Research Article

Psychological Science
1–15
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09567976251384640
www.psychologicalscience.org/PS

Invariant Recognition Memory Spaces for Real-World Objects Revealed With Signal-Detection Analysis

Igor Utochkin[1], Daniil Azarov[2], and Daniil Grigorev[3,4]
[1]Institute for Mind and Biology, University of Chicago; [2]Department of Psychological and Brain Sciences and Department of Statistics, Indiana University Bloomington; [3]Department of Psychology, Koç University; and [4]Department of Psychology, Warwick University

Abstract

Recognition memory refers to the process of distinguishing between previously experienced and novel events. Apart from the objective quality of stored memories, recognition depends on the retrieval context produced by all items (foils) presented together with actually memorized targets and causing confusion. Memory models often conceptualize target-foil confusability via distances in psychological spaces where greater confusability originates from shorter interitem distances. We tested whether recognition spaces change when other foils are added to the retrieval context or when target memory strength is changed (N = 1,311 adults). Using signal-detection modeling, we found that separately measured distances, d's, from each foil to the target provide a good linear prediction of those distances for all foils being presented together against that target. Those predictions stay accurate even when the absolute distances are scaled up or down because of a change in memory strength. This suggests strong metric invariance of spaces used for recognition decisions under variable retrieval contexts.

Keywords

recognition memory, visual memory, signal-detection theory, representational spaces, decision making

Received 3/26/24; Revision accepted 9/11/25

Recognition memory, the ability to distinguish between previously experienced and new impressions, is one of the most common and broadly studied memory-retrieval processes. Recognition accuracy depends on how well a stimulus or episode is encoded and stored in memory: The better a target item is remembered, the easier it is to distinguish it from never-presented new foils. Recognition also depends on context created by other items in memory. For example, if we memorize a list of 30 personal names and one country name, the latter will be recognized particularly well (Hunt, 1995). If we study a list of words closely associated with a certain concept, this concept is likely to be falsely recognized even if never studied (Roediger & McDermott, 1995). Typically, people are more prone to mistake a foil for a target if these items are similar (Brady & Störmer, 2024; Migo et al., 2013; Tulving, 1981). In other words, the recognition outcome is a function of both memory strength and target-foil similarity. By carefully taking into account these factors, influential models of recognition memory provide remarkably accurate predictions about the probability of correct recognition and false alarms (Meagher & Nosofsky, 2023; Nosofsky, 1991; Nosofsky et al., 2011; Schurgin et al., 2020).

Apart from those factors, a retrieval situation also bears context variables that can significantly influence recognition outcomes. The list of such variables is quite broad. In the current work, we will specifically talk about the retrieval context —that is, the influence that multiple items can have on recognition of each other if they happen to be relevant parts of the same recognition situation. If a recognition decision involves a choice

Corresponding Author:
Igor Utochkin, University of Chicago, Institute for Mind and Biology
Email: isutochkin@gmail.com

between several options ("Which of these items do I remember?"), the decision would depend on all items presented at the same time. For example, suspect identification by a crime eyewitness usually requires the eyewitness to review a lineup of photos. The ability to recognize the perpetrator would be modulated by the number of alternatives and their mutual similarity (Colloff et al., 2021; Lam & Wixted, 2024; Shen et al., 2023). Although individual item recognition typically involves comparing each item to memory contents, the retrieval context introduced by multiple test options can alter this process by enabling direct comparisons among the options. For example, it may prompt observers to place less weight on shared features and focus instead on more diagnostic, distinctive ones (Migo et al., 2009; Shen et al., 2023; Tulving, 1981; Wixted & Mickes, 2014).

Important ideas about representational changes underlying contextual effects on choices come from research on value-driven preferences, such as a consumer's choice between a pricier laptop with a speedy CPU and a cheaper laptop with a slower CPU. Here contextual effects are particularly pronounced, so that introducing a third item, C, can change or even reverse the preference between A and B, even if C is not offered as an option (see reviews: Busemeyer et al., 2019; Spektor et al., 2021). Decision theories suggest that preference evolves as a dynamic competitive process of comparison between items in a representational space that changes because of various interactions between the items and information loss during comparisons (e.g., Dumbalska et al., 2020; Roe et al., 2001; Usher & McClelland, 2004), which results in context-driven distortions.

In the present study, we tested whether recognition-memory spaces underlying discrimination between old and new items vary across different test contexts. By *recognition-memory space*, we mean locations of test items on a hypothetical psychological dimension underlying the impression of memory match between these items and memory content (see next paragraph for details). The closer the representations on this dimension, the more similar memory-match effects they produce and, therefore, the more confusable the items are. Will the distance between target A and foil B substantially change if a third item C is introduced, thus causing a change in confusability between A and B? Alternatively, do these spaces remain relatively invariant despite the retrieval-context changes, when other foils come into play? Previous research showed considerable consistency in recognition of individual stimuli (old/new judgments) in various populations of observers (Bainbridge et al., 2013; Isola, Parikh, et al., 2011). Because this consistency not only preserves across people but also survives memory-list variations, it is often ascribed to stimulus-intrinsic memorability and can be considered a form of context invariance in memory (Bainbridge, 2020; Kramer et al., 2023). *Memorability* refers to invariance relative to the context of the studied material; the present study investigates invariance relative to the retrieval context.

Statement of Relevance

Recognition memory—the ability to distinguish between old and new experiences—depends not only on objective memory strength for remembered things but also on retrieval context imposed by other items (foils) present at the time of testing (e.g., how many foils there are, or how similar they are to the target). Although the behavioral pattern (distributions of choices between the target and foils) clearly changes as a function of such context, our study shows that the deep psychological structure determining the contribution of each foil to the outcome of a recognition test is remarkably stable. Specifically, we show that separately measured pairwise distances between arbitrary targets and various foils are strongly linearly predictive of those distances in more context-rich situations in which all foils are present together. Our findings suggest the existence of deep psychological invariants governing memory retrieval even when the test context is variable.

To test the degree of recognition-space invariance, we sought to answer two questions. First, are these spaces invariant between observers when their memory is tested with identical target-foil combinations (i.e., in identical retrieval contexts)? Second, to what extent are pairwise distances between a target and various foils measured separately (in a two-alternative forced-choice task, or 2-AFC) preserved when all those foils are presented together against the target (e.g., in a four-alternative forced-choice task, or 4-AFC)? We used signal-detection theory (SDT; Hautus et al., 2021) to obtain distances in the recognition-memory space. SDT links stimulus confusability in a recognition test to the representational separation between the stimuli along a psychological memory-match dimension that determines how strong the impression of a remembered stimulus would be. This separation, termed d', determines the overlap between the distributions of memory-match effects caused by each item on any random trial. In other words, the d'-dependent overlap between the target and foil distributions determines how often the memory-match magnitude of a foil happens to be larger than that of a target, which is exactly the probability of target-foil confusion. This allows one to recover target-foil distances in a recognition space from the observed probabilities of choosing each of the test items; the distances can then be compared across tasks. Previous work has shown that SDT metrics of target-foil confusability at the memory test indeed track psychological distances between test alternatives (Schurgin et al., 2020).

To estimate the degree of invariance, we tested our data against three different models: strong, weak, and no-invariance. *Strong invariance* suggests that, when test context changes, metric relationships between distances in the recognition space stay intact—that is, all d's undergo the same linear transformation. Weak invariance suggests that the context change preserves the rough ordering of distances (a more confusable foil in 2-AFC remains such in 4-AFC), but the distances change to a different extent for each foil (confusability can change dramatically for some foils but only slightly for others). Finally, *no-invariance* suggests that the recognition spaces randomly change between test contexts in such a way that neither distances nor orders are preserved.

We ran three experiments. In Experiment 1, we tested the strength of invariance between 2-AFC and 4-AFC contexts for target-foil spaces consisting of arbitrarily chosen visual objects, some from the same and some from different categories. In Experiment 2, we tested exactly the same spaces as in Experiment 1, but encoding time was shortened. This manipulation aimed to reduce memory strength and cause the overall d' decrement or recognition-space shrinkage. We asked to what extent the "shrunk" target-foil distances (Experiment 2) remained invariant relative to their counterparts from the originally measured spaces (Experiment 1). In Experiment 3, we used the same procedure as in Experiment 1, but on a new stimulus set. In this set, target-foil combinations consisted of the objects of the same category that varied either in identity (e.g., coffee mug A vs. B) or state (full mug vs. empty mug). This allowed us to test the generalizability of invariance to recognition spaces that require finer discrimination.

