Practice with uncertainty integration improves performance on a qualitatively different task and with new visualizations

Practice with uncertainty integration

Benjamin T. Files*

DEVCOM Army Research Laboratory, Army Research Directorate, Los Angeles, benjamin.t.files.civ@army.mil

Ashley H. Oiknine

DCS Corp, Los Angeles, aoiknine@dcscorp.com

Tiffany Raber

DEVCOM Army Research Laboratory, Army Research Directorate, Los Angeles, tiffany.raber.civ@army.mil

Bianca Dalangin

DCS Corp, Los Angeles, bdalangin@dcscorp.com

Kimberly A. Pollard

DEVCOM Army Research Laboratory, Army Research Directorate, Los Angeles, kimberly.a.pollard.civ@army.mil

Background: Every day, people must reason with uncertain information to make decisions that affect their lives and affect the performance of their jobs and organizations. Visualizations of data uncertainty can facilitate these decisions, but visualizations are often misunderstood or misused. Decision-making with uncertainty visualizations involves two major steps: 1) understanding the data from the visualization (*extraction*) and 2) preforming probabilistic reasoning to make decisions based on those data (*inference*). Previous research has demonstrated that deliberate practice with uncertainty visualizations can improve decision-making in abstract conditions, with performance gains even transferring to different visualization types (e.g., from ensemble to summary displays). This suggests that such practice may improve a person's facility with the inference step. However, we must first exclude two other possibilities. First, we must establish that such practice does not simply improve a person's ability to work with a specific display platform (e.g., static 2D vs. dynamic 3D) or specific context (e.g., abstract vs. concrete), and we must furthermore establish that practice does not simply improve the implementation of some rule-of-thumb heuristic that yields tolerable performance only with simple datasets. This study addresses whether the learning gains from abstract practice in one platform will transfer to more to concrete tasks on a different platform and with more complex data.

Objective: Here, we test the degree to which practice integrating multiple sources of uncertain information with abstract 2D summary or ensemble displays improves performance on transfer tasks involving decision-making with uncertainty displays on a 3D virtual sand table (3D tabletop terrain map) and with more sources of uncertain information to integrate.

Method: We conducted an online study with 378 participants who completed an uncertainty integration task in a 3D virtual sand table (3D tabletop terrain map) with story context using either summary or ensemble displays of uncertainty at different levels of data complexity. Participants had previously practiced in abstract 2D with the same display type (summary or ensemble), the other display type, or received no opportunity to practice. We analyzed response accuracy and speed and how these changed throughout the task.

Results: Results suggest that deliberate practice with abstract 2D uncertainty visualizations allows faster decision making in the new context and platform but does not improve accuracy. In the 3D task, the summary display generally yielded similar or better performance than the ensemble display. Learning gains from practice transferred to both same-type and different-type visualizations in the 3D platform and context, suggesting that practice in abstract 2D improved participants' ability to reason with uncertainty data, not just their ability to extract the data from the display platform, particular context, or specific visualization type. The faster performance also transferred to the higher complexity condition, suggesting that practice improved reasoning beyond use of simple rules of thumb.

Conclusions: The results suggest that practice with the abstract 2D task enhanced facility with the underlying probabilistic reasoning, which transferred to the new concrete, 3D, and more complex context, rather than just increasing visualization-specific understanding. This implies that deliberate practice can be a beneficial tool to improve reasoning with uncertainty, including across contexts, across visualization types, and across different levels of data complexity.

Materials: Stimuli, stimulus software, anonymized data, and analysis scripts and related code are available online at https://osf.io/5xdsg/.

CCS CONCEPTS • Human-centered computing~Visualization~Empirical studies in visualization • Human-centered computing~Visualization~Visualization theory, concepts and paradigms • Human-centered computing~Human computer interaction (HCI)~Interaction paradigms~Virtual reality

Additional Keywords and Phrases: Uncertainty visualization, reasoning with uncertainty, training, deliberate practice

1 INTRODUCTION

People often need to make consequential decisions based on evidence and forecasts that are uncertain. This occurs in domains such as healthcare (Han et al., 2019), weather (Joslyn & LeClerc, 2012), and military intelligence (Dhami et al., 2015). Increasingly, decisions need to be made based on multiple uncertain estimates (Padilla, Dryhurst, et al., 2021), but there has been limited past work examining how to communicate uncertain estimates in a way that supports optimal combination of the estimates (for exceptions, see Greis et al., 2018; Hegarty et al., 2016; Kusumastuti et al., 2022). What can be done to help people make decisions when they need to account for multiple uncertain estimates?

Well-designed visualizations of uncertainty can help people make better decisions (Padilla et al., 2020), but uncertainty is often omitted in reports of estimates and forecasts, in part because of concern that people will misinterpret or mistrust information presented with uncertainty (Hullman, 2020). Instructions and other explanations of how to use visualized uncertainty can help (Boone et al., 2019; Fiore et al., 2019; Song et al., 2019), but instruction might be insufficient. People without specific education or training in understanding uncertainty (Joslyn & LeClerc, 2012), as well as those with extensive training (Belia et al., 2005), often misinterpret common forms of uncertainty communication. When multiple sources of information are combined, the complexity of the task grows exponentially, potentially exacerbating any misunderstandings and overwhelming the decision-maker.

One promising line of research has aimed to help people make better use of visualizations of multiple, uncertain estimates in decision-making (Fernandes et al., 2018). In this work, decision-making performance improved over time when participants received immediate feedback after working with a variety of uncertainty communication techniques (Fernandes et al., 2018). Subsequent work showed that repeated practice paired with post-trial feedback improved

performance of both practiced and two un-practiced uncertainty visualizations (Kusumastuti et al., 2022). These observations fit in with the more general assessment that people might tend to make sub-optimal decisions when they have had no or limited opportunity to improve, but that probabilistic decision-making can improve notably when people receive feedback that lets them improve (Hertwig & Grüne-Yanoff, 2017; Lejarraga & Hertwig, 2021).

A cognitively inspired account of decision-making from visualized data lays out several steps in making a decision, distinguishing between message assembly (i.e., extracting meaning from the graphic) and inference (i.e., answering a question leading to a decision or action) (Padilla et al., 2018). Past work has found that practice with one type of visualization (e.g., ensemble or summary type) improves performance with an un-practiced visualization type, at least for abstract, 2D data visualizations (Kusumastuti et al., 2022). This suggests that at least some of the practice-driven improvement is taking place in the inference phase. However, it is possible that the observed gains were not the result of improvements in probabilistic inference but rather in other skills agnostic to visualization type, such as increased facility with the display platform and abstract context. It may also be the case that the practice afforded not a true improvement in probabilistic reasoning *per se* but rather greater facility in implementing rule-of-thumb heuristics that were just good enough to yield passable performance with the fairly simple data estimates they were shown (Kusumastuti et al., 2022). Here, we seek to rule out these alternative explanations and push this transfer farther, to see if performance boosts from practice transfer to new contexts and display platforms as well as to a change in visualization type and change in data complexity.

