1. Clue generation procedure (Pretest)
   1. Clues are constructed from a bank of concepts (11 Stolen Objects, 11 Crime Scenes, 33 Names, 22 descriptions, 22 articles of clothing, 22 tools, and 22 vehicles) and set of relationships (e.g. {Name} owns a {vehicle}, {A witness thought they saw {stolen object} in {vehicle}) that forms a complete network between all concepts. These are randomly shuffled such that different clues can be generated for each game.
   2. The bank of concepts was constructed by starting with a pool of candidate concepts approximately 3-4x the size of concepts eventually used. A pretest survey was conducted in which Amazon Mechanical Turk workers were asked to assess how likely a number of concepts was to have been used in a generic burglary. Individuals saw a subset of the concepts and were asked to give their gut reactions using a slider from Extremely Unlikely to Extremely Likely, as illustrated in Fig 10 below. In total, 139 participants rated each of 403 candidate concepts between 20 and 30 times. Participants in the pretest were paid $1.25 for a task which took each participant an average of about 4 minutes.

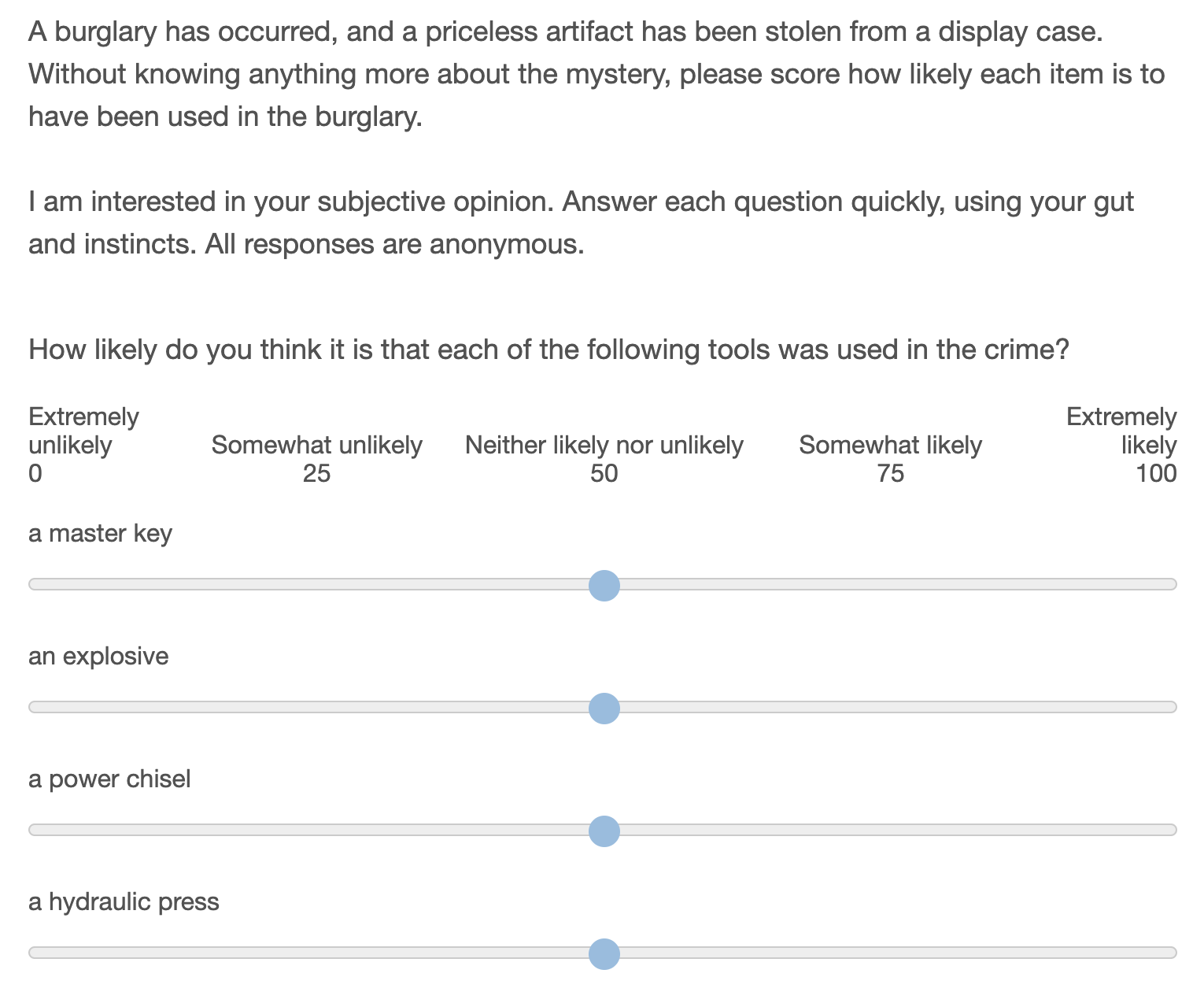
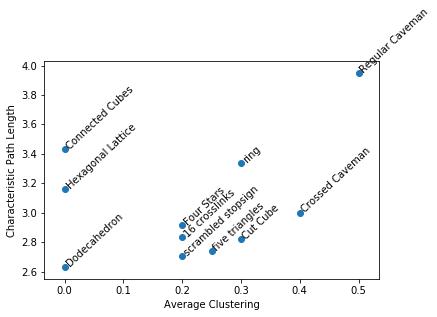


Fig 10: Pretest the perceived likelihood of each item being used in a crime

* 1. The pool of candidate names in the pretest represents the subset of the 200 most popular last names in the United States with a racial composition of between 50% and 80% ‘White’, as recorded in the 2000 US census. This selection is made to minimize the possibility of racial biases in the results. Additionally, names which are also common first names are excluded (e.g. “Stewart” or “Ross”) as are names which also serve as descriptors or adjectives in other clues (e.g. “Green”, “White”, or “Young”).
  2. The remaining candidate concepts were written such that they would be as independent from one another as possible (e.g. I do not include both “a fat man” and “an overweight man” as these are synonymous, nor both “an old man” and “a man with grey hair” as these are perceived to go together.)
