of teamwork, and coded the descriptions or attributes of the tasks reported.

Our analysis included a total of 21 meta-analyses, which includes all papers cited in Larson (2010)'s review, augmented with recently-published work obtained via a literature search (using the keywords "teamwork," teams," "groups," and "meta-analysis" on Google Scholar).

Of these 21 published papers, three provided no information about the task at all (Bond Jr. & Van Leeuwen, 1991; Dennis & Williams, 2005; Mullen et al., 1991). The remaining meta-analyses vary widely in their level of detail. Several encoded only a single attribute of the task, such as task complexity (Bond Jr. & Titus, 1983) or task interdependence (Gully et al., 1995). However, we observed that, among papers that focused on the same dimension, coding schemes were not always consistent; for example, while Gully et al. (1995) and Marlow et al. (2018) used binary (High/Low) schemes to code task interdependence, DeChurch & Mesmer-Magnus (2010) used a ternary scheme (High/Moderate/Low), and LePine et al., (2008) used a 10-point scheme.

Among those that reported more than one dimension, every paper used one primary framework or coding scheme, such as McGrath's or Steiner's typologies. Some used variations of canonical frameworks; for example, De Dreu & Weingart (2003) used five categories rather than McGrath's typical eight, and wrote the following justification in their Method section:

> Too little information about the groups' tasks was provided in the research articles to directly code for task complexity, uncertainty, or routineness. Instead, group tasks were classified into more global categories using McGrath's (1984) group-task circumplex, and then assumptions about their complexity or uncertainty were made. The tasks performed in the studies fell into four categories: planning–production tasks, decision-making tasks, project tasks, and mixed tasks in which teams performing different tasks were combined into one sample.

Our findings highlight the need to more comprehensively measure different attributes of tasks. At present, decisions for how to categorize a task appear ad-hoc, as different papers focus on different typologies, and there is typically insufficient information to evaluate other attributes of the task. When descriptions of tasks are present, they are often extremely brief (e.g., "brainstorming") without any specification of the topic, objective, or other key information. Consequently, it would be difficult to characterize these tasks according to dimensions beyond those that were coded and reported in the study, and it would also be difficult to evaluate task attributes across empirical settings.

Table 14: Review of Task Descriptions in Teamwork Meta-Analyses

No.	Citation	Descriptions and Features of Tasks Provided
1	(Baltes et al., 2002)	McGrath Categorization
2	(Bell, 2007)	Study Setting (Lab, Field) Steiner Categorization (Additive, Disjunctive, Conjunctive, Compensatory) Team Type (Advisory, Design, Executive, Production)
3	(Bell et al., 2011)	Study Setting (Lab, Field)

Table 14: Review of Task Descriptions in Teamwork Meta-Analyses

No.	Citation	Descriptions and Features of Tasks Provided
		Devine Team Typology (Physical versus Intellectual, with subcategories) Performance Measures Coded (Efficiency, General Performance, Creativity, Innovation)
4	(Bond Jr. & Titus, 1983)	Classified Task Complexity as “Simple,” “Complex,” or “Unknown”
5	(Bond Jr. & Van Leeuwen, 1991)	N/A (Did Not Report)
6	(Bowers et al., 2000)	Task Performance Measure (Quality, Quantity, Accuracy) Task Difficulty (High, Medium, Low)
7	(DeChurch & Mesmer-Magnus, 2010)	Task Interdependence (High, Moderate, Low) Team Type (Action, Decision-Making, Project)
8	(De Dreu & Weingart, 2003)	Variation of McGrath Categorization: (1) Decision Making; (2) Project; (3) Production; (4) Planning; or (5) Mixed
9	(Dennis & Williams, 2005)	N/A (Did Not Report)
10	(Devine & Philips, 2001)	Brief Description *Description Examples:* “SouthEast Airlines top management simulation,” “Generating a short fictional story,” “Brainstorming”
11	(Gully et al., 1995)	Task Interdependence (High, Low)
12	(Gully et al., 2002)	0-3 Scale for Task, Goal, and Outcome Independence
13	(Halfhill et al., 2005)	Classified as “Lab” or “Field,” with a classification of Team Types within the setting: (Lab) Problem solving, Decision Making, Brainstorming, Management Simulation (Field) Production, Management, Action/Performing, Service
14	(LePine et al., 2008)	Study Setting (Lab, Field) Task Interdependence (1-10 scale, 1 = *pooled interdependence with members performing their work alone*; 10 = *comprehensive interdependence with all task work performed in the presence of the other members*)
15	(Marlow et al., 2018)	Simple Task Type (Cognitive-Based, Action-Based) Wildman et al. (2012)’s categorization: (1) managing others, (2) advising others, (3) human service, (4) negotiation, (5) psychomotor activity, (6) defined problem solving, (7) ill-defined problem solving Task Interdependence (High, Low)
16	(McEwan et al., 2017)	Study Setting (Healthcare, Laboratory, Aviation, Military, Industry) Brief Description of Team Type and Participants (e.g., “undergraduate students assigned to intervention or control”) Categorized Teamwork Dimension (“Reflection,” “Preparation,” “Execution”) Categorized Criterion Measure (“Performance,” “General Execution,” “Conflict Management,” “General Teamwork,” etc.)