As the technological landscape evolves and increasingly offers alternative and enticing ways to display data, there is motivation in the training domain to understand how skills are obtained across different platforms and how they transfer from one type of platform to another. Platforms that utilize 3D displays are of particular interest and have been used to examine performance in terrain understanding (Smallman and Cook, 2010), representations in navigation (Pollard, Siriwardena, Krum, and Files, 2022; Liao, Dong, Peng, and Liu, 2016), surgical operations (Beattie, Hill, Horswill, Grove, and Stevenson, 2020; Ashraf, Collins, Whelan, O'Sullivan, and Balfe, 2015), weather forecasting (Hegarty, Smallman, Stull, and Canham, 2009), as well as many other situations. This popularity is due to 3D displays being perceived by users as more realistic (Hegarty, Smallman, Stull, and Canham, 2009), aesthetically appealing (Hegarty, Smallman, Stull, and Canham, 2009), and familiar with their respective field (Smallman and Cook, 2010). Notably, despite the rising interest in and preferred use of 3D environments, a number of studies have shown a decrease in performance rates, such as response time (Hegarty, Smallman, Stull, and Canham, 2009) and accuracy (Hegarty, Smallman, Stull, and Canham, 2009) when using 3D platforms as compared to the performance with their 2D counterparts. Nevertheless, there are still use cases in which 3D displays may be more beneficial than 2D displays. 3D displays have demonstrated their effectiveness in situations that involve shape understanding, in which users need to understand 3D spatial relationships (Ashraf, Collins, Whelan, O'Sullivan, and Balfe, 2015; Smallman and Cook, 2010), whereas situations that require relative positioning may be more effective for 2D displays (Smallman and Cook, 2010; Hegarty, Smallman, Stull, and Canham, 2009). This suggests that situations that require both 2D and 3D understanding may benefit from using both 2D and 3D platforms. Furthermore, as accessibility and distribution of hardware may differ across training sites, it is imperative to understand the extent to which skills learned from training transfer between 2D and 3D displays to ensure skill proficiency.

To begin to answer this question, we examined whether practice with an abstract 2D uncertainty integration task resulted in better performance in a similar task within a novel context and different display platform: a 3D virtual sand table (3D tabletop terrain map) environment in which participants were asked to identify the best place on a virtual terrain representation to drop a supply package, given two or four independent estimates of the ideal location (with visualized uncertainty for each estimate).

This virtual environment context placed additional demands on the user, requiring them to carry out the uncertainty integration task while managing these other demands. The first additional demand was that participants needed to navigate 3D space to get themselves into a position that afforded a preferred view of the task. The second additional demand was that the task was in the context of a 3D terrain. The verticality of the terrain was not directly relevant to the inference task, but it potentially distorted the visualized uncertainty and/or occluded some elements of the visualization. These additional demands combined to make a task with the same underlying probabilistic inference at its core as the 2D training task, but in a qualitatively different context.

We also examined whether practice-derived performance improvements would extend to performance of a task with increased complexity: during practice, participants were shown either two or three independent estimates of some abstract two-dimensional location, but in the 3D task, participants had to integrate either two or four independent estimates. If practice in a two and three-estimate task were reinforcing some generalizable inference, then we would expect performance benefits to transfer to a four-estimate task. However, if the practice is reinforcing the use of heuristics (i.e., approximations using visualization-based stand-ins or shortcuts) that function reasonably well in lower complexity problems, the benefits of such practice might begin to break down under higher complexity. For example, increasing the number of estimates increases the probability that at least one estimate has an error that is in the tails of its probability distribution (i.e., has more extreme levels of uncertainty). This would challenge heuristics that only work when estimates are clustered at the bulk of their distributions. If practice in the two or three-estimate task reinforced low-dimensional or other non-generalizable heuristics, then we would expect performance benefits to transfer to the virtual 3D task with two estimates (simple task) but not four estimates (complex task).

If practice with a visualized uncertainty integration task improves the speed and fluency of the underlying inference, then practice with such a task should lead to improvements on another task with the same underlying inference, regardless of the nature of the visualization itself (its visualization type, platform, context, or complexity). If, however, practice only improves the speed and fluency of the process of converting a visual array into a set of abstracted messages (i.e., data extraction), then practice might not lead to better performance on a task that demands a different message-assembly process, due to a difference in visualization type, context/visualization platform, or data complexity.

This leads us to the following hypotheses:

Hypotheses: Allowing a person to practice uncertainty integrations from data visualizations (with feedback) improves their underlying ability to perform probabilistic inferences. This improvement in probabilistic reasoning might carry over to other visualization types (H1), other visualization platforms or contexts (H2), and to more complex data sets (H3). If this hypothesis is true, practice-derived improvement should transfer to different visualization and decision-making tasks that nonetheless require the same underlying probabilistic inference. We thus would expect practiced participants to perform better (greater speed and/or accuracy) on a transfer task as compared to un-practiced participants, even when the visualization type is different (H1), and the visualization platform and context is different (H2), and the level of data complexity is greater (H3).

The alternative hypothesis (HA) is that practicing uncertainty integrations from data visualizations (with feedback) merely improves other underlying skills without also improving the ability to perform probabilistic inferences. If HA is true, we would expect that practice would not improve performance on other visualization types (HA1), and/or not on other visualization platforms/contexts (HA2), and/or not on more complex (more difficult) data sets (HA3).

To test these hypotheses, we conducted an experiment in which two thirds of the participants were assigned to practice uncertainty integration of two and three independent, bivariate normal estimates of a location in abstract, 2D space. Practice used one of two visualizations of uncertainty: an ensemble representation showing a random sample from the underlying

probability distribution or a summary representation showing 50% and 95% confidence ellipses around the estimates (Figure 1). The remaining third were assigned to the unpracticed condition. Both practiced and unpracticed participants then did a similar uncertainty integration task in the context of a 3D virtual sand table environment (Figure 2). Participants were also randomly assigned to do the 3D task with either an ensemble/scatter visualization or a summary/ellipse visualization. Performance on the 3D task was assessed in terms of accuracy and response time to better understand the generalizability of practice effects on uncertainty integration with visualized uncertainty.