  3. From the pretest results, I selected a subset of concepts that are perceived to be as likely as one another to be used in a burglary. (This helps to ensure that we do not see games in which all participants adopt “a set of lock picks” as a tool in the burglary, and reject “a machette”, just because lock picks are easier to imagine being used in a burglary.) The final selection was made by taking the subset of beliefs that minimized the difference in mean value of pretest survey responses when responses are normalized for each individual, and cross-checking against the means of the raw responses. 11 concepts were selected for each time the concept (name, vehicle, etc.) is used in a game.
  4. A similar pretest survey was conducted to select ‘spur’ clue concepts from a pool of candidates.

1. Display considerations
   1. Presentation of social information**:** All at once display
      1. A ‘scrolling feed’ type information display has recency and primacy effects, and opens questions about how we should aggregate social information from multiple players. Showing all information at once, in the order that it is sorted by the neighbor, eliminates the effect of alternate ordering sequences.
   2. Number of neighbors**:** 3
      1. The number of neighbors is limited by the size of the screen and an individual’s ability to process information. The minimum number of neighbors for a non-trivial social network is 3, and is also a reasonable number for managing the cognitive load in the game.
   3. Number of starting clues**:** 4
      1. Fewer starting clues are preferred for minimizing cognitive load on individuals. With three neighbors, individuals see 16 clues all at once when they start playing. This takes about 30 seconds to read through and understand. The next increment (5 starting clues) gives 20 items for an individual to process at game start, which starts to be cognitively overwhelming.
   4. Number of players**:** 20
      1. Larger numbers of players are better for generalizability and seeing an effect size. Smaller numbers mean we can afford more replications. There needs to be enough players that the mean shortest-path-length is greater than two, to realistically represent multi-stage diffusion.