Table 14: Review of Task Descriptions in Teamwork Meta-Analyses

No.	Citation	Descriptions and Features of Tasks Provided
17	(Mullen et al., 1991)	N/A (Did Not Report)
18	(Riedl et al., 2021)	Detailed Description of Task and McGrath Categorization
19	(Salas et al., 2008)	Study Setting (*scale not reported*) Task Interdependence (*scale not reported*)
20	(Schmutz et al., 2019)	Routine or Non-Routine
21	(Weber & Hertel, 2007)	Brief Description and Steiner Categorization *Description Examples:* “Physical persistence task (bar-holding paradigm);” “Cognitive maximizing task in PC-supported environment” Steiner Categories are (1) Conjunctive, (2) Additive, or (3) Coactive

Table 14. A summary of the 21 reviewed meta-analyses on teamwork, in which we coded the descriptions or dimensions of the task reported in each meta-analysis. We found that the dimensions reported are very heterogeneous, and the level of detail in describing the task also varies widely, from no information at all to a relatively detailed paragraph.

Appendix I: Task Space Robustness Checks

To investigate the extent to which our findings are robust to changes in the Task Space, we conduct two analyses that estimate the following:

- The extent to which adding noise to columns in the Task Space changes our substantive findings (i.e., our ability to use the Task Space to predict outcomes; Section 4.3 in the main text);
- The extent to which using different subsets of task columns changes the relationship between the tasks in the space (i.e., how consistently do we find that Tasks A and B are "similar?")

I.1. Robustness of Substantive Findings with Noise Added to Task Columns

In our first robustness check, we add noise to the task dimensions, then replicate our primary analysis from Section 4.3 (in which we use ElasticNet models to predict strong and weak group advantage, training on data from Waves 1-2 and evaluating on Wave 3).

Specifically, we add normally distributed noise to task columns in the training data, then compute the RMSE by using the "perturbed" columns to predict the dependent variables (strong and weak group advantage) in the test data. This simulates a scenario in which there is measurement error in the training data, but the prediction target is still determined by the "true," unperturbed values.

Our analyses vary the amount of noise added along two axes:

- **Magnitude of Noise:** We vary the extent of noise by changing the standard deviation of the noise distribution (σ). We present five levels of σ: 0 (no noise), 0.25, 0.5, 0.75, and 1. Given that task dimensions are all in the range (0, 1), these values represent a substantial amount of noise.
- **Number of Columns Affected by Noise:** We vary the number of columns in the training data affected by the noise, from 0 (no noise added) to 24 (all columns).

Figure 47 presents the results of this analysis. At low levels of noise ($\sigma = 0.25$), we see that the Task Space outperforms both of the *unperturbed* McGrath baselines even after perturbing all 24 task columns. At high levels of noise ($\sigma = 1$), we see that the Task Space underperforms the McGrath Subspace baseline, but consistently outperforms McGrath Categorical. Thus, while adding noise unsurprisingly does degrade the quality of our predictions, our overall finding—that the Task Space is useful for making out-of-sample predictions about group advantage, and represents an improvement over using task categories—is robust to measurement error.