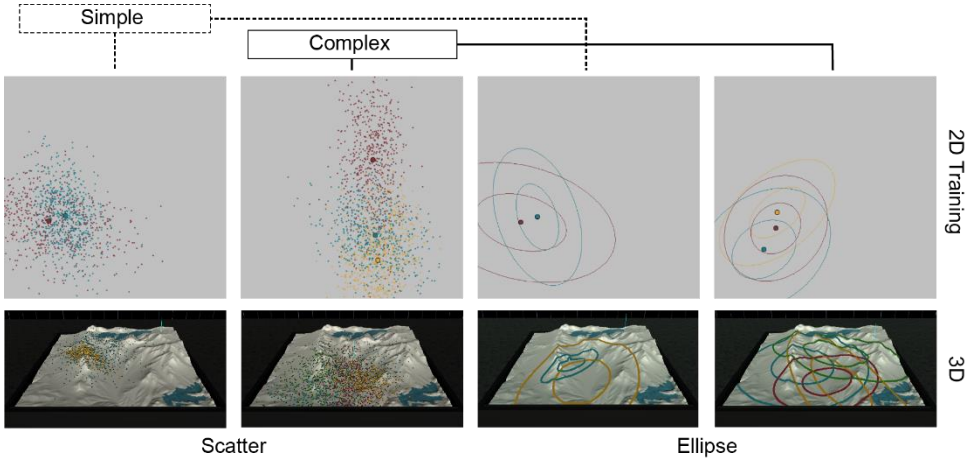

Figure 1. Examples of stimuli in the 2D training task (top row) and the 3D task (bottom row). Scatter stimuli (left half) and ellipse stimuli (right half) represent the underlying 2, 3, or 4 independent estimates and their uncertainties.

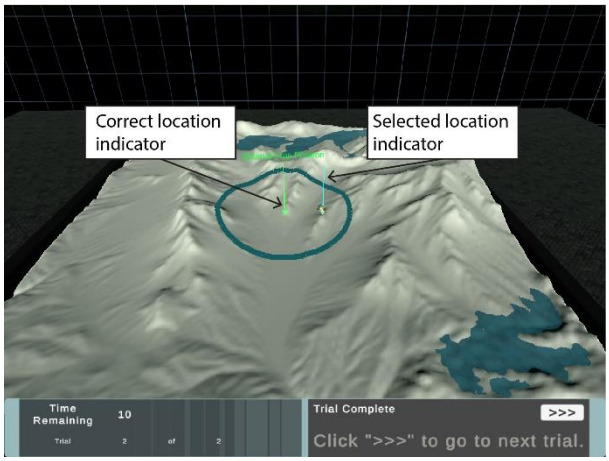

Figure 2. Image of 3D qualifying task with sand table and visualization developed using WebGL 2.0. Turquoise rod indicates the user's cursor and selection. Green rod shows the optimal drop position after the user selects a location.

2 METHOD

This work was not pre-registered, so it can be considered exploratory. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

2.1 Participants

This study was conducted online with participants recruited via the web-based service Prolific Academic. People who participated in our previous related studies were excluded from recruitment. Individuals who reported being 18 years of age or older, being fluent in English, and having normal hearing and vision were invited to join the qualifying session of the study. The qualifying session (described below) was used to gather additional demographic and background information and to screen for additional inclusion criteria for the main session: normal color vision as confirmed via a web-based 14-plate Ishihara color vision test and using a desktop or laptop computer with a suitable pointing device (i.e., not a touch screen) that was capable of running the web-based virutal environment used in the main session. Participants who qualified were invited to take part in the main study. Participants were paid for participating in the qualifier and again if they participated in the main study.

Our goal was to obtain data from 60 participants per condition. This number came from a simulation-based power analysis in which we found that number to be sufficient to have a >80% chance to detect a between-subjects effect in initial task performance due to differences in visualization used in earlier work with a similar paradigm (Kusumastuti et al., 2022).

2.2 Qualifier

Participants in the qualifying session responded to demographic questions, provided information about computer hardware used to complete the experiment, and completed a 14-plate Ishihara color vision test. Participants also completed the Regulatory Focus Questionnaire (Higgins et al., 2001), the Game Use Inventory (Metcalfe et al., 2015), The Big 5 Inventory (John et al., 1999) and a task in the same virtual environment used in the main experiment. The qualifying task was used to determine that the participant's computer could adequately display our virtual environment and that the user could successfully interact with it. The task itself was to mark a location at the center of a ring displayed on the terrain of a 3D virtual sand table (Figure 2). After the qualifying task, participants completed a 9-item exit questionnaire regarding their experience with the task and asking about any technical difficulties. Those who met eligibility criteria, successfully completed the qualifying task (completed all three trials of the qualifier task, responded to at least one trial before the 20-second deadline, and maintained at least 30 frames per second throughout), and reported no technical difficulties were invited to participate in the main experiment.

2.3 Study Design

Our study used a 3 x 2 x 2 between-by-within design with three between-subjects practice conditions, two between-subjects transfer conditions, and a within-subjects complexity condition. Participants practiced in three ways: with summary (ellipse) displays of uncertainty in an abstract 2D plane, with ensemble (scatter) displays of uncertainty in an abstract 2D plane, and a no-practice control. The transfer conditions involved a terrain model shown as a virtual sand table, on which either summary (ellipse) displays were shown on the 3D terrain or ensemble (scatter) displays were shown on the 3D terrain. Participants were randomly assigned to one of the six between-subjects conditions. Additional within-subjects variables included two levels of difficulty experienced by each participant during the practice and transfer phases.

2.4 Stimuli

The 2D practice task and the 3D task (both described below) both used stimuli generated with the same process. Underlying each stimulus is a set of 2, 3, or 4 estimates each having bivariate normal probability distributions. The covariance matrix for each probability distribution was randomly generated with constraints to produce a set of stimuli with a variety of variances and covariances. The center of each estimate was offset from a common point (selected randomly from a uniform

distribution over a square with sides half the length of the stimulus and centered on the middle of the stimulus) by a random draw from its respective probability distribution. These stimuli are veridical in the sense that the stimulus generating process is consistent with the visualized probability distributions.

Each estimate was displayed either using an ensemble of points drawn randomly from the probability distribution of the estimate (scatter stimuli) or using summary ellipses that enclosed the central 50% and 95% of the probability mass of the probability distribution (ellipse stimuli). Example stimuli appear in Figure 1.

2.5 Main Experiment Procedure

The main experiment consisted of an uncertainty integration practice task that used 2D abstract stimuli (2D practice task), a 3D virtual environment-based uncertainty integration transfer task (3D task), and a battery of questionnaires administered throughout the experiment. The questionnaires were the Graph Literacy Scale (Galesic & Garcia-Retamero, 2011), the Berlin Numeracy Scale (Cokely et. al, 2012), Cognitive Reflection Task (Frederick, 2002), and a Computer Self Efficacy Questionnaire (Compeau & Higgins, 1995). Participants were randomly assigned to a practice condition (2D abstract scatter stimuli, 2D abstract ellipse stimuli, or no practice control) and a transfer condition (3D scatter sand table or 3D ellipse sand table). All participants, including those in the unpracticed control condition, viewed an 8-minute instruction video describing the 2D practice task. This video described how to interpret both the scatter and the ellipse stimuli, and it explained how to combine two or more estimates to arrive at a final judgement. After the video, participants took a comprehension test and did three multiple-choice practice trials. Incorrect answers were given corrective feedback; participants moved on when they got all questions correct or after three tries. Participants then completed the 2D practice task with either the scatter or ellipse visualization or skipped practice (unpracticed control condition). Finally, participants read on-screen instructions for the 3D virtual environment-based transfer task and then completed that task. Following the task, participants took the Intrinsic Motivation Inventory (Ryan, 1982) and answered an exit questionnaire.

