Comparison of Resource-Rational Observer Models of Individual and Ensemble Spatial Perception

Yanina E. Tena Garcia (yanina.tena-garcia@psychol.uni-giessen.de)[1,2], Bianca R. Baltaretu[1], Katja Fiehler[1] & Dominik M. Endres (dominik.endres@uni-marburg.de)[2]

[1]Department of General Psychology, Justus-Liebig University Giessen

[2]Theoretical Cognitive Science, Dept. Psychology, Marburg University

Abstract

To better understand the underlying mechanisms of individual and ensemble perception in naturalistic scenes, we compared three bayesian resource-rational models on experimental data (from 27 healthy adults): the 'Individual Encoding Model' (IEM), a variant of the summation model; the 'Ensemble Encoding Model' (EEM), related to the automatic averaging model; and the 'Task Adapted Encoding Model' (TAEM), a flexible combination of both models that adapts to task demands. In the experiment, participants encoded and reproduced either an individual object position or an ensemble position (group centroid) in a 3D-rendered scene using a computer mouse. In both tasks, we manipulated set size (3, 6, 10 objects) and presentation time (50, 100, 800 ms). The EEM and TAEM generally explained the human behavioral data best. We conclude that, in naturalistic scenes, the choice between individual versus ensemble perception is likely driven by the more compact scene representation of the ensemble model.

Keywords: ensemble perception; spatial perception; scene perception; cognitive model; bayesian model

Introduction

Extracting object locations is a critical task in daily life. It requires determining the precise position of a specific object (e.g., picking up a coffee mug), referred to as individual object perception, and/or identifying the average position of a group of objects (e.g., locating where the group of mugs is placed in a kitchen), known as ensemble perception. Behavioral findings suggest that individual and ensemble perception rely on distinct processing mechanisms, especially reflected in their differential effects of set size and presentation time (Ariely, 2001; Chong & Treisman, 2003, 2005; Melcher et al., 2021; Ward et al., 2016). Extracting individual information is supposed to be a sequential and resource-intensive process (Perry & Fallah, 2014), whereas ensemble information is considered to be extracted automatically prior to individual object representations (Haberman & Whitney, 2007; Melcher et al., 2021; Oriet & Hozempa, 2016; Yildirim et al., 2018). Recent computational modeling studies challenge this traditional view and instead suggest that ensemble perception relies on pooling pre-encoded individual information (Harrison et al., 2021; Robinson & Brady, 2023; Utochkin et al., 2024).

Ensemble perception is thought to aid orientation in complex environments, though previous research into this topic was conducted using overly simplistic stimulus displays (e.g., Melcher et al., 2021; Neumann et al., 2018). Meaningful scene context can facilitate or impair individual perception (Draschkow et al., 2014; Ringer et al., 2021). However, its influence on ensemble perception remains unexplored.

Thus, we investigate the computations underlying human encoding of multiple object locations embedded in naturalistic scenes with limited working memory (WM).

Behavioral Experiment

The data from the behavioral experiment, along with the computational modeling scripts, are available here: https://doi.org/10.60834/tam-datahub-15.

To assess participants ability to locate individual and ensemble information, data from 27 participants were collected (all healthy students of Justus Liebig University Giessen; 17 females, mean age = 23.66 years). Participants were seated in front of a 25" monitor (60 Hz, 1920 × 1080 px) in a dark room. A chinrest maintained a constant 60-cm viewing distance.

Participants performed a total of four experimental blocks, two for each the individual and the ensemble task, consisting of 54 trials each. The task was to encode a test scene as accurately as possible and to reproduce the position of a predefined target. These test scenes (created in Blender (v2.9; https://www.blender.org/)) showed always the same kitchen environment with either 3, 6 or 10 readily movable kitchen-related target objects and were presented for one of three presentation times (50, 100 or 800 ms). After scene presentation, participants were instructed to either recall the position of one of the presented target objects as accurately as possible (individual task) or the 'average position of all presented target objects' (verbatim instruction of the ensemble task). Participants responded by viewing the empty kitchen environment and clicking on the remembered position on the monitor, with no time limit. Three practice trials were performed for each task condition to control if participants correctly understood the tasks. The experiment was conducted using PsychoPy (v2021.2.0; Peirce et al., 2019).

Locating accuracy was measured as the distance between the clicked and the actual (veridical) 2D target position. Individual target positions were defined relative to the center of mass of the respective object, and ensemble positions by the average of the veridical object locations.