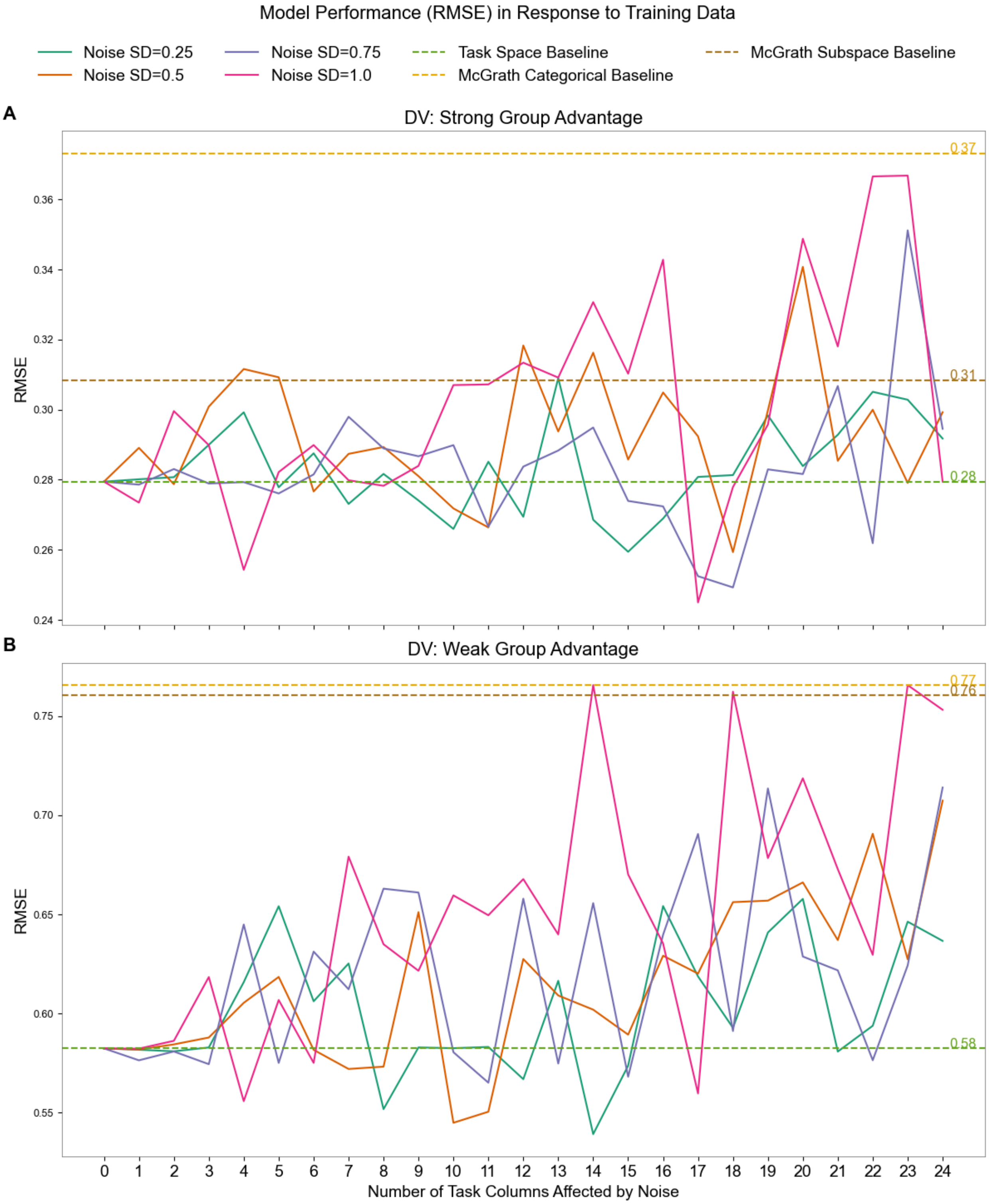


Figure 47. Robustness of model performance (RMSE) to adding noise to training data columns. Panel A shows results for strong group advantage, and Panel B shows results for weak group advantage. The *x*-axis shows the number of task columns in the training data affected by noise; the *y*-axis shows the resulting root mean squared error (RMSE) when predicting on the unperturbed test data. Dotted lines represent the original (unperturbed) results from the Task Space and McGrath baselines.

I.2. Consistency of Task Relationships with Different Subsets of Dimensions

Because of its concatenation-based design, the Task Space embeds natural "redundancies," in which the same general construct may be measured in one or more different ways. For example, Type 2 (Generate) asks raters to make a binary evaluation about whether a task involves brainstorming ("Is this a 'generation' or 'brainstorming' task?"), while Creativity Input asks raters to evaluate the extent to which creativity is required on a continuous scale ("What fraction of the effort required for this task is creative thinking?").

We exploit this natural redundancy in our second robustness check. At a high level, our second test examines the extent to which the Task Space consistently measures relationships between pairs of tasks, even when the number of dimensions is gradually dropped.

Specifically, we examine a task triad, A, B, and C, and compute the difference in cosine distance between (A, B) and (A, C) as follows:

$$\textit{difference} = \textit{cosine\_distance}(A, B) - \textit{cosine\_distance}(A, C)$$

If the difference is a positive value, then (A, C) are closer together than (A, B); if the difference is a negative value, then (A, B) are closer together than (A, C).

We then randomly sample dimensions from the Task Space, and compute the extent which the signs of different samples agree. Specifically, for a given triad of tasks and a subspace of d dimensions, let k be the number of trials (for randomly drawing a feature set), m be the modal sign across the k trials, and si be the sign of the ith trial. Then the agreement, A_d, for the three tasks can be defined as:

$$A_d = \frac{\sum_{1}^{k}(s_i = m)}{k}$$

Semantically, this can be interpreted as: *if k people use different random subsets of the Task Space, how consistently would they find that A and B are closer than A and C?* If the resulting value of agreement (A_d) is high, this would mean that it does not matter which *d* dimensions a researcher selects from the task space in order to make judgments about task similarity. In other words, different subspaces are "consistent."

We conduct this analysis progressively, from dropping one dimension (*d* = 23), to dropping all but two dimensions. *k* is an arbitrary hyperparameter that we set to 10, although the analysis can be repeated for any number of trials. At each level of *d*, we analyze all possible combinations of 3 tasks (102 choose 3, or 171,700) and obtain the average value of agreement (A_d).

Our results are presented in Figure 48. Our findings demonstrate that the Task Space is, indeed, robust to changes in the number and operationalizations of its dimensions; at one extreme, randomly selecting just 2 dimensions will still consistently characterize the relative relationships in a task triad 64% of the time. Of course, as the number of dimensions selected approaches the full 24-dimensional space, this value monotonically approaches 100%. However, we find it highly encouraging that selecting just 7 dimensions will yield more than 80% agreement — suggesting that individuals using different subsets of the space are very likely to obtain consistent results.