2.5.1 Abstract 2D practice task

For each trial during the 2D practice task, participants were presented with a flat plane which contained a hypothetical hidden location. Two or three independent estimates of the hidden location were shown in different colors, and the participant's task was to incorporate information from all displayed estimates to select the most likely location and click on that location with their mouse. This uncertainty integration practice task is identical to the practice task we have used previously (Kusumastuti et al., 2022); additional details, including additional images of the stimuli used and code to generate these stimuli, are available online https://osf.io/5xdsg/. A demonstration version of the task is available online https://osf.io/txqk9/wiki/home/.

Practice used only the assigned visualization type and contained 8 blocks of 30 trials each. In the first three blocks, trials contained 2 estimates. In the subsequent three blocks, there were 3 estimates, and the last two blocks had both 2 and 3 estimates randomly intermixed. Participants were given 10s to respond with a mouse click. After each trial, feedback displayed the participant's selected location and the most likely (best answer) location. Points for accuracy (out of 100 possible) and speed (out of 20 possible) were displayed, as was a cumulative total score for the block. Accuracy points were computed as 100 times the ratio of the joint likelihood (i.e., the product of the probability density evaluated at a given location for all estimates) at the selected point divided by the joint likelihood at the best point (referred to as the *accuracy score*). Speed points were computed as:

$$speed\ points = 20\left(1 - \frac{rt}{rt_{max}}\right)^{2.5}$$

The exponent in this function was set so it would have a steep early drop-off to encourage fast responding but not so steep as to vanish to zero too early in the response window.

2.5.2 3D virtual environment-based transfer task

For the 3D virtual environment-based transfer task, participants were randomly assigned to one of two conditions, 3D scatter or 3D ellipse, after which they viewed on-screen instructions for the 3D task. After reviewing the task instructions, participants were presented with the same virtual sand table interface from the qualifying task. The 3D task shared the same underlying inference task with the 2D practice task: combine two or more independent estimates to select the best location. In other ways, the 3D task differed from the 2D practice task. The 3D task was in the context of a virtual sand table area where participants could use keyboard and mouse (or other pointing device) to move their perspective relative to the sand table. Participant viewpoints were not allowed to pass within or through the virtual sand table. However, participants were free to move the camera up as high as they cared to. For each trial, we recorded the integrated distance moved and camera angle change. Participants started the trial 4 units of distance (virtual meters) from the center of the sand table. Participants used their mouse to rotate the camera as well as to select a location on the terrain. Participants could confirm a selected location on the terrain by pressing the space bar, ending the trial. After the participant confirmed the location, a small package was animated falling onto the selected location. Post-trial feedback showed the selected location, the best location, points for accuracy and speed, and cumulative scores. Because of the requirement to maneuver the viewpoint and the more complex method of selecting a location, participants had up to 20 seconds to confirm their response. A countdown timer indicated time left, and this timer turned red when less than 5 seconds remained. The 3D task contained two blocks of 60 trials. The first block used 2 estimates for each trial and the second block used 4 estimates. The experiment software and a demonstration version of the task is available online, https://osf.io/txqk9/wiki/home/.

All testing was approved by the Institutional Review Board of the U.S. Army Research Laboratory under protocols ARL-21-070 and ARL-22-062.

2.6 Analysis

The main goal of analyses was to assess effects of the 2D practice task on performance of the 3D task, with emphasis on effects at the start and end of each block as well as how performance changed over time. Performance was analyzed in terms of three separate performance models: two for accuracy (accuracy score, error distance), and one for speed (response time). The accuracy score was the basis for performance feedback to participants, and it ranges from zero to one, with a score of one representing maximum accuracy. In another model, accuracy was analyzed in terms of the Euclidean distance between the given response and the correct response. Although the points-based feedback emphasized the accuracy score, both the location of the given response and of the correct response were visible during the feedback screen, so some participants may have construed their performance as related to the linear distance from their response to the correct one. The feedback used a non-linear transform to convert response times into points for response speed, but here we analyze response time directly.

Each outcome was analyzed separately using Bayesian hierarchical models. The models were implemented in Stan (Stan Development Team, 2022), and code for the models can be found online at https://osf.io/5xdsg/. Samples were generated from these models using the No-U-Turn sampler (Hoffman & Gelman, 2014) with four independent chains each having 2000 warm-up samples and 1000 post-warm-up samples. All sample sets had no divergences, all chains had BFMI > 0.3, and all parameters had effective sample sizes > 100 and r-hat < 1.1.

Because we expected performance to change over time, the models all made use of an exponential learning function:

$$f(t|\pi,\delta,\alpha) = \pi + \delta(1 - e^{-\alpha t}) \quad (1)$$

where time *t* is in terms of the proportion of trials completed, π is the initial value of the function, $\pi + \delta$ is the value the function approaches after all learning is complete (i.e., at time infinity), and the α parameter, which is constrained to be greater than 1, is related to the learning rate. Note, because δ is unconstrained, this function can model performance that gets worse over time as well.

2.6.1 Accuracy score Model

To model accuracy score, we used a zero-inflated beta model. A zero-inflated beta model is an appropriate choice because haphazard or accidental clicks are likely to evaluate to zeroes, while the rest of the clicks will have ratio values between 0 and 1. Although the accuracy score is theoretically never exactly zero, in practice we recorded the ratio out to eight figures past the decimal point, so there was a reasonable possibility that an inaccurate response would evaluate to zero. A similar theoretical issue applies to ratios very close to one, but this occurred in only 22 trials (0.05%) of cases, so we subtracted 1e-8 from those rather than using a zero-one-inflated model. Under this zero-inflated beta model, the probability of an accuracy score *R* is given as:

$$p(R) = \begin{cases} Z, & if\ R = 0 \\ (1 - Z)\mathcal{B}(R|\mu,\kappa), & otherwise \end{cases} \quad (2)$$

The parameters of this model are the zero-inflation parameter *Z*, which is the probability a response will evaluate to zero, and the two parameters of the beta distribution; we used the mean μ, and sample size (or more concisely, *precision*) κ parameterization to facilitate interpretation. Each of these parameters changed with time following the exponential learning function defined in (1) above:

$$logit(Z) = f(t|\pi_Z,\delta_Z,\alpha_Z) + m_Z + c_Z \quad (3)$$
$$logit(\mu) = f(t|\pi_\mu,\delta_\mu,\alpha_\mu) + m_\mu + c_\mu \quad (4)$$
$$\log(\kappa - 1) = f(t|\pi_\kappa,\delta_\kappa,\alpha_\kappa) + m_\kappa + c_\kappa \quad (5)$$