Research Transparency Statement

General disclosures

Conflicts of interest: The authors declare no conflict of interest. **Funding:** This research received no specific funding and was paid for by the authors' personal funds. **Artificial intelligence:** No artificial-intelligence-assisted technologies were used in this research or the creation of this article. **Ethics:** This research complies with the Declaration of Helsinki and received approval from a local ethics board (University of Chicago, Protocol No. IRB15-1290). **Open Science Framework (OSF):** To facilitate long-term preservation, we registered all OSF files at https://osf.io/bfgnc (study materials) and https://osf.io/7huwa (data and analysis scripts).

Study disclosures

Preregistration: No aspects of the research were preregistered. **Materials:** All study materials are publicly available (https://osf.io/ectdx/files/osfstorage). Image stimuli used in the experiments are parts of publicly available stimulus sets provided by another laboratory (Experiments 1 and 2: https://bradylab.ucsd.edu/stimuli/ObjectCategories.zip; Experiment 3: https://bradylab.ucsd.edu/stimuli/StateExemplar.zip). **Data:** All primary data are publicly available (https://osf.io/5whxp/files/osfstorage). **Analysis scripts:** All analysis scripts are publicly available (https://osf.io/5whxp/files/osfstorage). **Computational reproducibility:** The computational reproducibility of the results in the main article (but not the supplementary materials) has been independently confirmed by the journal's STAR team.

General Method

Sample-size note

The sample size for the purpose of our main statistical analysis was defined as the number of tested target-foil spaces because we were interested in correlations (taken as measures of space invariance) between performance rates provided by these combinations in different subsamples of people and under different recognition-task conditions. We tested 120 target-foil spaces (120 unique memory targets, each combined with three different foils yielding 360 target-foil discriminability indices d'), which allowed us to detect a significant Spearman's ρ correlation of at least 0.20 using the permutation test (given $\alpha = .05$ and $1 - \beta = .95$). For each target-foil test combination, we aimed to collect data from at least 100 participants. From the point of view of an individual test target-foil combination, each participant contributed a single trial. This allowed us to estimate a d' for each test combination (mean error: 0.14 – 0.37 depending on the d' magnitude range). Each participant completed 120 test trials, thus contributing to d' estimation of 120 pairwise target-foil combinations (2-AFC tasks) or 360 such combinations (4-AFC task).

Participants

Participants were recruited through the online platform Prolific.com. No prescreening criteria were applied. The participants were paid £9/hour. The median time to complete the experiment varied between 8 and 12 min.

Before the beginning of the experiments, participants gave online informed consent. The study design and the consent form were approved by the Ethics Committee of the University of Chicago. We aimed to collect data from at least 100 participants per group (each group used a different version of a memory test). Because the experiment included four versions of the memory test (three 2-AFC and one 4-AFC), the overall intended number of participants was at least 400 per experiment. If participants' performance was below a certain level (d' = 0.35; proportion of correct answers below 0.6 in 2-AFC and below 0.35 in 4-AFC) or at the ceiling (proportion of correct answers = 1), their data were not included in the analysis, and data collection continued until reaching the desired sample size. Overall, 426, 456, and 429 people participated in Experiments 1, 2, and 3, respectively. The final samples after the data exclusion included 401 (94%), 413 (91%), and 402 (94%) participants.

Stimuli and procedure

The experiment was run on Pavlovia.org. We restricted the allowed device type to computers or laptops. For Experiments 1 and 2, we randomly selected 120 basic object categories (e.g., cup, lamp, backpack) from the "Massive Memory – Object Categories" stimulus set (Konkle et al., 2010). One randomly chosen exemplar from each category (e.g., lamp A) was included in the target list. We also randomly chose one more exemplar from each of these categories to use as a relatively similar foil (e.g., lamp B) with corresponding targets. Additionally, 120 other categories were randomly selected (two exemplars in each). Each of these latter categories was randomly assigned to be a foil category for one of the target categories (for example, in Fig. 1a, the stroller is a foil category for the lamp target). As a result, each target was combined with three fixed foils—one relatively similar (always labeled as "Foil 1") and two dissimilar (labeled as "Foil 2" and "Foil 3"). This provided an unbiased set of test alternatives for the 4-AFC task so that each test display contained two categories with two exemplars. We made three parallel versions of the 2-AFC tests using the Latin square counterbalancing scheme: Each target-foil combination was assigned one of the versions, and each version included equal numbers of trials with each foil type (40 trials per foil type). These allowed us to balance the number of foils of different types and, as a result, balance overall task difficulty across participants. Assigning different test versions to different participants also guaranteed that each participant was tested for each target only once.

In Experiment 3, we tested recognition-memory spaces requiring finer target-foil discrimination, so that all foils assigned to each target were from the same object category but differed either in identity (different exemplar of a coffee mug), state (e.g., full mug vs. empty mug), or both (Fig. 1b). We used 120 categories from the exemplar-state set from Brady et al. (2013). Although the distinction between exemplars and states of real-world objects seems somewhat arbitrary in terms of underlying visual differences, they still can be functionally meaningful features that independently contribute to recognition (Brady et al., 2013; Utochkin & Brady, 2020). Each category included two exemplars in two different states. We randomly chose one exemplar in one of the states from each category for the target list. Then, the same exemplar in a different state was considered Foil 1, a different exemplar in the same state was considered Foil 2, and a different exemplar in a different state was considered Foil 3.

At the beginning of each experiment, participants were asked to memorize a series of upcoming images. In the study phase, the 120 targets were presented in a random order one by one (for 1,000 ms each in Experiments 1 and 3, and for 250 ms in Experiment 2) with a 500-ms interstimulus interval. The study phase was followed by either a version of the 2-AFC test or the 4-AFC test. In the 2-AFC task, test images were presented to the left and right of the screen center. In the 4-AFC task, test images were presented in four quadrants around the center. Participants had to click on the old item (target) in each trial, which was followed by feedback on correctness. The order of trials and the spatial positions of the items were randomized across participants.

Data analysis and modeling

Converting outcome proportions into SDT $\mathbf{d'}$. For each target-foil combination, we calculated the proportion of each possible response outcome among all the participants tested on this combination. Direct comparison between response frequencies in 2-AFC and 4-AFC tasks is problematic: For example, it is not intuitively obvious whether 75% correct answers in 2-AFC are comparable to 50% correct answers in 4-AFC, although both these numbers are 25 percentage points above the chance levels of 50% and 25%, respectively. SDT provides a tool to convert response frequencies into psychophysical discriminability metrics, d', independent of the number of test alternatives (Fig. 1c). We use d' as a metric of the target–foil distance in the hypothetical recognition space, which determines their confusability during the memory test. In the 2-AFC task, a single d' determines confusability