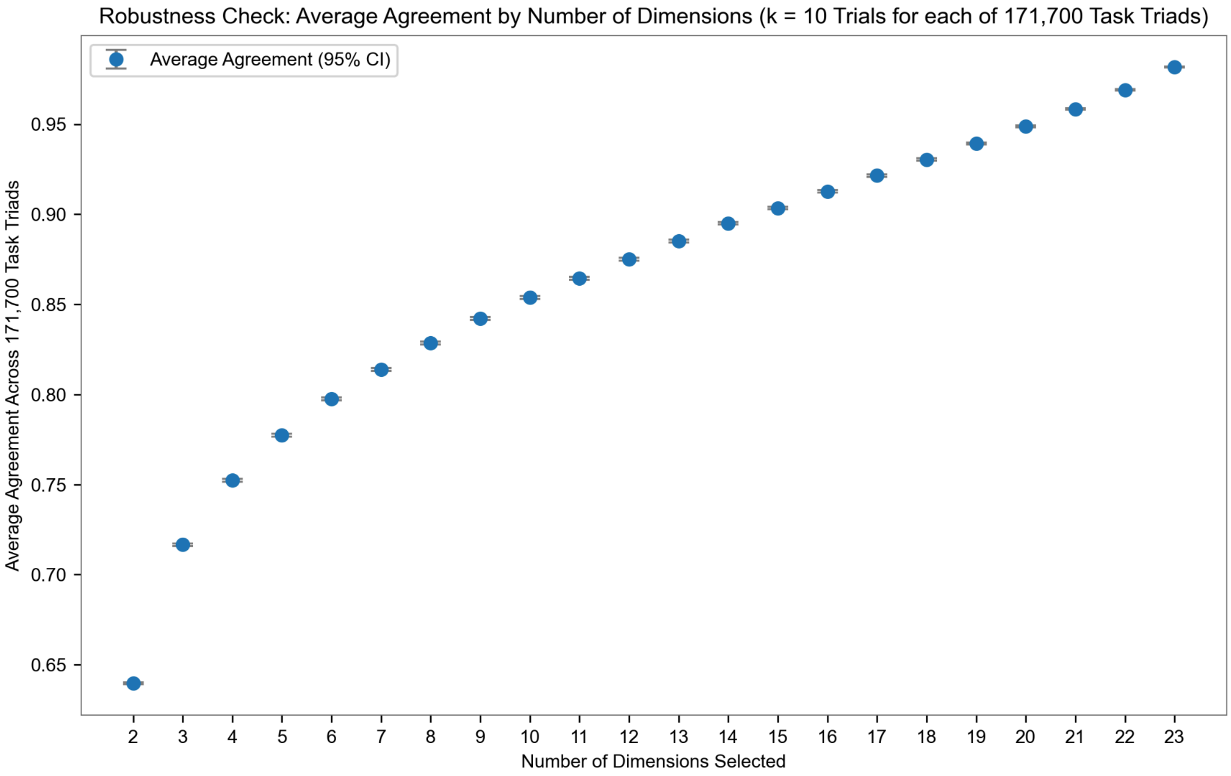


Figure 48. Results of the second robustness check. The x-axis represents the number of dimensions (d) randomly selected from the Task Space; for each possible combination of selecting three tasks (102 choose 3, or 171,700), we generate ten d-dimensional spaces and evaluate the percentage of them that agree that "Task A is more similar to Task B than Task C" (using our measure of cosine similarity). The y-axis shows the average agreement over all 171,700 trials for a given dimension d, with 95% confidence intervals.

These findings are remarkably robust when compared to humans' holistic judgments of whether two items are similar or different. For example, Tversky (1977)'s classic choice set studies find that human perceptions of similarity between objects can be easily influenced by contextual factors. When asked to compare a target country (Austria) to a choice set (Sweden, Hungary, and a distractor), participants in one condition judged that Austria is most similar to Sweden (58% to 42%) and participants in the other judged the reverse (81% to 19%)[3]. By decomposing a task into its constituent dimensions and applying a quantitative metric, our method is far less prone to bias and more robust than the typical approach of qualitatively adjudicating whether a given task belongs to a given category.

[3] In the original study, Tversky provided subjects with two sets of three countries, and asked them to select which one was most similar to the target country (Austria). In Set 1 (Sweden, Hungary, Poland), 49% chose Sweden and 36% chose Hungary; in Set 2 (Sweden, Hungary, Norway), 14% chose Sweden and 60% chose Hungary. After removing votes for the "distractors" (Poland and Norway), we compute the proportion of the remaining votes for "Austria is more similar to Sweden" against "Austria is more similar to Hungary" to obtain the above values.

References

Abimbola, G. A. (2006). *Effects of Task Structure on Group Problem Solving* (thesis). University of Waterloo, Waterloo, Ontario, Canada.