Each of the parameters of the exponential learning functions was indexed by condition (i.e., the combination of practice condition (unpracticed, practiced with ellipse, practiced with scatter), 3D stimulus condition (ellipse, scatter), and block (simple 2-estimate or complex 4-estimate), and these were modeled as drawn from a normal distribution with a common offset θ and standard deviation τ. Using $\pi_{\mu,i}$ as a representative example:

$$\begin{aligned} \pi_{\mu,i} &\sim \mathcal{N}(\theta_\mu, \tau_\mu) \\ \theta_\mu &\sim \mathcal{N}(0,3) \\ \tau_\mu &\sim \mathcal{N}^+(0,2) \end{aligned} \quad (6)$$

The terms *m* and *c* in (3-5) are random effects of stimulus and participant, respectively. The stimulus random effects were modeled with a normal distribution with mean zero and different standard deviations for 2-estimate stimuli (i.e., $m_{simple} \sim \mathcal{N}(0,\sigma_{simple})$, $\sigma_{simple} \sim \mathcal{N}^+(0,1)$) and 4-estimate stimuli (i.e., $m_{complex} \sim \mathcal{N}(0,\sigma_{complex})$, $\sigma_{complex} \sim \mathcal{N}^+(0,1)$). The participant random effects were modeled similarly, but the random effects in the 2-estimate and 4- estimate conditions were potentially correlated.

Priors for the common offset of the π parameter in (6) were normal with mean zero and standard deviation 3 and were normal with mean zero and standard deviation 2 for δ and $\log(\alpha - 1)$. All other priors were standard normal or standard half-normal, as appropriate.

2.6.2 Error Distance Model

Analysis of error distance used a normalized Cartesian coordinate system, where terrain extended from -0.5, -0.5 to 0.5, 0.5. Response error *D* was computed as the Euclidian distance between a response at location *X* and the best answer at location *Y*, i.e., $D = |X - Y|$. To account for responses very far from the correct answer, we used a mixture model that had some probability a response was selected effectively uniformly at random from the entire response area and otherwise came from a spherical normal distribution centered on the best response:

$$p(X) = \begin{cases} \mathcal{U}(X), with\ probability\ u \\ \mathcal{N}(D|0, \Sigma), with\ probability\ 1 - u \end{cases} \quad (7)$$

Where $\mathcal{U}$ is the bivariate uniform distribution over (-0.5, -0.5) to (0.5, 0.5) and $\mathcal{N}$ is the bivariate normal distribution with covariance matrix $\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$. The parameters of this model are therefore u, the probability of a uniform response, and σ, the standard deviation of the probability distribution. These were modeled with exponential learning functions (1), as in the accuracy score model:

$$\log \sigma = f(t|\pi_\sigma, \delta_\sigma, \alpha_\sigma) + m_\sigma + c_\sigma$$
$$logit(u) = f(t|\pi_u, \delta_u, \alpha_u) + m_u + c_u$$

The prior on the common offset of π_σ was normal with mean -2 and standard deviation 1, and on π_u was mean -3, standard deviation 1. These were chosen to reflect the expectation that the probability of uniform responding was low and that when they were non-uniform, the errors should be somewhat less than 0.5. Standard deviations on condition effects were half-normal with standard deviation 0.5 as were the standard deviations for random effects of stimulus and participant. The exception was the α parameters where the prior was normal with mean zero and standard deviation 0.2.

2.6.3 Response Time Model

Response times were modeled using a truncated lognormal distribution:

$$\log(RT) = \mathcal{N}(\mu, \sigma)T[, \log(RT_{max})] \quad (8)$$

If a participant did not give a response by the time-out of RT_{max}=20s, that response time was recorded as 20s, and the probability of a response at 20 was modeled as the total probability mass of the untruncated response time distribution above 20. As in the previous models, the parameters were modeled with exponential learning functions:

$$\mu = f(t|\pi_\mu, \delta_\mu, \alpha_\mu) + m_\mu + c_\mu$$
$$\log \sigma = f(t|\pi_\sigma, \delta_\sigma, \alpha_\sigma) + m_\sigma + c_\sigma$$

The prior for the common offset on π_μ was normal with mean 0 and standard deviation 1.5, and for $\log(\pi_\sigma)$ was normal with mean -0.5 and standard deviation 0.5. For both $\log(\alpha_\mu - 1)$ and $\log(\alpha_\sigma - 1)$, the prior was normal with mean -2 and standard deviation 1. Priors on the standard deviations of the random effects of stimuli were half-normal with mean zero and standard deviation 0.2; the priors on the standard deviations of the random effects of participant were half-normal with mean zero and standard deviation 0.1.

As an overall note, the priors we selected were chosen to provide enough information to encourage solutions at the correct scale rather than expressing some theoretically driven expectation about the solutions. Importantly, the priors for the different practice conditions were identical.

3 RESULTS

A total of 390 participants completed the experiment. Twelve were excluded from analysis (8 due to having an average frame rate lower than 20 fps, 3 due to failing to respond to 30 or more trials out of 120, and 1 due to achieving an overall average accuracy lower than 10%). The remaining 378 were included in analysis. Because data were collected asynchronously in parallel, different numbers of participants were included in the six experimental conditions (Table 1). The mean self-reported age of these participants was 27 (median: 25, range 18-67). Detailed demographics are available in an online supplement, https://osf.io/5xdsg/.

Table 1. Participant count per condition.

	Practice condition		
stimulus	Ellipse	Scatter	Unpracticed
Ellipse	61	62	66
Scatter	60	60	69

Participants in the practiced conditions took on average 89 min (std 27 min, range 43 to 209 min) from start to finish. Participants in the unpracticed condition took on average 69 min (std 25 min, range 33 to 190 min) from start to finish.

Performance on the uncertainty integration task was analyzed in terms of three separate performance metrics: two for accuracy (accuracy score, error distance), and one for speed (response time). For each of these three models, we summarize the main results with posterior predictive group mean time-courses and emphasize the initial and final performance within a block. Each model has two or three parameters that vary over time, and time-courses for these parameters are also shown.

3.1 Accuracy score

Accuracy, measured with the ratio of the likelihood at the selected location to the likelihood at the best location and averaged over time, appears in Figure 3. Collapsing over time and practice condition, posterior mean accuracy with ellipse stimuli and the simple 2-estimate task was 0.791, 95% credible interval [0.786, 0.797], ellipse stimuli and the complex 4-estimate task was 0.572 [0.566, 0.580], scatter stimuli and the simple 2-estimate task was 0.725 [0.719, 0.731], and scatter stimuli and the complex 4-estimate task was 0.567 [0.560, 0.574]. Average accuracy was lower for the more complex 4-estimate task than the simple 2-estimate task. The mean difference was 0.189, 95% credible interval [0.182, 0.195]. Considering only the simple 2-estimate task, accuracy was higher for ellipse stimuli than scatter stimuli, with a mean difference of 0.066 [0.059, 0.074], but accuracy was similar for ellipse and scatter stimuli in the complex 4-estimate task, with a mean difference of 0.006 [-0.004, 0.015]. Effects of practice condition within a given task complexity and stimulus type were small and included zero in their 95% credible intervals with the exception of the simple 2-estimate task with ellipse stimuli. There, accuracy of participants who did not practice was higher than the accuracy of those who practiced with scatter and ellipse, with mean differences of 0.013 [0.001, 0.025] and 0.03 [0.019, 0.044], respectively, and participants who practiced with scatter had higher accuracy than those who practiced with ellipses, with mean difference of 0.018 [0.005, 0.018].