   5. Network shape**:** Dodecahedron and regular connected caveman (k=5)
      1. Eleven symmetric candidate networks were evaluated with n=20 and degree=3. Of this set, the Dodecahedral network minimizes the average shortest path between individuals with no network clustering, and represents a social network we should expect to exhibit low polarization *a priori*. A regular connected caveman network maximizes the characteristic path length and exhibits strong clustering, and so we expect to exhibit more polarization *a priori.* Descriptions of each of these networks are included in the preregistration code.
   6. Number of unique clues in the game**:** 78
      1. From an information diversity perspective, more clues is better. With 4 starting clues and 20 players, we can have up to 80 unique clues in the game. 13 nodes yields 78 clues, and the two spots remaining can be filled with the given link between the crime scene and the stolen object.
   7. Number of times each clue is represented: 1\*
      1. Each clue should be represented an equal number of times so as not to bias the network to one particular outcome.
      2. \*The ‘given’ clue that the object was stolen from the crime scene is included 3 times to fill out the 80 slots in the game.
   8. Length of game: 8 minutes
      1. Pilot trials were conducted with durations of 5 and 8 minutes. It was observed that participants remained engaged for 8 minutes, and felt rushed with 5.
   9. Survey format: Empty sliders
      1. Rather than force individuals to make a discrete choice between suspects/vehicles etc., a slider allows individuals to assess a degree of confidence in their assessment of the solution to the mystery.
2. Choice of Measures
   1. Self-report similarity: Pearson Correlation
      1. Correlation is a natural measure when we have a fixed number of continuous measures of each subject, as is the case in the self-report, and there is precedence for this use in recent literature *(4)*. It is useful to have a measure with a fixed range (-1,1) and which is readily interpretable.
   2. Behavioral similarity: Phi coefficient
      1. The phi coefficient corresponds to Pearson correlation when measures are binary, and has the same interpretable (-1,1) range. This is appropriate for a universe in which there are a finite number of beliefs measured, but would be less appropriate as the number of adopted beliefs becomes a very small fraction of the total number of possible beliefs.
      2. Other measures of similarity are present
   3. Polarization
      1. Percent of Variance present in first principal component
         1. This measure corresponds to the notion of “constraint” articulated by Dimaggio et al. *(6)*. In their paper they describe Chronbach’s alpha and the PCA measure both providing similar measures of constraint. I have chosen the PCA measure here as more interpretable and well known among computational social scientists.
      2. 5TH and 95th percentile similarities
         1. There are a number of different measures in the literature that try to capture the notion that with polarization, the most similar individuals become more self-similar, and the least similar individuals move further away from one another. The fact that no single measure has emerged as the leader hints at problems with each. Variance *(see 1,6)* captures heterogeneity between individuals, but not clustering into camps. Kurtosis *(see 1,6)* is predicated on a bimodal distribution. The “gap” statistic *(4)* is one of dozens of ways of assessing the quality of a machine learning clustering algorithm. When the identities of camps are already known the difference of means between groups can be used *(6,7).*
         2. As I do not need to identify the groups themselves, or compare to external datasets, it is sufficient for me to merely report what each of these other measures is trying to approximate: the similarity that is found within groups, and that which is found across groups. As I am only interested in the relative differences between conditions (or for the same population over time) then I can arbitrarily designate a threshold for which comparisons will be considered ‘within-group’ or ‘across groups’. This provides a much more intuitive demonstration of increasing polarization than the measures found in literature.
         3. The closer the chosen thresholds are to the tails of the distribution, the more conservative the claim that the comparisons beyond this threshold are appropriately “within” or “across” groups. At the same time, we need enough samples included in the set to minimize noise due to the finite number of comparisons. In this 20-participant social network, the 95th and 5th percentiles correspond to 10 comparisons between individuals.
3. Recruitment