A total of 727 trials (12.47% of the 27 datasets) were excluded. 616 of these trials could be traced back to participants not fixating the countdown till the end and therefore not starting with a central fixation. Further exclusion cri-

teria (not mutually exclusive) were blinks during scene presentation in the 50- and 100-ms conditions (≈0.22%), not fixating the screen for more than 30% of the presentation time (≈0.12%), blinks longer 200 ms during scene presentation (≈0.15%), and high locating errors (>3 SD from the mean, computed separately for locating task and presentation time conditions) accompanied by short reaction times (<1 s;≈1.58%). These data were preprocessed using Jupyter Notebook (v6.2.0; https://jupyter.org/).

Based on previous studies (e.g., Melcher et al., 2021; Neumann et al., 2018), we hypothesized that locating performance increases with more presentation time in both individual and ensemble perception, see Eqn. 1. Larger set sizes increase error in the individual task but have no effect or reduce error in the ensemble task. In the 50 ms condition, participants showed large locating errors across all set sizes (avg. > 8.5°), consistent with random responses to any object they perceived. Because this condition could not be captured by the models, it was excluded from further analyses.

Models

We created and explored three Bayesian perception models for our behavioral data, grounded in current theoretical accounts of ensemble perception. The first two models differ in their assumptions about how individual and ensemble percepts are computed. The first model, here referred to as the 'Individual Encoding Model' (IEM), assumes that the ensemble location is only evaluated on demand, reflecting the emerging view that ensemble perception arises from individual object representations (Harrison et al., 2021; Robinson & Brady, 2023). During scene presentation, individual target positions are encoded up to sensory noise $\vec{V}$ and WM limitations WML. During recall of individual object locations, the encoded latent positions are reproduced. If presented with an ensemble query, this model summarizes the encoded information as the average location around which all target objects were arranged (Harrison et al., 2021; Robinson & Brady, 2023). In contrast, the second model posits a generative process for the visual scene that first chooses a (latent) ensemble location (Chong & Treisman, 2005; Ward et al., 2016), relative to which the individual objects are arranged (Lew & Vul, 2015). Within this framework, the ensemble percept is grounded in the early encoding of individual elements. However, these individual object representations are subsequently integrated and no longer maintained as independent entities (Alvarez, 2011). The sensory signals resulting from this process may also be corrupted by sensory noise and the WM limitations. We refer to this process as the 'Ensemble Encoding Model' (EEM). The third model, here referred to as the 'Task Adapted Encoding Model' (TAEM), integrates these first two models, proposing that both processes coexist and are flexibly recruited depending on task affordances, with the IEM engaged during individual perception task and the EEM during ensemble perception tasks. Percepts arise in all models through approximate bayesian inference about the latent variables.

The coordinate system's origin was set to the center of the screen, as participants were instructed to start the exploration of the scene at this position during encoding. Owing to the inherent limitations in human vision, spatial encoding and retrieval accuracy, we incorporated zero-mean visual uncertainty into the models. The magnitude of this uncertainty reflects the evidence accumulation process about the objects' locations. Assuming that sensory noise is independent across time given object locations, the total sensory variance σ_v^2 at the end of the evidence accumulation process will scale inversely with the stimulus presentation interval Δt – the longer the participant looks at the display, the more evidence is collected (c.f. drift-diffusion models (Myers et al., 2022)). Also, assuming that the evidence for multiple objects has to be collected sequentially, the total sensory variance should scale with the number of objects n_{obj} in the display, since the accumulation time per object is (at least) halved if the number of objects is doubled. Thus

$$\sigma_v^2 = \sigma_v^2 \frac{\Delta t_0}{\Delta t} \frac{n_{\text{obj}}}{n_{\text{obj0}}} \tag{1}$$

Since we cannot measure sensory variance directly, we use the locating error in the individual task as an upper bound proxy, ignoring other source of noise and switch costs when more than one object has to be attended. Using our data, we find $\sigma_{v0} = 1.8°$ for $\Delta t_0 = 800ms$ and $n_{\text{obj0}} = 3$, see also Figure 2.

Our aim was to determine which of these models best explains the observed locating behavior in the two different locating tasks under the three set size and two presentation times (50 ms excluded) conditions, thereby gaining further insight into the possible computations of individual and ensemble perception.