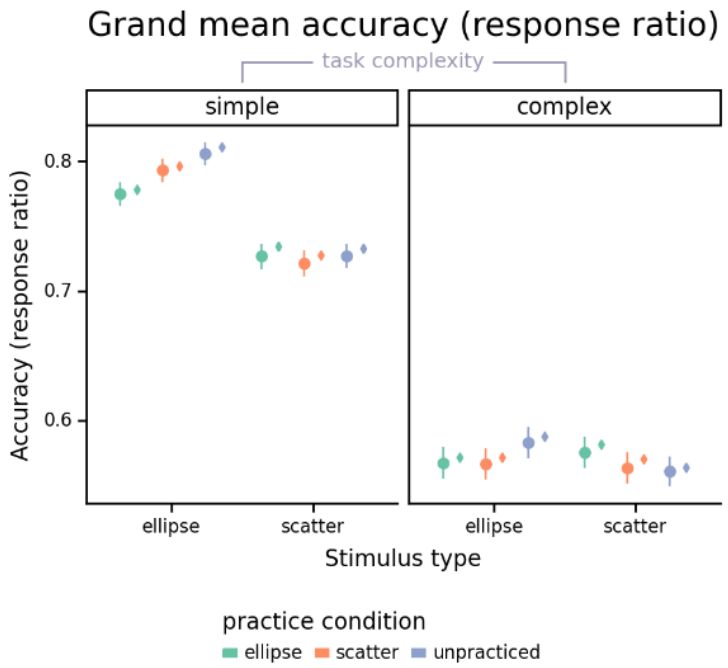

Figure 3. Grand mean accuracy (measured as the ratio of the likelihood at the selected location to the likelihood at the best location) by practice condition, stimulus type, and task complexity. Circles show the posterior predictive mean and credible interval. Adjacent diamonds show the sample grand means.

Accuracy over time, along with modeling results, are in Figure 4. Differences in accuracy score due to practice condition for a given stimulus type and task complexity were all small and the 95% credible interval included zero in all cases both initially and at the end of the block. When comparing effects of stimulus type by practice condition and task complexity, the largest difference arose at the end of the simple 2-estimate task block: accuracy with ellipse stimuli was higher than with scatter stimuli at the end of the simple 2-estimate task block for participants who practiced with scatter stimuli (mean difference 0.094, 95% credible interval [0.021, 0.171]) and who did not practice (mean difference 0.092 [0.024, 0.157]), and less so for those who practiced with ellipse stimuli (mean difference 0.045 [-0.030, 0.116]).

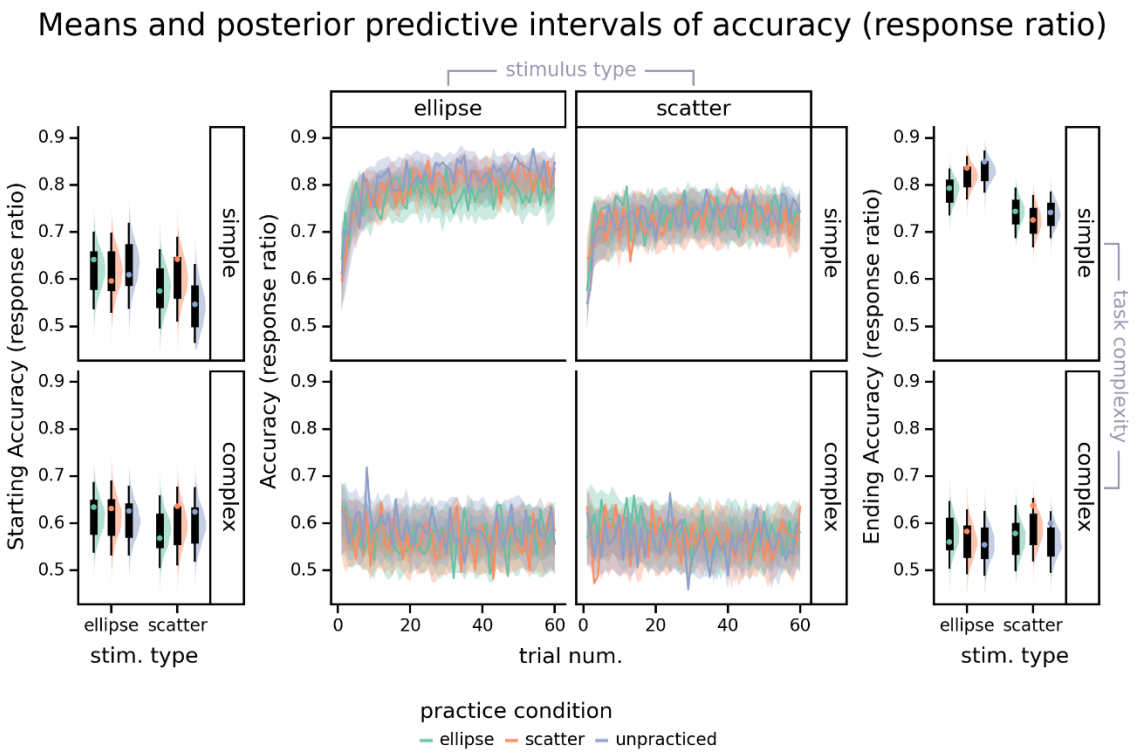

Figure 4. Posterior predictive summaries of accuracy, calculated as the ratio of the likelihood at the selected location to the likelihood at the most likely location. Posterior summaries include central 66% and 95% credible intervals of the posterior predictive distribution of starting accuracy (accuracy on the first trial of the block, left) and ending accuracy (accuracy on the last trial of the block, right). Dots show the experiment sample mean. The central panel shows how accuracy score changed over time, with light ribbons covering a 95% credible interval and darker ribbons covering a 66% credible interval. Lines show the experiment sample mean.