Adams, G. S., Converse, B. A., Hales, A. H., & Klotz, L. E. (2021). People systematically overlook subtractive changes. *Nature*, *592*(7853), 258–261. https://doi.org/10.1038/s41586-021-03380-y

Aggarwal, I., & Woolley, A. W. (2013). Do you see what I see? the effect of members' cognitive styles on team processes and errors in task execution. *Organizational Behavior and Human Decision Processes*, *122*(1), 92–99. https://doi.org/10.1016/j.obhdp.2013.04.003

Almaatouq, A., Alsobay, M., Yin, M., & Watts, D. J. (2021). Task Complexity Moderates Group Synergy. *Proceedings of the National Academy of Sciences*, *118*(36). https://doi.org/10.1073/pnas.2101062118

Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P. M., Moussaid, M., & Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, *117*(21), 11379–11386. https://doi.org/10.1073/pnas.1917687117

Arthur, W., Strong, M. H., Jordan, J. A., Williamson, J. E., Shebilske, W. L., & Regian, J. W. (1995). Visual attention: Individual differences in training and predicting complex task performance. *Acta Psychologica*, *88*(1), 3–23. https://doi.org/10.1016/0001-6918(94)e0055-k

Babcock, L., Recalde, M. P., Vesterlund, L., & Weingart, L. (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, *107*(3), 714–747. https://doi.org/10.1257/aer.20141734

Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L., & Girshick, R. (2019). 33rd Conference on Neural Information Processing Systems (NeurIPS).

Baltes, B. B., Dickson, M. W., Sherman, M. P., Bauer, C. C., & LaGanke, J. S. (2002). Computer-mediated communication and group decision making: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 87(1), 156-179.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42(2), 241-251.

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

Bell, S. T., Villado, A. J., Lukasik, M. A., Belau, L., & Briggs, A. L. (2011). Getting specific about demographic diversity variable and team performance relationships: A meta-analysis. *Journal of Management*, 37(3), 709-743.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Bernstein, E., Shore, J., & Lazer, D. (2018). How intermittent breaks in interaction improve collective intelligence. *Proceedings of the National Academy of Sciences*, *115*(35), 8734–8739. https://doi.org/10.1073/pnas.1802407115

Bhatia, N., & Gunia, B. C. (2018). "I was going to offer $10,000 but…": The effects of Phantom Anchors in negotiation. *Organizational Behavior and Human Decision Processes*, *148*, 70–86. https://doi.org/10.1016/j.obhdp.2018.06.003

Bond, C. F., & Titus, L. J. (1983). Social facilitation: a meta-analysis of 241 studies. *Psychological Bulletin*, 94(2), 265.

Bond, Jr, C. F., & Van Leeuwen, M. D. (1991). Can a part be greater than the whole? On the relationship between primary and meta-analytic evidence. *Basic and Applied Social Psychology*, 12(1), 33-40.

Bonner, B. L., Baumann, M. R., & Dalal, R. S. (2002). The effects of member expertise on group decision-making and performance. *Organizational Behavior and Human Decision Processes*, *88*(2), 719–736. https://doi.org/10.1016/s0749-5978(02)00010-9

Bornstein, G. (2003). Intergroup conflict: Individual, group, and collective interests. *Personality and Social Psychology Review*, *7*(2), 129–145. https://doi.org/10.1207/s15327957pspr0702_129-145

Bornstein, G., & Yaniv, I. (1998). Individual and Group Behavior in the Ultimatum Game: Are Groups More "Rational" Players? *Experimental Economics*, *1*(1), 101–108. https://doi.org/10.1023/a:1009914001822

Bowers, C. A., Pharmer, J. A., & Salas, E. (2000). When member homogeneity is needed in work teams: A meta-analysis. *Small Group Research*, 31(3), 305-327.

Brewer, M. B., & Chen, Y. R. (2007). Where (who) are collectives in collectivism? Toward conceptual clarification of individualism and collectivism. *Psychological Review*, 114(1), 133.

Brittain, J. W., & Sitkin, S. B. (1986). Carter Racing. http://learnmoore.org/MBA205/05CarterRacing.pdf

Butchibabu, A., Sparano-Huiban, C., Sonenberg, L., & Shah, J. (2016). Implicit coordination strategies for effective team communication. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(4), 595–610. https://doi.org/10.1177/0018720816639712

Camerer, C. F. (1997). Progress in behavioral game theory. *Journal of Economic Perspectives*, *11*(4), 167–188. https://doi.org/10.1257/jep.11.4.167

Chen, F., & Samroengraja, R. (2009). The stationary beer game. *Production and Operations Management*, *9*(1), 19–30. https://doi.org/10.1111/j.1937-5956.2000.tb00320.x

Choi, H.-S., & Thompson, L. (2005). Old wine in a new bottle: Impact of membership change on group creativity. *Organizational Behavior and Human Decision Processes*, *98*(2), 121–132. https://doi.org/10.1016/j.obhdp.2005.06.003

Cohen, T., Leonardelli, G. J., & Thompson, L. (2010). The agreement bias in negotiation: Teams facilitate impasse. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1612404

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, *100*(2), 193–201. https://doi.org/10.1016/j.obhdp.2005.10.001

Dawes, R. M. (1980). Social Dilemmas. *Annual Review of Psychology*, *31*, 169–193. https://doi.org/https://www.cmu.edu/dietrich/sds/docs/dawes/social-dilemmas.pdf

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: a meta-analysis. *Journal of Applied Psychology*, 95(1), 32.