Individual Encoding Model (IEM)

In the IEM the ensemble location $\vec{E}$ is computed by averaging individual object representations $\vec{X}_i$ (Harrison et al., 2021; Robinson & Brady, 2023). For K objects in a scene which are presented at locations $\vec{O}_i$, that differ from the internal representations $\vec{X}_i$ by independently drawn visual noise/uncertainty $\vec{V}_i$, the model assumptions are

$$\begin{aligned}
\vec{X}_i &\sim \mathcal{N}(\vec{0}, \Sigma_X) && (2)\\
\vec{V}_i &\sim \mathcal{N}(\vec{0}, \sigma_v^2 \cdot \mathbb{I}_2) && (3)\\
\vec{O}_i &= \vec{X}_i + \vec{V}_i && (4)\\
\vec{E} &= \frac{1}{K} \sum_{i=1}^{K} \vec{X}_i && (5)\\
\Sigma_X &\sim \mathcal{W}(4, 12/\sqrt{4} \cdot \mathbb{I}_2) && (6)
\end{aligned}$$

where $\mathcal{N}$ refers to a multivariate normal distribution and $\mathbb{I}_2$ is the 2×2 identity matrix. $\mathcal{W}(4, 12/\sqrt{4} \cdot \mathbb{I}_2)$ is a wide Wishart prior on the covariance matrix Σ_X with an expectation of 144 for the diagonal elements (parametrization via `scale_tril` in `PyTorch`), which reflects the experimental design: a standard

deviation of 12° was used for object placement around the screen center, across all test scenes used in the behavioral experiment.

Ensemble Encoding Model (EEM)

The EEM describes a generative process for scenes that independently draws the ensemble position $\vec{E}$ and the object positions $\vec{X}_i$ relative to $\vec{E}$ (Lew & Vul, 2015). In terms of our kitchen scenes (see Fig. 1), the ensemble position can be thought of as a reference location, similar to selecting a shelf, relative to which the individual objects are encoded and positioned. Visual uncertainty is the same as for the IEM. The model is therefore specified as

$$\vec{E} \sim \mathcal{N}(\vec{0}, \Sigma_E) \tag{7}$$
$$\vec{X}_i \sim \mathcal{N}(\vec{0}, \Sigma_X) \tag{8}$$
$$\vec{V}_i \sim \mathcal{N}(\vec{0}, \sigma_v^2 \cdot \mathbb{I}_2) \tag{9}$$
$$\vec{O}_i = \vec{E} + \vec{X}_i + \vec{V}_i \tag{10}$$
$$\Sigma_X \sim \mathcal{W}(4, 9/\sqrt{4} \cdot \mathbb{I}_2) \tag{11}$$
$$\Sigma_E \sim \mathcal{W}(4, 9/\sqrt{4} \cdot \mathbb{I}_2) \tag{12}$$

The prior diagonal expected covariances of 81 reflect the experimental design here, too. Both the standard deviation of the average object position from the screen center, as well as the average deviation of the individual object from the average position was 9°.

Task Adapted Encoding Model (TAEM)

The TAEM represents the idea that IEM and EEM coexist and are used depending on task-specific affordances. Therefore, the IEM is applied to the individual task and the EEM to the ensemble task.

In all three models, the ensemble percept is given by the posterior distribution of (the latent) $\vec{E}$ after encoding, i.e. we evaluate $P(\vec{E}|\vec{O}_{1,\dots,K}, \Sigma_X)$ for the IEM, and $P(\vec{E}|\vec{O}_{1,\dots,K}, \Sigma_X, \Sigma_E)$ for the EEM and TEAM. Similarly, the individual object percepts are given by $P(\vec{X}_{1,\dots,K}|\vec{O}_{1,\dots,K}, \Sigma_X)$ etc.. All posteriors can be computed analytically, details can be found in the model scripts: https://doi.org/10.60834/tam-datahub-15.

Working Memory Limitations

Given that human WM capacity is limited (Luck, 2008; Sims et al., 2012), performance might not only be constrained by the resources available for evidence accumulation, but also by WM limitations WML. If WML is less than the memory capacity required for storing the information encoded in the (exact) posterior relative to the prior, then a degradation of the internal representation of the scene might be expected. To implement WML, we follow an approach commonly used in resource-rational modeling (Binz et al., 2022): first, restate the inference as an optimization problem of a lower bound $\mathcal{L}$ on the model evidence. Let $\Theta = (\vec{X}_{1,\dots,K}, \vec{E})$ be the collection of latent variables of the model, $P(\Theta)$ the prior (see model descriptions above) and $Q(\Theta)$ the posterior, which represents the percepts. Then, for one trial

$$\mathcal{L} = \int d\Theta Q(\Theta) \log\left(P(\vec{O}_{1,\dots,K}|\Theta)\right) + D_{KL}(Q(\Theta) \,||\, P(\Theta)) \tag{13}$$