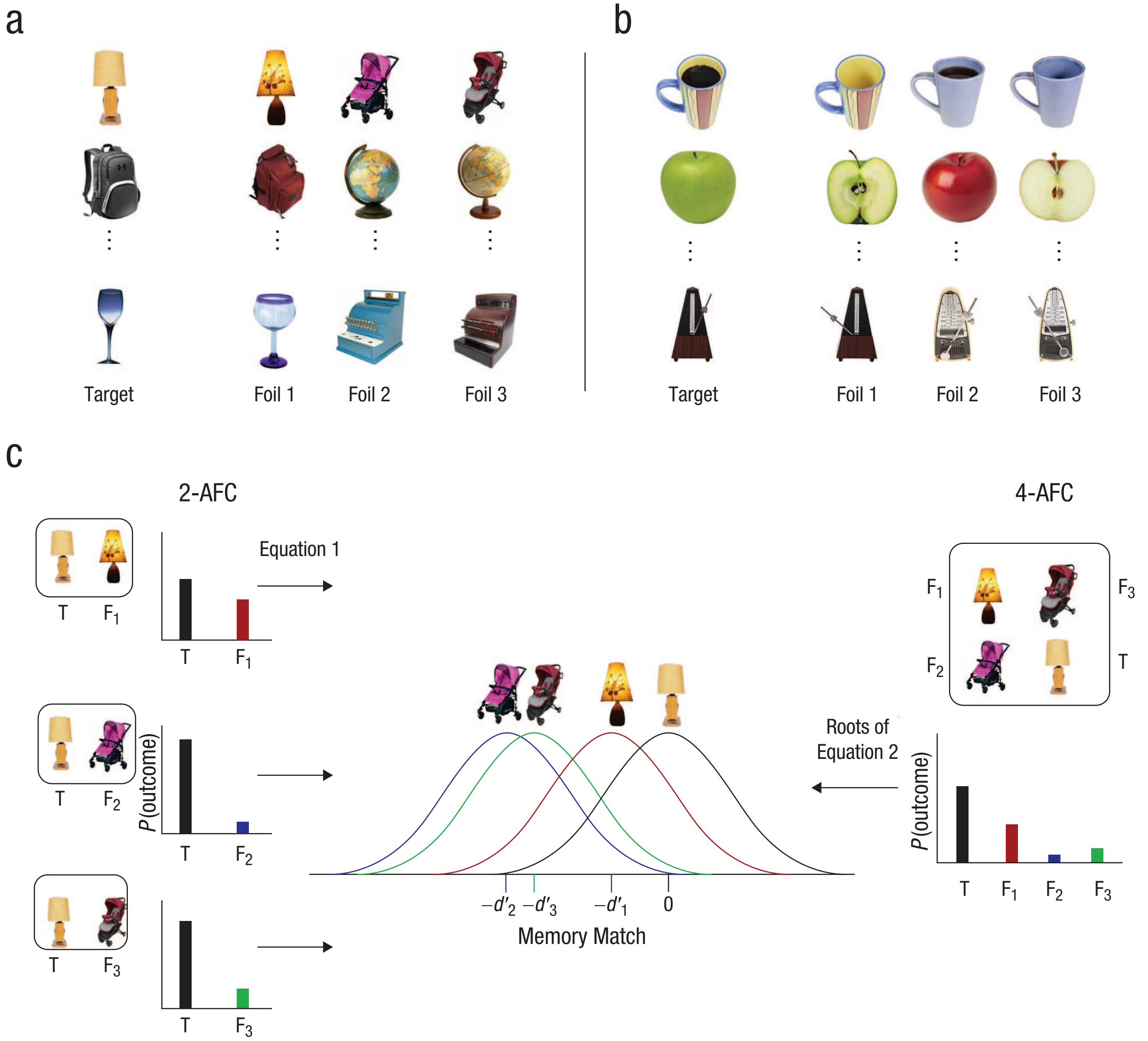

Fig. 1. The structure of stimulus sets used. In (a) are illustrated the sets used in Experiments 1 and 2; Foil 1 represents another exemplar from the same category as the target, and Foils 2 and 3 represent two exemplars from a different category. In (b) are illustrated the sets used in Experiment 3; Foil 1 represents the same exemplar as the target (but in a different state), and Foil 2 represents a different exemplar but in the same state as the target. Foil 3 represents different exemplar in a different state. In (c) is shown a signal-detection recognition space of an example combination of a target (T) and three foils ($F_{1–3}$). In the 2-AFC task, in the left side of (c), the target was tested separately against one of the foils (each pair was tested in different participants). This allows us to recover each target-foil distance (d') from the proportion of correct answers (Equation 1) separately. Because the target is a shared item between all 2-AFC test pairs, we can locate all foil d's to the same memory-match line, referencing them to the location of the target distribution in the middle of (c). The d's determine how memory-match effects are distributed for each test item and how confusable they are in a forced-choice memory test. In the 4-AFC task at right, the target and all three foils are presented simultaneously. This allows us to find all d's from the choice distribution among four alternatives by solving a system of Equation 2, in which we substitute the probability of each outcome to the left side. 2-AFC = two-alternative forced choice; 4-AFC = four-alternative forced choice.

between the target and the only foil. The probability of a correct answer to a target, p(target), in 2-AFC is the probability of a random sample from the target distribution ($\mu = d'$, $\sigma = 1$) being greater than another sample from the foil distribution ($\mu = 0$, $\sigma = 1$). Therefore, p(target) is defined by the difference distribution ($\mu = d'$, $\sigma = \sqrt{2}$) as the cumulative probability density Φ being greater than 0. If we rereference the mean of this difference distribution to 0, then d' can be obtained using the inverse z-function of the Gaussian distribution:

$$d' = -z[1 - p(\text{hit}); \mu = 0, \sigma = \sqrt{2}] = z\left[p(\text{hit}); \mu = 0, \sigma = \sqrt{2}\right] \quad (1)$$

The distribution of outcomes in a 4-AFC task, p(target) and $p(\text{foil}_i)$, where i is a foil type index from 1 to 3, is determined by a unique combination of three distances between the target and each of the foils, d'_i. (as in Fig. 1c, middle). As our model implies three separate noise distributions—one for each foil—instead of a single noise distribution, we can rereference the target distribution to 0. As a result, the foil distributions are positioned at $-d'_i$ (Fig. 1c). The probability of each 4-AFC outcome is defined as the probability that a sample drawn from the corresponding distribution exceeds samples drawn from all other distributions, leading the observer to judge that item as the most familiar. The probability of a correct answer to the target, p(target) in this model, is as follows (modified from DeCarlo, 2012)—

$$p\left(target\right) = \int_{-\infty}^{\infty} \varphi\left(x; \mu = 0\right) \cdot \prod_{i=1}^{3} \Phi\left(x; \mu = -d'_i\right) \cdot dx, \quad (2)$$

—where x is a memory-match magnitude produced by a stimulus and ϕ and Φ are the Gaussian probability density and cumulative density, respectively, of the x given the μ and $\sigma = 1$. The probability of each of the foil outcomes, $p(\text{foil}_i)$, can be computed accordingly. Finding a triplet of d'_i for a given 4-AFC combination from the data could be possible by solving a system of Equations 2 for d'_i given the observed p(target) and $p(\text{foil}_i)$. We searched for this solution using a general optimization by differential evolution algorithm (the *DEoptim* package for R; Mullen et al., 2011).

Testing invariance within the task: split-half reliability analysis. To estimate recognition consistency across target-foil combinations within each task (2-AFC and 4-AFC), we have run 10,000 iterations of split-half correlation analysis of the d'. In each iteration of this analysis of the 2-AFC data, we randomly split each of the three 2-AFC groups (parallel test versions) into two halves and concatenated all the first halves to make one half-sample with all possible target-foil pairs and all the second halves to make another half-sample. We calculated 2-AFC d' for each target-foil combination in each half-sample (Equation 1) and calculated Spearman-Brown (SB) correlations (ρ_{SB}) between the half-samples. For the 4-AFC task, we also randomly split the 4-AFC group into two random halves, calculated 4-AFC d's (based on Equation 2) for each half-sample, and ρ_{SB} between the half-samples. To test the significance of these correlations, we compared the average ρ_{SB}s against chance distributions obtained from 10,000 permutation tests correlating d's from one half-sample with d's from another half-sample with shuffled labels of target-foil combinations.

We also applied Deming regression (a linear regression method used when both the predicting and dependent variables are measured with an error) on each iteration of our split-half test. This method provides an estimate of the *measurement standard error* (*MSE*, which equals a standard deviation, or *SD*, of residuals) that can be interpreted as the amount of random deviation of the group sample value from the best-fit regression line between two groups. Conceptually, the perfect fit to the regression line (when all data points lie exactly on this line) can be taken for the ideal model of strong recognition invariance. Then the *MSE* gives us an estimate of how far the observed recognition pattern is from that predicted by ideal invariance (both due to a sampling error and genuine variation of d's across observers). These error estimates show the minimum possible d' variation expected from observers doing identical memory tasks and will be further used for testing the amount of invariance between more remote tasks, such as 2-AFC and 4-AFC. Because the half-splits were random and the task was the same for both half-samples, we set a default prior *MSE* ratio of 1 (equal errors) for our Deming regression model.