De Dreu, C. K., & Weingart, L. R. (2003). Task versus relationship conflict, team performance, and team member satisfaction: a meta-analysis. *Journal of Applied Psychology*, 88(4), 741.

Dennis, A. R., & Williams, M. L. (2005). A meta-analysis of group side effects in electronic brainstorming: More heads are better than one. *International Journal of e-Collaboration* (IJeC), 1(1), 24-42.

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

Engel, D., Woolley, A. W., Jing, L. X., Chabris, C. F., & Malone, T. W. (2014). Reading the mind in the eyes or reading between the lines? theory of mind predicts collective intelligence equally well online and face-to-face. *PLoS ONE*, *9*(12). https://doi.org/10.1371/journal.pone.0115212

Finlay, F., Hitch, G. J., & Meudell, P. R. (2000). Mutual inhibition in collaborative recall: Evidence for a retrieval-based account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1556–1567. https://doi.org/10.1037/0278-7393.26.6.1556

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25-42.

Frick, B., Rose, A., & Kolle, A. (2017, February 2). *Gender Diversity is Detrimental to Team Performance: Evidence from a Field Experiment*. Working Papers Dissertations. https://ideas.repec.org/p/pdn/dispap/23.html

Gully, S. M., Devine, D. J., & Whitney, D. J. (1995). A meta-analysis of cohesion and performance: Effects of level of analysis and task interdependence. *Small Group Research*, 26(4), 497-520.

Gully, S. M., Incalcaterra, K. A., Joshi, A., & Beaubien, J. M. (2002). A meta-analysis of team-efficacy, potency, and performance: interdependence and level of analysis as moderators of observed relationships. *Journal of Applied Psychology*, 87(5), 819.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, *95*(1), 384–394. https://doi.org/10.1257/0002828053828662

Hackman, J. R., Jones, L. E., & McGrath, J. E. (1967). A set of dimensions for describing the general properties of group-generated written passages. *Psychological Bulletin*, *67*(6), 379–390. https://doi.org/10.1037/h0024647

Halfhill, T., Sundstrom, E., Lahner, J., Calderone, W., & Nielsen, T. M. (2005). Group personality composition and group effectiveness: An integrative review of empirical research. *Small Group Research*, 36(1), 83-105.

Hemenway, D., Moore, R., & Whitney, J. (1987). Teaching tools: The oligopoly game. *Economic Inquiry*, *25*(4), 727–730. https://doi.org/10.1111/j.1465-7295.1987.tb00776.x

Hertz, U., Romand-Monnier, M., Kyriakopoulou, K., & Bahrami, B. (2016). Social influence protects collective decision making from equality bias. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(2), 164–172. https://doi.org/10.1037/xhp0000145

Isenberg, D. J. (1981). Some effects of time-pressure on vertical structure and decision-making accuracy in small groups. *Organizational Behavior and Human Performance*, *27*(1), 119–134. https://doi.org/10.1016/0030-5073(81)90042-8

Jiang, L.-L., Perc, M., & Szolnoki, A. (2013). If cooperation is likely punish mildly: Insights from economic experiments based on the Snowdrift Game. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0064677

Johansson, N. O., Andersson, J., & Rönnberg, J. (2005). Compensating strategies in collaborative remembering in very old couples. *Scandinavian Journal of Psychology*, *46*(4), 349–359. https://doi.org/10.1111/j.1467-9450.2005.00465.x

Kearns, M., Suri, S., & Montfort, N. (2006). An experimental study of the coloring problem on human subject networks. *Science*, *313*(5788), 824–827. https://doi.org/10.1126/science.1127207

Kelly, J. R., & Karau, S. J. (1999). Group decision making: The effects of initial preferences and time pressure. *Personality and Social Psychology Bulletin*, *25*(11), 1342–1354. https://doi.org/10.1177/0146167299259002

Kennedy, N. S. (2009). Wolf, Goat, and Cabbage: An Analysis of Students' Roles and Cognitive and Metacognitive Behaviors in Small Group Collaborative Problem-Solving. *Analytic Teaching*, *29*(1), 39–52. https://doi.org/https://journal.viterbo.edu/index.php/atpp/article/view/1024

Krafft, P. M., Hawkins, R. X. D., Pentland, A. "Sandy," Goodman, N. D., & Tenenbaum, J. B. (2015). Emergent Collective Sensing in Human Groups. https://people.csail.mit.edu/pkrafft/papers/krafft-et-al-2015-emergent.pdf

Laughlin, P. R., & Shippy, T. A. (1983). Collective induction. *Journal of Personality and Social Psychology*, *45*(1), 94–100. https://doi.org/10.1037/0022-3514.45.1.94

Laughlin, P. R., Hatch, E. C., Silver, J. S., & Boh, L. (2006). Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and Social Psychology*, *90*(4), 644–651. https://doi.org/10.1037/0022-3514.90.4.644

Leavitt, H. J. (1951). Some effects of certain communication patterns on group performance. *The Journal of Abnormal and Social Psychology*, *46*(1), 38–50. https://doi.org/10.1037/h0057189

LePine, J. A., Piccolo, R. F., Jackson, C. L., Mathieu, J. E., & Saul, J. R. (2008). A meta‐analysis of teamwork processes: Tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel Psychology*, 61(2), 273-307.