Second, constrain the Kullback-Leibler divergence $D_{KL}(Q(\Theta) \,||\, P(\Theta))$ between posterior and prior, which measures the number of bits needed for encoding the posterior relative to the prior, to be less than WML. We carry out this constrained optimization with the method of Lagrange multipliers, i.e. we maximize

$$\mathcal{L}(\lambda) = L + \lambda D_{KL}(Q(\Theta) \,||\, P(\Theta)) \tag{14}$$

where λ is the Lagrange multiplier. For a fixed λ, the maximization can be carried out analytically. If $D_{KL}(Q(\Theta) \,||\, P(\Theta)) \leq WML$ for $\lambda = 0$, then the scene encoding fits into the WM, and $Q(\Theta)$ equals the exact posterior. In this case, the model behaves like an ideal observer. If $D_{KL}(Q(\Theta) \,||\, P(\Theta)) > WML$, then we use interval bisection to find the λ for which $D_{KL}(Q(\Theta) \,||\, P(\Theta)) = WML$. Then, $Q(\Theta)$ will be an approximate posterior, representing a degraded scene encoding. Since WML might differ substantially between participants, we fit individual WMLs estimated from the location-task data.

Model Fitting

We fitted the models to simulated data (for recovery tests) and real data (for model evaluation) by maximizing the posterior probability of the model's predictions and parameters with respect to Σ_X, Σ_E and WML. To this end, we implemented all models in Python (version 3.12; Python Software Foundation, 2023) using the machine learning framework PyTorch (version 2.5; PyTorch Contributors, 2023) for automatic gradient computation and optimization. We performed 1000 steps with the Adam optimizer (Kingma & Ba, 2017), using a learning rate of 0.01. To compare the models, we used a Laplace approximation to the model evidence (Bishop, 2006).

Recovery tests

To verify model behavior, we tested the models on simulated data reflecting the experimental design. We assessed the recovery of latent variables, model parameters and predefined WM limits (raging from 20 to 50 bit). We further assessed model recovery using approximate bayesian model comparison to determine whether the correct model type could be identified. All tests showed that models behaved as expected (see the aforementioned repository for all test scripts and results), with model parameters well recovered (IEM: $r = 0.95$; EEM: $r = 0.92$; TAEM: $r = 0.87$). In terms of model recovery, the IEM was recovered unambiguously, with posterior probabilities of 100%. In contrast, recovery of the EEM and TAEM was less distinct, with posterior probabilities exceeding 90% in favor of the EEM in both cases, making it difficult to differentiate between these two models which warrants cautious interpretation.

Results

We illustrate the fit of the models to the behavioral data for an example trial in Figure 1. For individual object location reproduction, all three models made predictions close enough to the participant's estimation, i.e. within their covariance ellipses (upper panel of Fig. 1). We also computed the average angular distances between model posterior means and participants reports (see Fig.2). The results indicate that the model predicts participants' reports for individual object locations as well as the veridical position. In all cases the locating errors were smaller for smaller set sizes and for longer presentation times.

For the ensemble locating task, the EEM and TAEM covariance prediction contained the participant's response (lower panel of Fig. 1), the IEM was 'overconfident' as a consequence of Eqn. 5. Figure 2 shows that, in the 100 ms presentation condition, participants' reported ensemble locations were predicted equally well by all models and by the average of the veridical object locations, although the EEM and TAEM showed slightly larger deviations in the 3-object condition. However, this pattern changed in the 800 ms condition, particularly at higher set sizes. While for the 3-object condition average location predictions of the IEM continued to predict participants' reports similarly well as the average of the veridical locations, the EEM and TAEM deviated stronger, which is in line with the predicted deviations (squares) of these models. However, the IEM strongly underestimated its own locating error. In the 6- and 10-object condition, the IEM's overconfidence worsened, while EEM and TAEM showed good estimates of both mean location and expected errors.

In a next step, we computed an approximation to the model evidence via −2BIC (Bayesian Information Criterion) for each model at the individual participant level, both within and across experimental conditions. Across all conditions, the EEM provided the best fit for participants' behavior (see Fig. 3). The IEM showed the poorest fit overall, with model evidence values approximately 45 nats lower than those of the EEM and TAEM, indicating strong evidence against the IEM. This pattern was primarily driven by the IEM's performance in the ensemble task, where it produced systematically overconfident estimates and deviated from participants' reports (cf. Figure 2). In contrast, the TAEM appears to benefit from combining both approaches by flexibly engaging each model where it performs best, which may explain its only slightly poorer fit compared to the EEM (4 BIC in favor of the EEM) despite the stronger complexity penalty imposed by the −2BIC. However, given the limited discriminability between the EEM and TAEM (see section 'Recovery tests'), this comparison should be interpreted with caution.