Testing the strength of invariance between 2-AFC and 4-AFC. We estimated a slope, intercept, and Spearman's correlation ρ between the full-sample 2-AFC d's and 4-AFC d's across 360 target-foil combinations. As in the split-half analysis, we used the Deming regression to estimate the slope and intercept of 4-AFC d' as a function of 2-AFC d'. The ratio of the average 4-AFC to the average 2-AFC measurement error estimated in the split-half analysis was taken as prior for that regression. To test whether the relationships between 2-AFC and 4-AFC d' show strong, weak, or no invariance, we compared our data against data simulated from three models, respectively. Our general modeling approach was based on Monte Carlo simulations. In these simulations, we generated random 2-AFC d's, mapped them onto 4-AFC d' according to the model's rule, and corrupted both sets of d's by a random measurement error based on the *MSE* estimates from the split-half analysis. Then we calculated a slope, intercept, and ρ between the simulated 2-AFC d' and 4-AFC d'. As a result of multiple simulation runs, we obtained distributions of the slopes, intercepts, and ρs produced by each model. These distributions gave us information as to how likely the combination of slope, intercept, and ρ observed in our data could be produced by each of the models. These likelihood estimates were used for the formal evaluation of model fits to the data.

Model 1: strong invariance. According to this model, 4-AFC d's keep the same metric relationships as 2-AFC d's (Fig. 2a). In other words, the functional relationship between all 2-AFC and 4-AFC d's is a single linear function with a nonzero slope. We term the slope and intercept of this linear function *generative model parameters*, as they generate a true mapping rule between 2-AFC and 4-AFC d's. We distinguish the generative parameters from the *output parameters*—namely, observed slope, intercept, and ρ calculated between the 2-AFC and 4-AFC d's simulated by the model as a result of applying the linear function with the generative parameters and measurement error. We simulated the data from this model for a grid of generative slopes and intercepts (both with step = 0.01) to estimate how the output slope, intercept, and ρ should be distributed. The range of generative slopes and intercepts varied between experiments and was based on a preliminary coarse grid search. The range of simulated slopes was set from 0.5 to 1.5. The range of intercepts was set from −0.5 to 0.5 for the models simulating comparisons between 2-AFC and 4-AFC within each experiment, and from −1 to 0 for the models simulating comparisons between Experiments 1 and 2. In each simulation (10,000 per grid cell), the model sampled a set of 360 d's from normal distributions approximating the distributions of observed predictor 2-AFC d's for each type of foils. For example, 120 d's for Foil 1 were sampled from a distribution with the same mean as the observed average (*M*) 2-AFC d' for Foil 1 and the observed standard deviation (*SD*) of 2-AFC d' for Foil 1, with measurement error (*MSE*) removed on the basis of estimates from the split-half analysis. The *MSE* was scaled by the factor of 1/√2 to acknowledge the fact that the full sample, rather than half, is now used. These initially sampled d's were considered the model's source d' (d'_{source}), or "true" values of d's in the recognition space chosen as predictors (2-AFC d' in most cases). Therefore, in a general case, d'_{source} sampling method is described by the following rule:

$$d'_{\text{source}} \sim N(\mu = M(2-\text{AFC } d'_{\text{foil type}\,i}),$$
$$\sigma = \text{sqrt}\,(SD^2(2-\text{AFC}d'_{\text{foil type}\,i}) - (MSE_{2\text{AFC}} / \sqrt{}\,2)^2). \quad (3)$$

The model then mapped the d'_{source} onto the space where the dependent variable belonged (4-AFC in this case) by applying a linear function of the d'_{source} with a given combination of generative slope and intercept. The resulting values termed d'_{mapped} were taken for the true d' in the 4-AFC space. Next, the model added a random measurement error (from a corresponding normal distribution with *SD* = *MSE*/√2) to each d'_{source} and d'_{mapped} to simulate the observed data in both conditions. The output slope, intercept, and ρ were estimated between these sets of observed data in the same way as in the real data. As a result of 10,000 runs of this model, we obtained predicted distributions of output slopes, intercepts, and ρs that we further used for model likelihood estimation.

Model 2: weak invariance. In this model, 4-AFC d' ordering within each target-foil combination is kept in the same order as in 2-AFC, but their metric relationships are not (Fig. 2b). In other words, more confusable objects are closer together than less confusable ones, but the distances between them in the 4-AFC space can be different from the 2-AFC space. It is easy to see that such a model communicates a trivial intuition about target-foil memory confusability: If Foil A is more similar to this target than Foil B, then A will always be more confusable than B at the memory test—whether it is 2-AFC or 4-AFC. However, unlike the strong invariance model, the present model assumes that change of the retrieval context (that is, introducing four alternatives instead of two) can change the distances from the target to each individual foil, d'_i, randomly such that any of the d's is not predictive of any other except for their magnitude order. The implementation of this model was similar to that of Model 1, with one exception. Specifically, instead of calculating d'_{mapped} for the 4-AFC task as a linear function of d'_{source} for 2-AFC, we randomly generated triplets of d'_{mapped} (three target-foil distances in simulated 4-AFC spaces) from three normal distributions approximating d' distributions for each foil type from our 4-AFC data with a measurement error removed (as in Equation 3, but with 2-AFC replaced with 4-AFC). We kept generating the triplets of d'_{mapped} until their order matched the order of the 2-AFC triplet of d'_{source}. The rest of the modeling steps were the same as in Model 1.

Note that, although Model 2 is less constrained than Model 1 (any orderly vs. strictly linear relationships), our implementation of Model 2 has no free parameters, whereas Model 1 has two (generative slope and intercept). This may seem counterintuitive, as less constrained models are typically expected to have more free parameters. However, this contradiction resolves if we consider that different 4-AFC d'_{mapped} in Model 2 originates from a multitude of linear functions of corresponding 2-AFC d'_{source} with various generative slopes or intercepts. In this interpretation of Model 2, the number of its free parameters would be double the number of the generative linear functions and would overcome the number of free parameters in Model 1. However, we did not fit these implicit parameters to the data because (a) the number of such generative functions cannot be set a priori and (b) our crucial model comparison specifically aimed to distinguish between the