There were some small differences in the parameters of the model. The zero-inflated beta model of accuracy (2) has three parameters: μ, the mean of the beta distribution, indicates a typical response ratio; κ, the precision parameter of the beta distribution, increases as responses become more tightly clustered around that mean; and Z, the zero-inflation parameter, which is the probability that the response ratio is zero (which is theoretically impossible in a pure beta model). The change in these parameters over time was modeled with exponential growth functions. Parameter time series estimates of the zero-inflated beta model appear in Figure 5. For participants who did the simple 2-estimate 3D task with ellipse stimuli, both the beta distribution mean, μ, and the beta precision parameter, κ, increased fastest for participants who practiced with ellipse stimuli, compared to those who practiced with scatter or did not practice, with posterior mean differences in the rate parameter for μ of -21.8 [-56.6, -1.1] and -27.4 [-61.6, -8.7], respectively and for κ of -20.3 [-58.0, -1.7] and -24.5, [-61.0, -6.5], respectively. This shows that for participants who practiced with ellipse stimuli, performance plateaued rapidly compared to those in the other practice conditions. However, examining performance parameters at the end of the block on ellipse stimuli and the simple 2-estimate task revealed that the mean parameter, μ, was higher for those who did not practice compared to those who practiced with ellipse stimuli, with mean difference of 0.04 [0.02, 0.06]. Similarly, the precision parameter, κ, was higher for those who did not practice compared to those who practiced with ellipse stimuli, with mean difference of 0.860 [0.055, 1.697]. This suggests that with ellipse stimuli, unpracticed

participants ended up with slightly more accurate and more precise performance compared to participants who practiced with ellipse stimuli.

Unpracticed participants had a higher initial zero-inflation parameter, Z, in the simple 2-estimate block relative to those who practiced with ellipse stimuli in both the ellipse condition (mean difference 0.044 [0.003, 0.098]) and the scatter condition (mean difference 0.071 [0.021, 0.138]), but zero-inflation, Z, for all practice conditions was similar by the end of the simple 2-estimate block. This shows that unpracticed participants had a higher risk, relative to trained participants, of providing a wildly incorrect response near the beginning of the simple 2-estimate block, but that this risk differential was eliminated over the course of the block.

Taken together, the differences in modeled performance parameters over time showed that in the simple 2-estimate task with ellipse stimuli, unpracticed participants had an initially relatively high risk of very inaccurate responding and improved a bit slower than those in the other conditions, but their eventual performance parameters were higher compared to those in the other conditions. However, these differences in model parameters did not lead to a substantial difference in posterior predicted group mean accuracy.

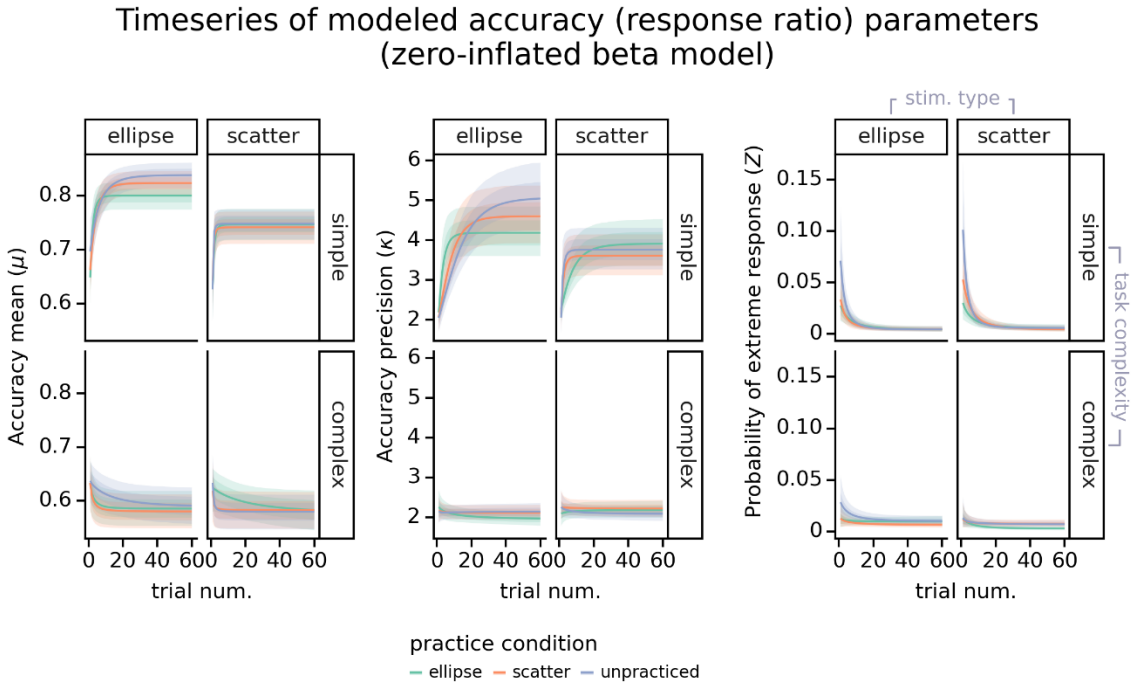

Figure 5. Posterior timeseries of the parameters of the zero-inflated beta model by task complexity (rows) and stimulus type (columns). The zero-inflated beta model includes a mean parameter μ, modeling participants' typical accuracy in terms of accuracy score (left), a sample size parameter κ, modeling the dispersion of their accuracy values (middle), and a zero-inflation parameter Z, modeling the probability of an extremely erroneous response (right). Time series show mean, central 66% (dark ribbons) and 95% (light ribbons) credible intervals. Trial time is the proportion of the 60-trial block elapsed.

The accuracy model incorporated random effects of participant, with correlations among the random effects for a participant when they were in the simple 2-estimate task block and in the complex 4-estimate task block. Random effects were highly correlated in each of the three parameters of the model: beta distribution mean, μ, r = .819, 95% credible

interval [.766, .862], beta distribution precision, κ, r = .728 [.637, .809], and zero-inflation, Z, r = .522 [.288, .732]. These high correlations indicate that participants' individual differences—perhaps arising from preexisting skill levels or dispositional traits—consistently affect performance. Future work should endeavor to identify what individual traits may underlie these effects.

3.2 Error Distance

Error distance is the distance between the best answer and the given answer in terms of normalized screen units. It differs from the accuracy analysis, above, because error distance ignores the likelihood at the selected location. Group mean and posterior predictive intervals for error distance appear in Figure 6. Averaging over practice condition, posterior predicted errors were smaller for ellipse stimuli than scatter stimuli in the simple 2-estimate task, mean difference .011, [.008, .013]. In the complex 4-estimate task block, the two stimulus types had similar posterior predicted errors, mean difference .000 [-.002, .002]. Average posterior predicted errors were smaller in the simple 2-estimate task than in the complex 4-estimate task with ellipse stimuli, mean difference .012 [.009, .014], but average errors were similar in the simple 2-estimate and complex 4-estimate tasks with scatter stimuli, mean difference .001, [-.002, .003]. Within a given complexity block and stimulus type, differences due to practice condition included zero in their 95% credible intervals with the exception that in the complex 4-estimate task with scatter stimuli, those who practiced with ellipses had smaller errors than those who did not practice, mean difference 0.003 [0.000, 0.007]. Consistent with the very high overall accuracy on the simple 2-estimate task with ellipse stimuli described in the previous section, error distances were also very low for that combination of complexity and stimulus, and this was most extreme in the participants who did not practice. This accounts for the mismatch between observed mean and posterior predicted mean in that case, because the priors on the model represent the expectation that such small errors were unlikely.