To assess the variability associated with the latent variables X and E in the IEM and EEM models, we calculated the average estimated Σ across and within each combination of set size and presentation time condition (see Fig. 4 for the estimates across all conditions). In the EEM, the total variance of the object locations is divided between Σ_X and Σ_E,

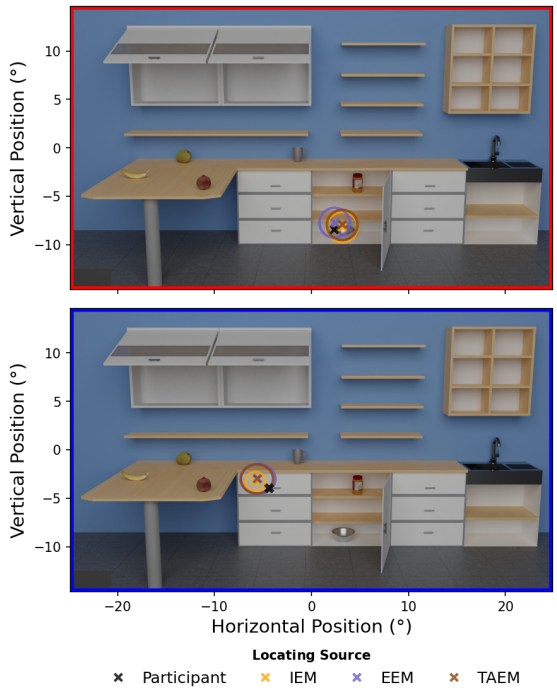

Figure 1: Example of a comparison of participants' locating behavior (black crosses) with predictions from the IEM (orange crosses), EEM (purple crosses) and TAEM (brown crosses) in an individual reproduction trial (first row, target: bowl) and an ensemble reproduction trial (second row) in the 6-object condition and with 800 ms presentation time. Colored circles indicate the estimated $\sqrt{\Sigma}$ of each model for the target position.

hence these (co)variances are smaller than Σ_X in the IEM. To assess the extent to which WM limitations influence the latent variables, we also report model estimates obtained without WM constraints (see Fig. 4). For the EEM, the estimated covariance parameter Σ was comparable between the limited and unlimited versions. In contrast, the IEM showed smaller values of Σ_X under WM limitations.

Lastly, using the ideal WM capacity estimated for each participant (see section 'Working Memory Limitations'), we examined whether individual WM capacity was related to localization performance (see Fig. 5). We emphasize that this correlation is not an out-of-sample prediction, as WM is inferred from the same dataset. To contextualize these estimates, we first assessed the WM requirements of the IEM and EEM under an effectively unlimited capacity constraint (set to 1000 bits, i.e. ideal observer models), revealing higher memory demands for the IEM (≈38 bits) than for the EEM (≈33 bits). When fitting WM capacity at the individual level, estimated capacities ranged from 17 to 35 bits, with a mean of 27.04 bits (std = 5.37). Participant's WM capacity and localization performance only correlated weakly negatively ($r = -0.24$), indicating a slight tendency for participants with higher WM capacity to perform better.

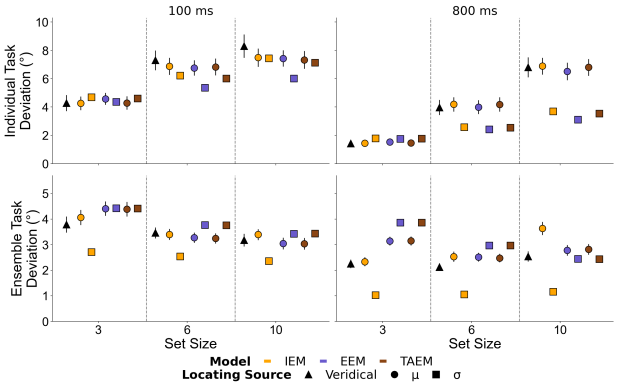

Figure 2: Mean locating errors (black) for the individual and ensemble tasks for all three set sizes with confidence intervals as error bars. Individual errors are relative to veridical positions, ensemble errors are relative to average veridical positions. Also shown are the median angular distances between model posterior means and participants reports for the IEM (orange), EEM (purple) and TAEM (brown).

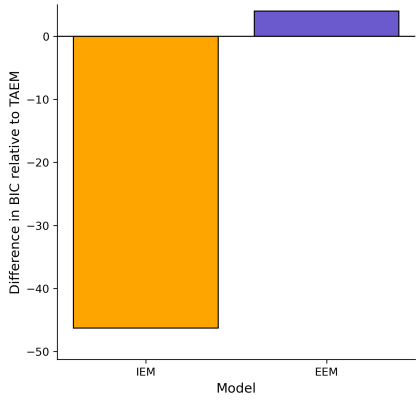

Figure 3: Mean difference in BIC across all experimental conditions between the TEAM and the IEM (orange) and the TEAM and the EEM (purple).