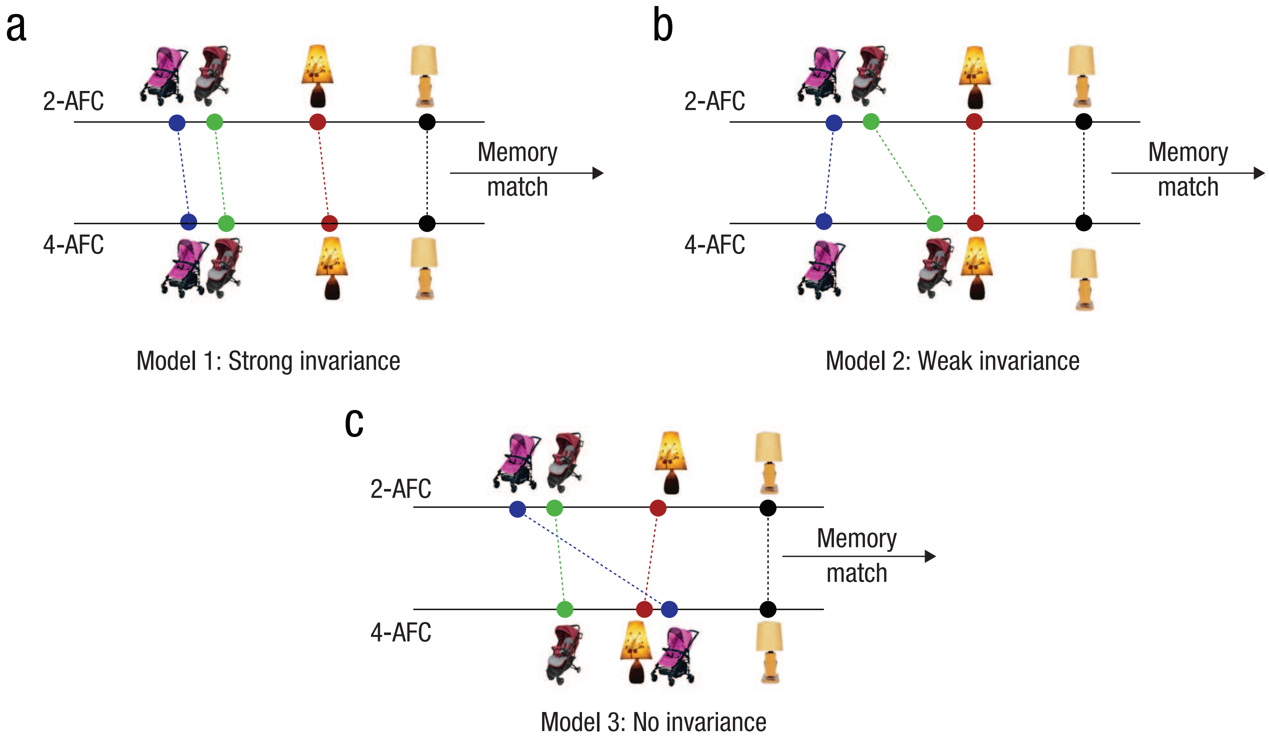

Fig. 2. Three models of recognition-space invariance between memory tests (for example, 2-AFC and 4-AFC). Parallel horizontal lines depict the SDT memory-match dimension (each line for one of the compared memory tests). Colored dots show the mean location of each item's memory distributions: The distance from the target black dot to each of the foil dots is one d'. In Model 1 (a; strong invariance), all d's in a 4-AFC task to a large extent keep the same metric relationships as their counterparts in a 2-AFC task. The preservation of the metric relationships between d's makes it possible to map 2-AFC d' onto 4-AFC d' with a single metric function (for example, a linear equation whose slope and intercept exactly determine each 4-AFC d' as a function of their counterpart 2-AFC d'). Although the example figure shows only one recognition space, Model 1 generalizes to all recognition spaces, such that all d's in all trials are scaled by the same linear function. In Model 2 (b; weak invariance), the order of d's in a 4-AFC space stays the same as in a 2-AFC space (more confusable foils from 2-AFC are also more confusable in 4-AFC). However, metric relationships between the d's are not preserved, and the direction and magnitude of their change between the tasks look random. In Model 3 (c; no invariance), item locations in the 2-AFC recognition space are completely not predictive of locations in the 4-AFC space. 2-AFC = two-alternative forced choice; 4-AFC = four-alternative forced choice.

model with a single generative function (strong invariance) and the models where the number of such functions is not determined (weak or no-invariance). The same note equally applies to Model 3 below.

Model 3: no invariance. This model was used as a null model, suggesting that the recognition spaces are not invariant, and consequently that target-foil distances in 4-AFC displays have nothing in common with those in 2-AFC (Fig. 2c). The implementation of this model was similar to that of Models 1 and 2, but d'_{mapped} for 4-AFC was generated randomly from the respective normal distributions (as in Equation 3, but with 2-AFC replaced with 4-AFC). No additional order alignment between the generated sets of d'_{source} and d'_{mapped} was made.

Model comparison. We evaluated the likelihood of the three models as a joint probability of observing a combination of ρ, slope, and intercept as in our data under each of these models. We estimated these probabilities using output parameter distributions from the model simulations. As these distributions were based on a finite number of simulations rather than a theoretical density function, these distributions were discrete: The output parameters were binned (bin size = 0.01), and the proportion of each bin to all simulations was taken as an approximation of its probability mass, *Pr*. The log-likelihood, *L*, of a model M was calculated as follows:

$$L(\text{M} \mid \rho, \beta_0, \beta_1) = \log(Pr(\rho \mid \text{M})) + \log(Pr(\beta_0 \mid \text{M})) + \log(Pr(\beta_1 \mid \text{M})), \quad (4)$$

where ρ, β_0, and β_1 are the simulated output bins for the observed correlation, intercept, and slope. For Model 1, which had two free parameters—the slope and intercept of d'_{source} vs. d'_{mapped} function—we used maximum-likelihood estimation to find the best-fit function. Finally, we compared Model 1 with the two best-fitting free parameters, Models 2 and 3 (0 free parameters), using the Akaike information criterion (AIC): AIC = $-2\ L(\mathrm{M}) + 2k$, where k is the number of free parameters. A model with the lowest AIC was accepted as the best model explaining the data.

Results

Experiment 1

Split-half reliability. We found high split-half correlations ρ_{SB} both for 2-AFC d' (average ρ_{SB} = .80, 95% confidence interval, or CI = [.75, .83]; Fig. 3a) and 4-AFC d' (average ρ_{SB} = .77, 95% CI = [.73, .80]; Fig. 3b). The average ρ_{SB} for both tasks was above chance correlation levels ($p < .0001$) established by the permutation half-split tests (2-AFC: mean ρ_{SB} = −.006, 95% CI = [−.23, .18]; 4-AFC: average ρ_{SB} = 0, 95% CI = [−.23, .19]; see Figs. 3a and 3b), suggesting highly consistent recognition performance produced by the same target-foil combinations across participants. The average half-sample *MSE*s, as estimated by the Deming regression, were 0.359 for 2-AFC and 0.362 for 4-AFC, which converted into the full-sample *MSE*s of 0.254 and 0.256. The average slope and intercept of a linear function between the half-samples were 1 and 0, respectively, for both tasks, which indicates that the d's were generally equal between the half-samples.

Invariance strength between 2-AFC and 4-AFC. The 2-AFC and 4-AFC d's showed a high correlation (Spearman's ρ = .76; Fig. 3c). The slope and intercept of the Deming regression model were 1.03 and 0.03, respectively, suggesting that 2-AFC d's and 4-AFC d's were close to equal. In other words, the overall memory discriminability of targets and foils did not substantially change as a function of test complexity, even though decision-making in 4-AFC is harder than in 2-AFC. Most important, we found that our data was better explained by Model 1 (AIC = 18.34, with the best-fit slope = 1.02 and intercept = 0.02) than by Model 2 (AIC = 35.01) or Model 3 (AIC = 41.08; Fig. 3d). Therefore, we concluded that 2-AFC d's for individual target-foil combinations map onto 4-AFC d's with strong invariance.

Experiment 2

Experiment 2 was a replication of Experiment 1 but with a shorter encoding time (250 ms per image) at the study phase; this change was supposed to decrease the overall memory strength. We did the same set of analyses as in Experiment 1 to estimate the degree of invariance between 2-AFC and 4-AFC under the same encoding conditions. An important addition was testing whether the change in memory strength between Experiments 1 and 2 could also change the recognition spaces. To this end, we tested whether d's in Experiment 1 can predict d's in Experiment 2 and whether these predictions fitted the strong, weak, or no-invariance model.

Split-half reliability. We found high Spearman-Brown correlations between half-sample d's in both 2-AFC (average ρ_{SB} = .84, 95% CI = [.81, .87]; see Fig. S1a in the Supplemental Material available online) and 4-AFC (average ρ_{SB} = .79, 95% CI = [.75, .82]; see Fig. S1b in the Supplemental Material). These correlations were greater than chance (average ρ_{SB} = 0, 95% CI = [−.23, .19], $p < .0001$). The estimated half-sample *MSE*s were 0.305 for 2-AFC and 0.326 for 4-AFC, which converted to the full-sample *MSE*s of 0.216 and 0.230.

Invariance strength between 2-AFC and 4-AFC. The 2-AFC and 4-AFC d's within Experiment 2 showed a high correlation (Spearman's ρ = .72; see Fig. S1c in the Supplemental Material). The slope and intercept of the Deming regression model were 0.92 and 0.08, respectively. Model comparison showed that this data pattern was better explained by Model 1 (AIC = 23.02, with the best-fit slope = 0.88 and intercept = 0.08) than Model 2 (AIC = 33.74) or Model 3 (AIC = 47.63; see Fig. S1d). Therefore, we concluded that 2-AFC d's for individual target-foil combinations map onto 4-AFC d's with strong invariance.

Comparison between long and short encoding times. In this analysis, we used the data from Experiment 1 (long encoding time) as a predictor and data from Experiment 2 (short encoding time) as a dependent variable. We tested two Deming regression models for different encoding times but similar test complexities (2-AFC vs. 2-AFC and 4-AFC vs. 4-AFC) and one model for different encoding times and test complexities (2-AFC vs. 4-AFC). In all cases (Fig. 4a), we found high correlations (ρ = .77, .76, and .68, respectively, in order of mentioning in the previous sentence). The slopes of the regression functions were 1.01, 0.91, and 0.94, and the intercepts were −0.55, −0.46, and −0.44, respectively. The negative intercepts suggest that overall target-foil discriminability in Experiment 2 was lower than in Experiment 1, as could be expected from the reduced encoding time. Consequently, this discriminability decrement can be interpreted as a result of lower memory strength.