Littlepage, G. E. (1991). Effects of group size and task characteristics on group performance: A test of Steiner’s model. *Personality and Social Psychology Bulletin*, *17*(4), 449–456. https://doi.org/10.1177/0146167291174014

London, K., & Nunez, N. (2000). The effect of jury deliberations on jurors’ propensity to disregard inadmissible evidence. *Journal of Applied Psychology*, *85*(6), 932–939. https://doi.org/10.1037/0021-9010.85.6.932

Lorge, I., & Solomon, H. (1959). Individual performance and group performance in problem solving related to group size and previous exposure to the problem. *The Journal of Psychology*, *48*(1), 107–114. https://doi.org/10.1080/00223980.1959.9916346

Maier, N. R. (1930). Reasoning in humans. I. On Direction. *Journal of Comparative Psychology*, *10*(2), 115–143. https://doi.org/10.1037/h0073232

Mao, A., Mason, W., Suri, S., & Watts, D. J. (2016). An experimental study of team size and performance on a complex task. *PLOS ONE*, *11*(4). https://doi.org/10.1371/journal.pone.0153048

Marks, M. A., Sabella, M. J., Burke, C. S., & Zaccaro, S. J. (2002). The impact of cross-training on team effectiveness. *Journal of Applied Psychology*, *87*(1), 3–13. https://doi.org/10.1037/0021-9010.87.1.3

Marlow, S. L., Lacerenza, C. N., Paoletti, J., Burke, C. S., & Salas, E. (2018). Does team communication represent a one-size-fits-all approach?: A meta-analysis of team communication and performance. *Organizational Behavior and Human Decision Processes*, 144, 145-170.

Mason, W., & Watts, D. J. (2011). Collaborative learning in Networks. *Proceedings of the National Academy of Sciences*, *109*(3), 764–769. https://doi.org/10.1073/pnas.1110069108

Mayo, A. T., Woolley, A. W., & Chow, R. M. (2020). Unpacking participation and influence: Diversity’s countervailing effects on expertise use in groups. *Academy of Management Discoveries*, *6*(2), 300–319. https://doi.org/10.5465/amd.2018.0044

McEwan, D., Ruissen, G. R., Eys, M. A., Zumbo, B. D., & Beauchamp, M. R. (2017). The effectiveness of teamwork training on teamwork behaviors and team performance: a systematic review and meta-analysis of controlled interventions. *PloS One*, 12(1), e0169604.

McLeod, P. L., Baron, R. S., Marti, M. W., & Yoon, K. (1997). The eyes have it: Minority influence in face-to-face and computer-mediated group discussion. *Journal of Applied Psychology*, *82*(5), 706–718. https://doi.org/10.1037/0021-9010.82.5.706

Moussaïd, M., Herzog, S. M., Kämmer, J. E., & Hertwig, R. (2017). Reach and speed of judgment propagation in the Laboratory. *Proceedings of the National Academy of Sciences*, *114*(16), 4117–4122. https://doi.org/10.1073/pnas.1611998114

Mullen, B., Johnson, C., & Salas, E. (1991). Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and Applied Social Psychology*, 12(1), 3-23.

Naber, M., Vaziri Pashkam, M., & Nakayama, K. (2013). Unintended imitation affects success in a competitive game. *Proceedings of the National Academy of Sciences*, *110*(50), 20046–20050. https://doi.org/10.1073/pnas.1305996110

Nielsen, J. (2000). Why you only need to test with 5 users. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, *118*(25). https://doi.org/10.1073/pnas.2022340118

Overbeck, J. R., Neale, M. A., & Govan, C. L. (2010). I feel, therefore you act: Intrapersonal and interpersonal effects of emotion on negotiation as a function of social power. *Organizational Behavior and Human Decision Processes*, *112*(2), 126–139. https://doi.org/10.1016/j.obhdp.2010.02.004

Rabin, M. (1993). Incorporating fairness into Game Theory and Economics. *The American Economic Review*, *83*(5), 1281–1302. https://doi.org/https://www.jstor.org/stable/2117561

Riedl, C., Kim, Y. J., Gupta, P., Malone, T. W., & Woolley, A. W. (2021). Quantifying collective intelligence in human groups. *Proceedings of the National Academy of Sciences*, 118(21), E.2005737118.