Discussion

Previous research has primarily examined individual and ensemble perception without accounting for the visual complexity of realistic scene contexts, which can either facilitate or hinder object perception (Draschkow et al., 2014; Ringer et al., 2021). To bridge this gap, we investigated how the perception of both individual objects and ensemble spatial information within naturalistic scenes is influenced by variations in set size and presentation time. Our behavioral data suggest that, consistent with previous studies using uniform backgrounds (Melcher et al., 2021; Neumann et al., 2018), individual object perception benefits more from smaller set sizes and longer presentation times, whereas ensemble perception remains unaffected by set size but also improves with longer presentation times.

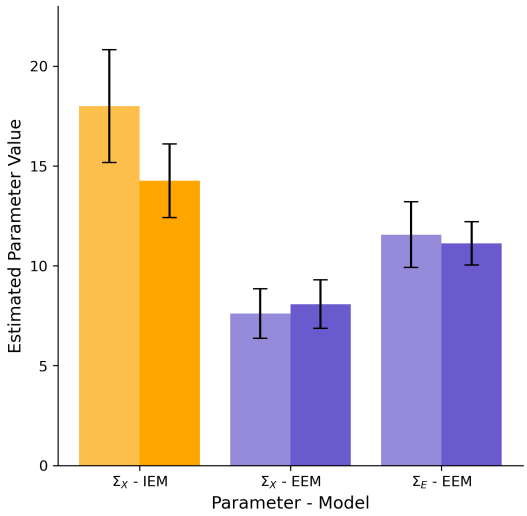

Figure 4: Estimated Σ along the horizontal coordinate of the IEM (orange) and EEM (purple) averaged across all set sizes and presentation time conditions, with standard deviations as error bars. Lighter colors represent the results of the respective models without WM limitations and darker colors with WM limitations.

Furthermore, we compared three computational models to address the ongoing debate about the computational goals of ensemble perception (Harrison et al., 2021; Robinson and Brady, 2023; Yildirim et al., 2018; for a review see Corbett et al., 2023): the Individual Encoding Model (IEM), the Ensemble Encoding Model (EEM) and a combination of both, the Task Adapted Encoding Model (TAEM). The IEM suggests that ensembles are constructed only when needed, relying on the averaging of individually encoded object percepts. In contrast, the EEM proposes that the ensemble's average position is directly processed and used to encode individual representations. In the TAEM, these two models are integrated by proposing that both processes coexist and are flexibly recruited depending on task demands, with the IEM engaged during individual perception tasks and the EEM during ensemble perception tasks. Contrary to recent studies (Harrison et al., 2021; Robinson & Brady, 2023; Utochkin et al., 2024), the EEM demonstrated a better fit for the observed locating behavior than the IEM, questioning the superior account of the IEM in ensemble perception.

Both, the IEM and EEM, captured performance in the individual localization task reasonably well, with participants' responses to the veridical object positions falling within the predicted confidence intervals (Fig. 2). In the ensemble task, however, the models diverged more clearly. Although the IEM produced accurate predictions in the 3-object condition at both presentation times, it consistently generated overly confident estimates and strongly deviated from participants' responses in the 10-object condition, particularly in the 800 ms condition. These systematic deviations likely account for the

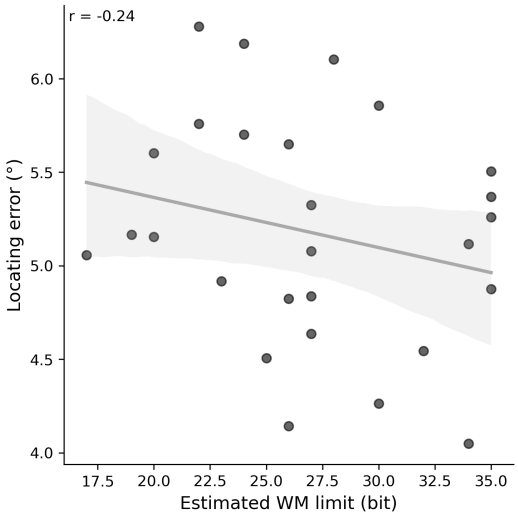

Figure 5: Correlation of the estimated working memory capacity of each participant with their overall locating error.

substantially poorer fit of the IEM relative to the EEM (exceeding 45 BIC points; Fig. 3), suggesting that ensemble-based representations provide a more robust account of ensemble localization especially under higher WM load.