Even though the memory strength decreased in Experiment 2 compared with Experiment 1, we found evidence for the preserved strong invariance, as shown by the comparison of our three invariance models (Fig. 4b). In the 2-AFC versus 2-AFC case, Model 1 (AIC = 17.89,

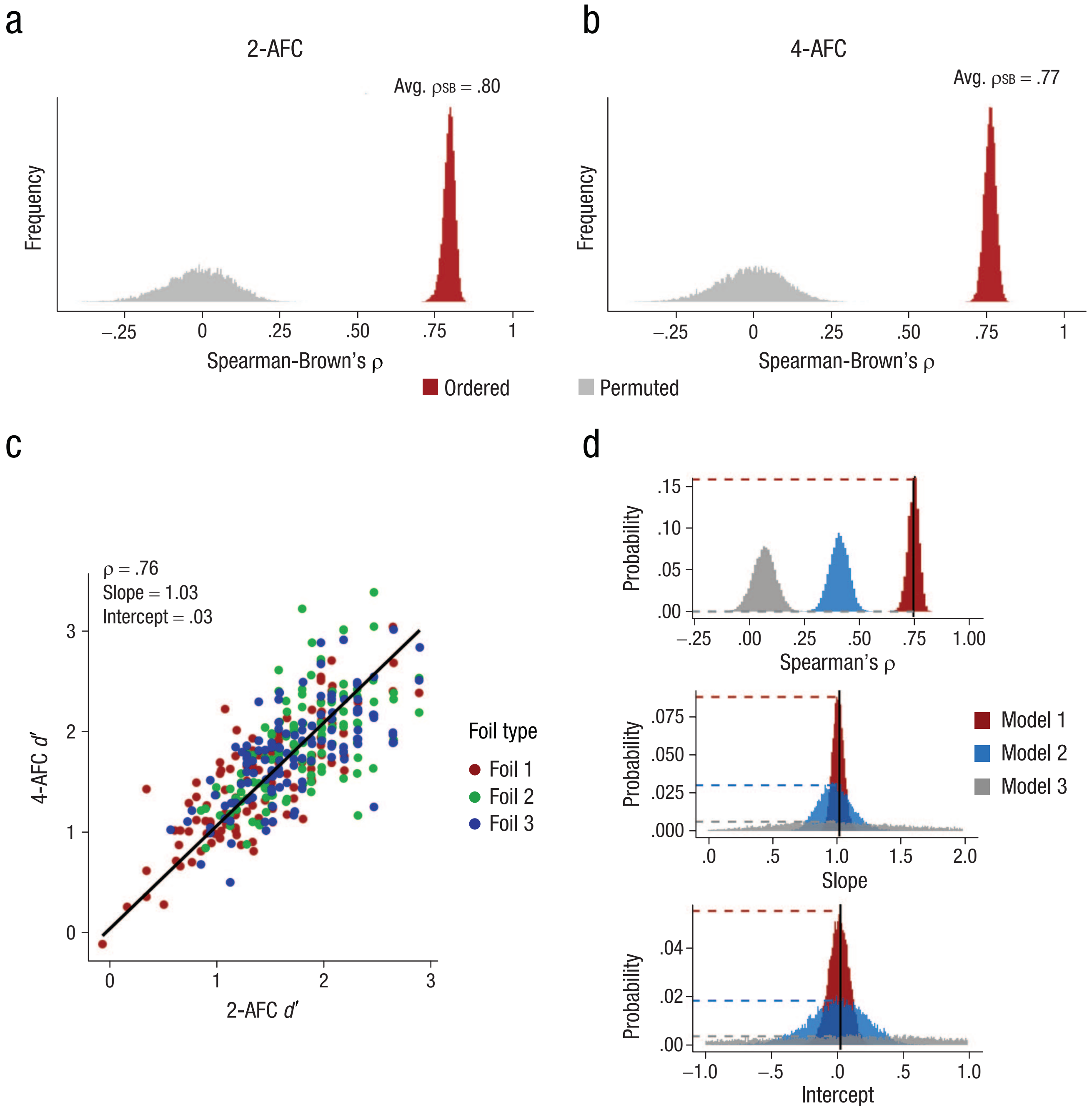

Fig. 3. Results of Experiment 1. Split-half correlations across target-foil combinations are shown for (a) the 2-AFC tests and (b) the 4-AFC test; in (c) is shown a scatterplot of 4-AFC *d′* as a function of 2-AFC *d′* with a Deming regression best-fit trend line; and in (d) is shown likelihood estimation of the three invariance models. Each color distribution represents predicted probabilities of an observed Spearman's ρ (top), slope (middle), and intercept (bottom) under Model 1 (strong invariance with the best-fit slope and intercept), Model 2 (weak invariance), or Model 3 (no invariance). Black vertical lines show the location of the actually observed parameters. Colored dashed lines indicate the probability of observing these parameters under each model (the height of a distribution at the location of the observed parameter). It can be seen that in all panels, the observed parameters are much more probable under Model 1 (red distribution), making this model the most likely explanation for this data. 2-AFC = two-alternative forced choice; 4-AFC = four-alternative forced choice; SB = Spearman-Brown.

with the best-fit slope = 1.01 and intercept = –0.54) explained the data better than Model 2 (AIC = 38.02) or Model 3 (AIC = 50.66). In the 4-AFC versus 4-AFC case, Model 1 (AIC = 17.97, with the best-fit slope = 0.92 and intercept = –0.48) explained the data better than Model 2 (AIC = 38.56) or Model 3 (AIC = 42.35). Finally, in the 2-AFC vs. 4-AFC case, Model 1 (AIC = 24.85, with the best-fit slope = 0.93 and intercept = –0.43) explained the data better than Model 2 (AIC = 38.09) or Model 3 (AIC = 43.49).

Experiment 3

Split-half reliability. We found high Spearman-Brown correlations between half-sample d's in both 2-AFC (average ρ_{SB} = .87, 95% CI = [.85, .89]; see Fig. S2a in the Supplemental Material) and 4-AFC (average ρ_{SB} = .85, 95% CI = [.82, .87]; see Fig. S2b). These correlations were greater than chance (average ρ_{SB} = 0, 95% CI = [−.23, .19], $p < .0001$). The estimated half-sample *MSE*s were 0.317 for 2-AFC and 0.339 for 4-AFC, which converted into full-sample *MSE*s of 0.224 and 0.240.

Invariance strength between 2-AFC and 4-AFC. The 2-AFC and 4-AFC d's within Experiment 3 showed a high correlation (Spearman's ρ = .83). The slope and intercept of the Deming regression model were 0.98 and 0.09, respectively. Model comparison showed that this data pattern was better explained by Model 1 (AIC = 16.03, with the best-fit slope = 0.98 and intercept = 0.10) than by Model 2 (AIC = 33.64) or Model 3 (AIC = 45.11; see Fig. S2c in the Supplemental Material). Again, we concluded that 2-AFC d's for individual target-foil combinations map onto 4-AFC d's with strong invariance. Therefore, we show that strong invariance generalizes to recognition spaces of objects requiring fine discrimination of object variations within a category.

Discussion

We found evidence for strong invariance of target-foil discriminability across observers, memory tests, and memory-strength manipulations at encoding. This invariance generalizes to memory discrimination between categories, objects of the same category, and different states of the same object. To summarize, our results consistently demonstrated that all target-foil d's recovered from one test condition (2-AFC) are mapped onto d's from another condition (4-AFC) via a single linear function. This suggests the preservation of metric relationships between the target and foils despite the retrieval-context change. Even when all recognition spaces are globally reduced because of the shorter encoding time, this linear mapping is still evident, suggesting that the shrinkage does not violate the geometry of the recognition space.

Therefore, in contrast to previous research on context sensitivity—which has suggested that there are substantial representational space changes in value-driven decision-making (Busemeyer et al., 2019; Spektor et al., 2021) or inference (Trueblood, 2012)—we concluded that memory-driven decisions are based on the context-invariant representational space. Note that the invariance of the representational space does not imply that recognition is context independent: After all, 4-AFC is more difficult than 2-AFC, and target-foil similarity dramatically influences recognition accuracy (Brady & Störmer, 2024; Migo et al., 2013). Rather, the invariance implies that the relationship between the target and each of the foils underlying their mnemonic discrimination is not changed when other foils show up in the common retrieval context. Note that our demonstration of strong invariance is based on a paradigm in which we artificially restricted context variables (i.e., other than the context imposed by test alternatives). For example, our participants memorized stimuli that were presented one at a time on a neutral background. Therefore, our conclusion about strong invariance can be limited to this paradigm. Future research is needed to investigate whether the invariance of recognition-memory spaces generalizes to more contextually rich conditions (e.g., when objects are encoded or retrieved as part of a meaningful scene).