Salas, E., DiazGranados, D., Klein, C., Burke, C. S., Stagl, K. C., Goodwin, G. F., & Halpin, S. M. (2008). Does team training improve team performance? A meta-analysis. *Human Factors*, 50(6), 903-933.

SanPietro, L. (2019, June 11). *Negotiation games to develop Dispute Resolution Skills*. PON. https://www.pon.harvard.edu/daily/teaching-negotiation-daily/how-negotiation-games-can-help-you-develop-skills-to-resolve-business-and-commercial-disputes/

Schmutz, J. B., Meier, L. L., & Manser, T. (2019). How effective is teamwork really? The relationship between teamwork and performance in healthcare teams: a systematic review and meta-analysis. *BMJ Open*, 9(9), e028280.

Schulz-Hardt, S., Brodbeck, F. C., Mojzisch, A., Kerschreiter, R., & Frey, D. (2006). Group decision making in hidden profile situations: Dissent as a facilitator for decision quality. *Journal of Personality and Social Psychology*, *91*(6), 1080–1093. https://doi.org/10.1037/0022-3514.91.6.1080

Sellier, A.-L., Scopelliti, I., & Morewedge, C. K. (2019). Debiasing training improves decision making in the field. *Psychological Science*, *30*(9), 1371–1379. https://doi.org/10.1177/0956797619861429

Shaw, M. E. (1963). Scaling Group Tasks: A Method for Dimensional Analysis: (532082008-001). *American Psychological Association*. https://doi.org/10.1037/e532082008-001

Shirado, H., Crawford, F. W., & Christakis, N. A. (2020). Collective communication and behaviour in response to uncertain 'danger' in network experiments. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *476*(2237). https://doi.org/10.1098/rspa.2019.0685

Shore, J., Bernstein, E., & Lazer, D. (2015). Facts and figuring: An experimental investigation of network structure and performance in information and solution spaces. *Organization Science*, *26*(5), 1432–1446. https://doi.org/10.1287/orsc.2015.0980

Silver, I., Mellers, B. A., & Tetlock, P. E. (2021). Wise teamwork: Collective confidence calibration predicts the effectiveness of group discussion. *Journal of Experimental Social Psychology*, *96*, 104157. https://doi.org/10.1016/j.jesp.2021.104157

Sommer, S. C., Bendoly, E., & Kavadias, S. (2020). How do you search for the best alternative? experimental evidence on search strategies to solve complex problems. *Management Science*, *66*(3), 1395–1420. https://doi.org/10.1287/mnsc.2018.3247

Straub, V. J., Tsvetkova, M., & Yasseri, T. (2023). The cost of coordination can exceed the benefit of collaboration in performing complex tasks. *Collective Intelligence*, *2*(2). https://doi.org/10.1177/26339137231156912

Straus, S. G. (1999). Testing a typology of tasks: An empirical validation of McGrath's (1984) group task circumplex. Small Group Research, 30(2), 166-187.

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, *223*(5), 96–102. https://doi.org/10.1038/scientificamerican1170-96

Takahashi, M. (2007). Does collaborative remembering reduce false memories? *British Journal of Psychology*, *98*(1), 1–13. https://doi.org/10.1348/000712606x101628

Tomassini, M., & Antonioni, A. (2020). Public goods games on coevolving social network models. *Frontiers in Physics*, *8*. https://doi.org/10.3389/fphy.2020.00058

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147-168.

Traeger, M. L., Strohkorb Sebo, S., Jung, M., Scassellati, B., & Christakis, N. A. (2020). Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences*, *117*(12), 6370–6375. https://doi.org/10.1073/pnas.1910402117

Tversky, A. (1977). Features of Similarity. *Psychological Review*, 84(4), 327.

Van Huyck, J. B., Wildenthal, J. M., & Battalio, R. C. (2002). Tacit Cooperation, strategic uncertainty, and coordination failure: Evidence from repeated dominance solvable games. *Games and Economic Behavior*, *38*(1), 156–175. https://doi.org/10.1006/game.2001.0860

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161

Watts, D. J. (2014). Common sense and sociological explanations. American Journal of Sociology, 120(2), 313-351.

Weber, B., & Hertel, G. (2007). Motivation gains of inferior group members: a meta-analytical review. *Journal of Personality and Social Psychology*, 93(6), 973.

Weidmann, B., & Deming, D. (2020). *Team Players: How Social Skills Improve Group Performance*. https://doi.org/10.3386/w27071

Whiting, M. E., Blaising, A., Barreau, C., Fiuza, L., Marda, N., Valentine, M., & Bernstein, M. S. (2019). Did it have to end this way? *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–23. https://doi.org/10.1145/3359311

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688. https://doi.org/10.1126/science.1193147

Yahosseini, K. S., & Moussaïd, M. (2020). Comparing groups of independent solvers and transmission chains as methods for collective problem-solving. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-59946-9

Yetton, P., & Bottger, P. (1983). The relationships among group size, member ability, social decision schemes, and performance. *Organizational Behavior and Human Performance*, *32*(2), 145–159. https://doi.org/10.1016/0030-5073(83)90144-7