The TAEM demonstrates that combining both approaches and flexibly engaging each where it performs best yields a substantially better fit than the IEM alone, even under the stronger complexity penalty imposed by the −2BIC. The relatively small differences between the TAEM and EEM underscore the importance of considering both processes rather than assuming a single dominant mechanism for multiple-object processing when processing naturalistic stimuli. At the same time, model discriminability between the EEM and TAEM was limited under the present experimental design. Our next step will be to focus on isolating the conditions under which the TAEM generates predictions that are distinct from those of the EEM.

From an efficient coding and resource-rational perspective (Binz et al., 2022; Sims et al., 2012), the relative utility of ensemble- versus item-based representations depends on set size and object distribution. Although ensemble perception can emerge with as few as two objects (Whitney & Yamanashi Leib, 2018), small and spatially dispersed sets may not benefit from ensemble processing (Klinghammer et al., 2017; Lew & Vul, 2015). In the present 3-object condition, the wide spatial distribution of items (up to 40°) likely reduced the advantage of extracting a shared ensemble position, resulting in larger discrepancies between EEM predictions and participants' ensemble reports compared to the IEM. Under such conditions, encoding individual object locations may constitute a more efficient strategy. As set size increases, however, the benefits of ensemble encoding become more pronounced, favoring summary representations over item-level encoding. This pattern aligns with resource-rational accounts, in which perceptual systems flexibly adopt representations that minimize information-processing demands while preserving task-relevant information (Binz et al., 2022), a principle captured by the TAEM's adaptive use of item-based and ensemble-based processes.

Estimated WM capacity showed substantial inter-individual variability, ranging from approximately 17 to 35 bits, with a mean of about 27 bits. Information-theoretic estimates suggest that representing a single object requires several bits (e.g., about five; (Sims et al., 2012)). Accordingly, the average capacity of about 27 bits would support reliable storage of only a limited number of objects. This is consistent with high performance in the 3-object condition, increasing difficulty in the 6-object condition and low performance in the 10-object condition. Notably, estimated memory demands were higher for the IEM (≈38 bits) than for the EEM (≈33 bits), indicating that item-based representations are more strongly affected by capacity constraints. Within this framework, the TAEM provides a natural account of how observers may flexibly shift between item-based and ensemble-based processing as WM demands increase across set size and presentation time.

One potential limitation of the present design is that participants were aware in advance of whether they would reproduce an individual or an ensemble position, which may have biased strategy selection in favor of ensemble perception. An argument for why our findings might still hold if participants were not primed for the reproduction task is given by the log-likelihoods calculated per task: the EEM accounted for individual localization performance comparably to the IEM ($EEM - IEM = 3.21$), while providing a superior explanation over the IEM of ensemble localization behavior ($EEM - IEM = -61.50$). This suggests that the advantage of the EEM is not solely driven by task-specific expectations but reflects a more general explanatory strength across task demands. Nevertheless, future experiments that manipulate or remove task predictability will be important to assess the flexibility of representational strategies more directly. Importantly, the explanatory scope of the EEM extends beyond ensemble perception, as it has also been proposed as a plausible mechanism for allocentric scene coding in spatial reaching tasks (Khoozani et al., 2019), highlighting its broader relevance for spatial cognition.

Our findings highlight the ensemble-based encoding of the EEM as the strongest and most efficient explanation of spatial behavior, extending beyond classic ensemble perception tasks. Nevertheless, the high performance of the TAEM underscores that perceptual strategies can be best understood as components of a flexible representational system, rather than mutually exclusive explanations. With larger or more informative datasets, the TAEM may prove superior in formal model comparisons. Future work should continue to develop models for resource-rational representations by integrating item- and ensemble-based perception and flexibly weighting these processes in a continuous fashion, while testing these models in carefully designed experimental scenarios.

Acknowledgements: This work was supported by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy (EXC 3066/1 "The Adaptive Mind", Project No. 533717223) and Collaborative Research Center SFB/TRR 135 (projects A4 and C6, under grant agreement no. 62202647).

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131. https://doi.org/https://doi.org/10.1016/j.tics.2011.01.003

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162. https://doi.org/10.1111/1467-9280.00327

Binz, M., Gershman, S. J., Schulz, E., & Endres, D. (2022). Heuristics from bounded meta-learned inference. *Psychological Review*. https://doi.org/10.1037/rev0000330

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. /bib / bishop / Bishop2006 / Pattern - Recognition - and - Machine - Learning - Christophe - M - Bishop.pdf,/bib/bishop/Bishop2006/978-0-387-31073-2_sm.pdf,https://www.microsoft.com/en-us/research/people/cmbishop/#!prml-book

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393–404. https://doi.org/10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, *45*(7), 891–900. https://doi.org/10.1016/j.visres.2004.10.004

Corbett, J. E., Utochkin, I., & Hochstein, S. (2023). *The pervasiveness of ensemble perception: Not just your average review*. Cambridge University Press.