Recognition-memory spaces, similarity spaces, and memorability

Considering many existing memory models (Gillund & Shiffrin, 1984; Hintzman, 1988; Nosofsky et al., 2011; Oberauer & Lin, 2017; Schurgin et al., 2020), it can be natural to think that the invariant spaces that we investigated in our study are simply psychological similarity spaces that can be recovered from psychophysical scaling (Daggett & Hout, 2025; Goldstone, 1994; Hebart et al., 2023; Hout et al., 2013; Schurgin et al., 2020; Son et al., 2021). Because study lists were identical across all participants of the compared groups, the invariance of recognition spaces could be explained by similar situational activation patterns in these psychological spaces. However, we use a broader term—"recognition-memory spaces," rather than "similarity spaces"—because, apart from all factors making given targets and foils more or less similar, there can be factors intrinsically determining individual memorability of each item (Isola, Xiao, et al., 2011; Kramer et al., 2023). That is, the confusability of a given target-foil combination can depend on how likely it is, per se, that this target causes correct recognition and that each foil causes false recognition, in addition to their relative similarity. In line with that, previous evidence has suggested that recognition decisions about individually tested items can generalize to forced choices between several items (Jang et al., 2009). Independent manipulations with item similarity and memorability in future research can help dissociate the contributions of these factors to recognition-memory spaces.

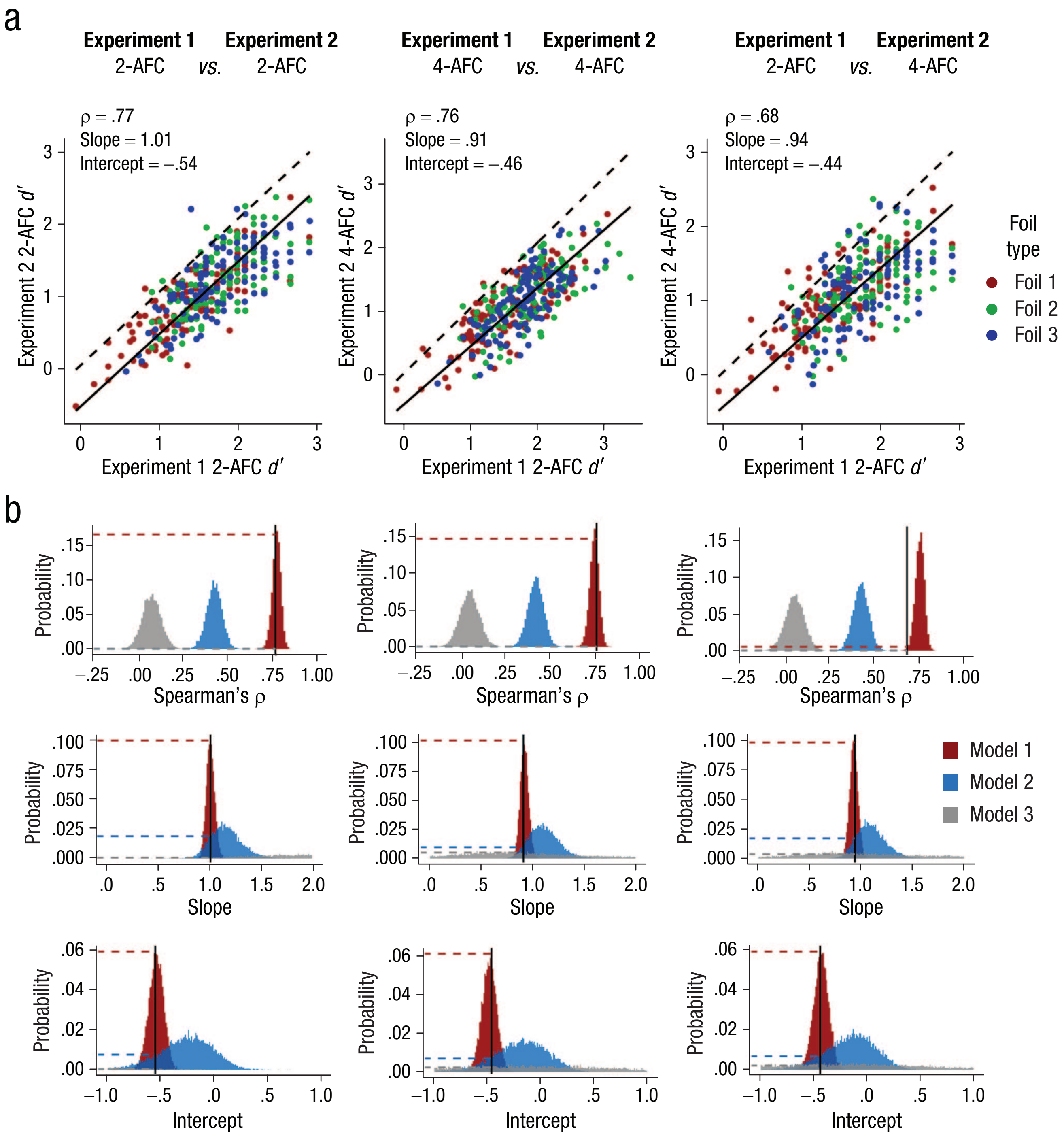

Fig. 4. Strong recognition invariance transfers from Experiment 1 (long encoding time) to Experiment 2 (short encoding time). In (a) are shown scatterplots of Experiment 2 d' as a function of Experiment 1 d', left to right: 2-AFC versus 2-AFC, 4-AFC versus 4-AFC, and 2-AFC versus 4-AFC. Spearman's ρ and Deming regression slopes and intercepts are shown at the top left corners. Best-fit linear functions between the variables presented on the scatterplots are shown with the solid trend lines. For reference, the dashed lines were added to show the slope and intercept of a 2-AFC vs. 4-AFC function from Experiment 1. As can be concluded from the solid lines always being below the dashed lines (as well as from the negative intercepts), performance in all tasks of Experiment 2 was always worse than in Experiment 1. However, the correlations between Experiment 1 and Experiment 2 remained high, suggesting invariance. Likelihood estimation of the three invariance models for each comparison is shown in (b) in the same left-right order. In all cases, the observed parameters are much more probable under Model 1 (red distribution), corroborating strong invariance. 2-AFC = two-alternative forced choice; 4-AFC = four-alternative forced choice.

The dimensionality of the recognition space

Although we used a unidimensional SDT model for measuring the target-foil distances in the recognition-memory spaces, we do not imply that the single dimension drives all underlying discriminations. We conceptualize this dimension as an integral effect of evidence accumulation from a multidimensional space with various dimensions representing individual encoded features of items and episodes. Upon test item presentation, these dimensions may produce variable amounts of mnemonic evidence giving access to specific episodic information about details of the retrieved material. Unidimensional SDT captures the cumulative effect of this evidence (i.e., how much information about the test item observers have; Wixted & Mickes, 2010) rather than the use of a single nonspecific memory signal. Although the structure of this multidimensional space is beyond the scope of the current article, this issue can be addressed using multidimensional SDT applied to orthogonal feature manipulations among targets and foils (as in our Experiment 3; see also Brady et al., 2013, and Balaban et al., 2020) or machine learning (Brady & Störmer, 2024; Sanders & Nosofsky, 2020).

To summarize, our research suggests that, even though items that co-occur together collectively create a unique retrieval context influencing the recognition of memory targets, representational spaces underlying mnemonic discriminations between these stimuli are strongly invariant. We have presented behavioral and model-based evidence that including additional foils into test alternatives preserves the relative discriminability between the target and other foils. Moreover, we showed that the SDT framework offers a convenient metric to recover recognition-memory spaces convertible between tasks with different choice complexity and demonstrating linear mapping between the recognition spaces of these choices.

Transparency

Action Editor: Tom Beckers
Editor: Simine Vazire
Author Contributions

Igor Utochkin: Conceptualization; Data curation; Formal analysis; Funding acquisition; Methodology; Project administration; Supervision; Visualization; Writing – original draft.

Daniil Azarov: Data curation; Investigation; Methodology; Software; Writing – review & editing.

Daniil Grigorev: Data curation; Investigation; Methodology; Software; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research received no specific funding and was paid for by the authors' personal funds.

Artificial Intelligence

No artificial-intelligence-assisted technologies were used in this research or the creation of this article.

Ethics

This research complies with the Declaration of Helsinki and received approval from a local ethics board (University of Chicago, Protocol No. IRB15-1290).

Open Science Framework (OSF)

To facilitate long-term preservation, all OSF files have been registered at https://osf.io/bfgnc (study materials) and https://osf.io/7huwa (data and analysis scripts).

Open Practices

Preregistration: No aspects of the research were preregistered. Materials: All study materials are publicly available (https://osf.io/ectdx/files/osfstorage). Image stimuli used in the experiments are parts of publicly available stimulus sets provided by another laboratory (Experiments 1 and 2: https://bradylab.ucsd.edu/stimuli/ObjectCategories.zip; Experiment 3: https://bradylab.ucsd.edu/stimuli/StateExemplar.zip). Data: All primary data are publicly available (https://osf.io/5whxp/files/osfstorage). Analysis scripts: All analysis scripts are publicly available (https://osf.io/5whxp/files/osfstorage). Computational reproducibility: The computational reproducibility of the results in the main article (but not the supplementary materials) has been independently confirmed by the journal's STAR team.

ORCID iDs

Igor Utochkin https://orcid.org/0000-0001-8433-446X
Daniil Azarov https://orcid.org/0000-0001-7241-2192
Daniil Grigorev https://orcid.org/0009-0006-5474-8518

Acknowledgments

The authors thank Varvara Mikhalishchina for the help with arranging stimuli for the experiments.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976251384640

References

Bainbridge, W. A. (2020). The resiliency of image memorability: A predictor of memory separate from attention and priming. *Neuropsychologia*, *141*, Article 107408. https://doi.org/10.1016/j.neuropsychologia.2020.107408

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334. https://doi.org/10.1037/a0033872

Balaban, H., Assaf, D., Meir, M. A., & Luria, R. (2020). Different features of real-world objects are represented in a dependent manner in long-term memory. *Journal of Experimental Psychology: General*, *149*(7), 1275–1293. https://doi.org/10.1037/xge0000716

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2013). Real-world objects are not represented as bound units: Independent forgetting of different object details from visual memory. *Journal of Experimental Psychology: General*, *142*(3), 791–808. https://doi.org/10.1037/a0029649

Brady, T. F., & Störmer, V. S. (2024). Comparing memory capacity across stimuli requires maximally dissimilar foils: Using deep convolutional neural networks to understand visual working memory capacity for real-world objects. *Memory & Cognition*, *52*(3), 595–609. https://doi.org/10.3758/s13421-023-01485-5

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in Cognitive Sciences*, *23*(3), 251–263. https://doi.org/10.1016/j.tics.2018.12.003

Colloff, M. F., Wilson, B. M., Seale-Carlisle, T. M., & Wixted, J. T. (2021). Optimizing the selection of fillers in police lineups. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(8), e2017292118. https://doi.org/10.1073/pnas.2017292118

Daggett, E. W., & Hout, M. C. (2025). A tutorial review on methods for collecting similarity judgments from human observers. *Attention, Perception, & Psychophysics*, *87*, 737–751 https://doi.org/10.3758/s13414-025-03044-3

DeCarlo, L. T. (2012). On a signal detection approach to *m*-alternative forced choice with bias, with maximum likelihood and Bayesian approaches to estimation. *Journal of Mathematical Psychology*, *56*(3), 196–207. https://doi.org/10.1016/j.jmp.2012.02.004

Dumbalska, T., Li, V., Tsetsos, K., & Summerfield, C. (2020). A map of decoy influence in human multialternative choice. *Proceedings of the National Academy of Sciences*, *117*(40), 25169–25178. https://doi.org/10.1073/pnas.2005058117

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1–67. https://psycnet.apa.org/doi/10.1037/0033-295X.91.1.1

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*, 381–386. https://doi.org/10.3758/BF03204653

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection theory: A user's guide* (3rd ed.). Routledge. https://doi.org/10.4324/9781003203636

Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., & Baker, C. I. (2023). THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, *12*, Article e82580. https://doi.org/10.7554/elife.82580

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551. https://psycnet.apa.org/doi/10.1037/0033-295X.95.4.528

Hout, M. C., Papesh, M. H., & Goldinger, S. D. (2013). Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(1), 93–103. https://doi.org/10.1002/wcs.1203

Hunt, R. R. (1995). The subtlety of distinctiveness: What von Restorff really did. *Psychonomic Bulletin & Review*, *2*, 105–112. https://doi.org/10.3758/BF0321441

Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011). Understanding the intrinsic memorability of images. *Advances in Neural Information Processing Systems*, *24*, 2429–2437. https://doi.org/10.21236/ada554133

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? *Journal of Vision*, *11*(11), Article 1282. https://doi.org/10.1167/11.11.1282

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*(2), 291–306.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558–578. https://doi.org/10.1037/a0019165

Kramer, M. A., Hebart, M. N., Baker, C. I., & Bainbridge, W. A. (2023). The features underlying the memorability of objects. *Science Advances*, *9*(17), eadd2981. https://doi.org/10.1126/sciadv.add2981

Meagher, B. J., & Nosofsky, R. M. (2023). Testing formal cognitive models of classification and old-new recognition in a real-world high-dimensional category domain. *Cognitive Psychology*, *145*, Article 101596. https://doi.org/10.1016/j.cogpsych.2023.101596

Migo, E., Montaldi, D., Norman, K. A., Quamme, J., & Mayes, A. (2009). The contribution of familiarity to recognition memory is a function of test format when using similar foils. *Quarterly Journal of Experimental Psychology*, *62*(6), 1198–1215. https://doi.org/10.1080/17470210802391599

Migo, E. M., Montaldi, D., & Mayes, A. R. (2013). A visual object stimulus database with standardized similarity information. *Behavior Research Methods*, *45*, 344–354. https://doi.org/10.3758/s13428-012-0255-4

Mullen, K. M., Ardia, D., Gil, D. L., Windover, D., & Cline, J. (2011). DEoptim: An R package for global optimization by differential evolution. *Journal of Statistical Software*, *40*(6), 1–26. https://doi.org/10.18637/jss.v040.i06

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception & Performance*, *17*(1), 3–27. https://doi.org/10.1037/0096-1523.17.1.3

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review*, *118*(2), 280–315. https://doi.org/10.1037/a0022494

Oberauer, K., & Lin, H. Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21–59. https://doi.org/10.1037/rev0000044

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, *108*(2), 370–392. https://psycnet.apa.org/doi/10.1037/0033-295X.108.2.370

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. https://psycnet.apa.org/doi/10.1037/0278-7393.21.4.803

Sanders, C., & Nosofsky, R. M. (2020). Training deep networks to construct a psychological feature space for a natural-object category domain. *Computational Brain & Behavior*, *3*(3), 229–251. https://doi.org/10.1007/s42113-020-00073-z

Schurgin, M., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*(11), 1156–1172. https://doi.org/10.1038/s41562-020-00938-0

Shen, K. J., Colloff, M. F., Vul, E., Wilson, B. M., & Wixted, J. T. (2023). Modeling face similarity in police lineups. *Psychological Review*, *130*(2), 432–461. https://doi.org/10.1037/rev0000408

Son, G., Walther, D., & Mack, M. L. (2021). Scene wheels: Measuring perception and memory of real-world scenes with a continuous stimulus space. *Behavior Research Methods*, *54*(1), 444–456. https://doi.org/10.3758/s13428-021-01630-5

Spektor, M. S., Bhatia, S., & Gluth, S. (2021). The elusiveness of context effects in decision making. *Trends in Cognitive Sciences*, *25*(10), 843–854. https://doi.org/10.1016/j.tics.2021.07.011

Trueblood, J. S. (2012). Multialternative context effects obtained using an inference task. *Psychonomic Bulletin & Review*, *19*, 962–968. https://doi.org/10.3758/s13423-012-0288-9

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 479–496. https://doi.org/10.1016/s0022-5371(81)90129-8

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, *111*(3), 757–769. https://psycnet.apa.org/doi/10.1037/0033-295X.111.3.757

Utochkin, I. S., & Brady, T. F. (2020). Independent storage of different features of real-world objects in long-term memory. *Journal of Experimental Psychology: General*, *149*(3), 530–549. https://doi.org/10.1037/xge0000664

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review*, *117*, 1025–1054. https://doi.org/10.1037/a0020874

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, *121*(2), 262–276. https://doi.org/10.1037/a0035940