Draschkow, D., Wolfe, J. M., & Võ, M. L.-H. (2014). Seek and you shall remember: Scene semantics interact with visual search to build better memories. *Journal of Vision*, *14*(8), 10. https://doi.org/10.1167/14.8.10

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751–R753. https://doi.org/10.1016/j.cub.2007.06.039

Harrison, W. J., McMaster, J. M., & Bays, P. M. (2021). Limited memory for ensemble statistics in visual change detection. *Cognition*, *214*, 104763. https://doi.org/10.1016/j.cognition.2021.104763

Khoozani, P., Schrater, P., Endres, D., Fiehler, K., & Blohm, G. (2019). Models of allocentric coding for reaching in naturalistic visual scenes. *Proceedings of CCN*. https://2019.ccneuro.org/proceedings/0000161.pdf

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. https://arxiv.org/abs/1412.6980

Klinghammer, M., Blohm, G., & Fiehler, K. (2017). Scene configuration and object reliability affect the use of allocentric information for memory-guided reaching. *Frontiers in Neuroscience*, *11*, 204. https://doi.org/10.3389/fnins.2017.00204

Lew, T. F., & Vul, E. (2015). Ensemble clustering in visual working memory biases location memories and reduces the weber noise of relative positions. *Journal of Vision*, *15*(4), 10. https://doi.org/10.1167/15.4.10

Luck, S. J. (2008). Visual short-term memory. In S. J. Luck & A. Hollingworth (Eds.), *Visual memory* (pp. 43–85). Oxford University Press.

Melcher, D., Huber-Huber, C., & Wutz, A. (2021). Enumerating the forest before the trees: The time courses of estimation-based and individuation-based numerical processing. *Attention, Perception, & Psychophysics*, *83*(3), 1215–1229. https://doi.org/10.3758/s13414-020-02137-5

Myers, C. E., Interian, A., & Moustafa, A. A. (2022). A practical introduction to using the drift diffusion model of decision-making in cognitive psychology, neuroscience, and health sciences. *Frontiers in Psychology*, *Volume 13 - 2022*. https://doi.org/10.3389/fpsyg.2022.1039172

Neumann, M. F., Ng, R., Rhodes, G., & Palermo, R. (2018). Ensemble coding of face identity is not independent of the coding of individual identity. *Quarterly Journal of Experimental Psychology*, *71*(6), 1357–1366. https://doi.org/10.1080/17470218.2017.1318409

Oriet, C., & Hozempa, K. (2016). Incidental statistical summary representation over time. *Journal of Vision*, *16*(3). https://doi.org/10.1167/16.3.3

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*, 195–203. https://doi.org/10.3758/s13428-018-01193-y

Perry, C. J., & Fallah, M. (2014). Feature integration and object representations along the dorsal stream visual hierarchy. *Frontiers in Computational Neuroscience*, *8*, 84. https://doi.org/10.3389/fncom.2014.00084

Python Software Foundation. (2023). Python (version 3.12) [Programming language]. https://www.python.org/

PyTorch Contributors. (2023). Pytorch (version 2.5) [Machine learning library]. https://pytorch.org/

Ringer, R. V., Coy, A. M., Larson, A. M., & Loschky, L. C. (2021). Investigating visual crowding of objects in complex real-world scenes. *I-Perception*, *12*(2), 1–23. https://doi.org/10.1177/2041669521994150

Robinson, M. M., & Brady, T. F. (2023). A quantitative model of ensemble perception as summed activation in feature space. *Nature Human Behaviour*, *7*, 1638–1651. https://doi.org/10.1038/s41562-023-01602-z

Sims, C. R., Jacobs, R. A., & 1, D. C. K. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*, 807–830. https://doi.org/10.1037/a0029856

Utochkin, I. S., Choi, J., & Chong, S. C. (2024). A population response model of ensemble perception. *Psychological Review*, *131*(1), 36–57. https://doi.org/10.1037/rev0000426

Ward, E. J., Bear, A., & Scholl, B. J. (2016). Can you perceive ensembles without perceiving individuals?: The role of statistical perception in determining whether awareness overflows access. *Cognition*, *152*, 78–86. https://doi.org/https://doi.org/10.1016/j.cognition.2016.01.010

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, *69*, 105–129. https://doi.org/10.1146/annurev-psych-010416-044232

Yildirim, I., Öğreden, O., & Boduroglu, A. (2018). Impact of spatial grouping on mean size estimation. *Attention, Perception, & Psychophysics*, *80*(7), 1847–1862. https://doi.org/10.3758/s13414-018-1560-5