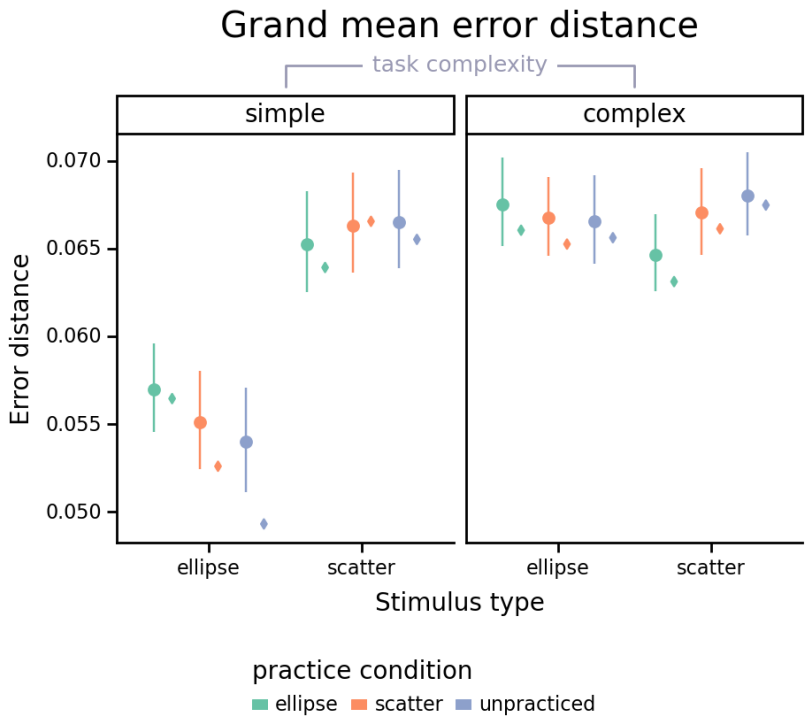

Figure 6. Posterior predicted error distance (lower is better) averaged over time is indicated by circles with a 95% credible interval. Diamonds show experimental sample mean.

The time series of posterior predictive error distance appears in Figure 7. All differences due to practice condition within a given task complexity and stimulus type were small and the 95% credible interval included zero at both the start of the block and at the end of the block.

The model of error distance was a mixture model (7) including responses drawn from a normal distribution centered on the correct answer for each trial (to represent legitimate attempts to do the task) and responses selected uniformly from the entire response space (to represent accidental, haphazard, or very inaccurate clicks). As such, the parameters of the model were the standard deviation of the normal error probability distribution, σ, and the probability of a response coming from the uniform distribution, u (Figure 8). In the early parts of both the simple 2-estimate and the complex 4-estimate blocks, there was a higher probability of a uniform response that decayed quickly. This suggests that in the initial trials of both the simple 2-estimate and the complex 4-estimate blocks, participants were relatively more likely to click far from the correct response. This is consistent with a similar profile in the accuracy score model showing an initially high risk of an accuracy score of zero.

Initial values of the standard deviation of the error probability distribution, σ, and the probability of a response coming from the uniform distribution, u, were similar across the different practice conditions, with the 95% credible interval containing zero for all practice contrasts. The same was true at the end of the block, with the exception that the error standard deviation, σ, was larger in the simple 2-estimate task with ellipse stimuli for those who practiced with ellipses

compared to those who practiced with scatter (mean difference 0.003 [0.001, 0.006]) and compared to those who did not practice (mean difference 0.005 [0.002, 0.007]).

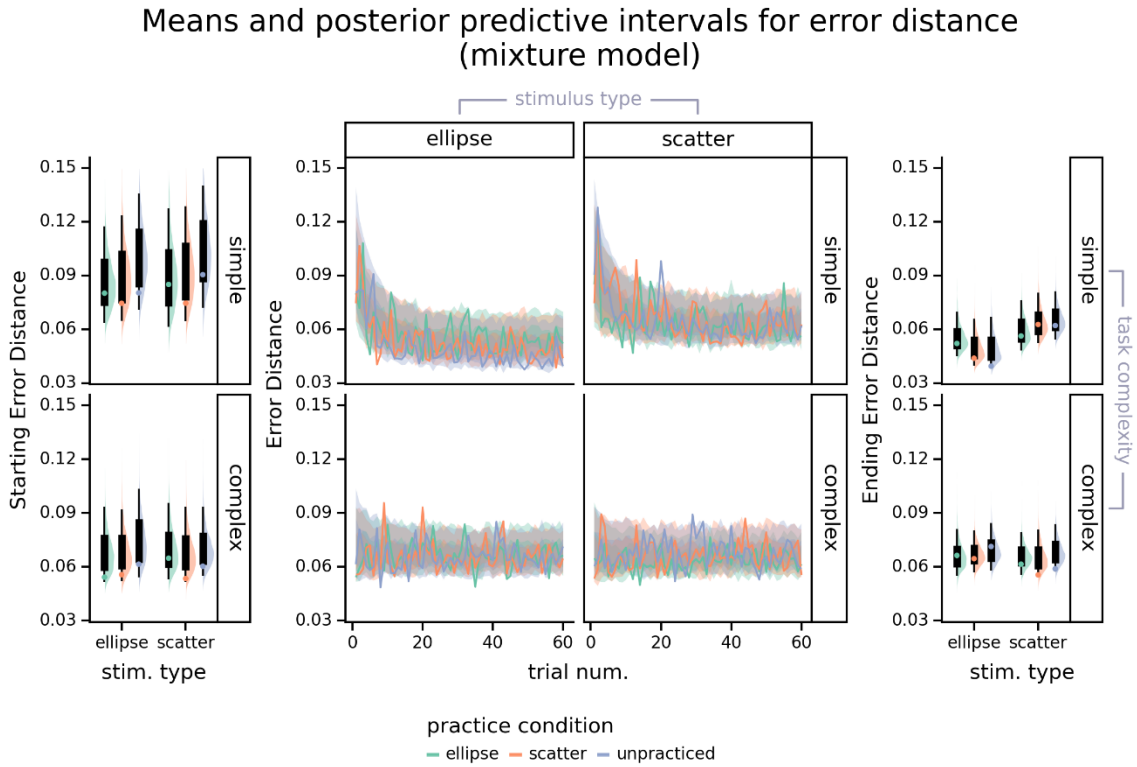

Figure 7. Posterior predictive time-series of error distance, a measure of inaccuracy. Distance units are relative to the width of the virtual environment. Posterior summaries include the central 66% and 95% credible intervals of the posterior predictive distribution of starting error distance (error distance on the first trial of the block, left) and ending error distance (error distance on the last trial of the block, right). Dots show the experimental sample mean. The central panel shows how the mean error distance changed over time, with light ribbons covering a 95% credible interval and darker ribbons covering a 66% credible interval. Lines show the experimental sample mean.

As in the accuracy score model, the random effects of participant in the simple 2-estimate and the complex 4-estimate blocks were highly correlated: for the normal error standard deviation, σ, the correlation was r = .746 95% credible interval [.688, .799], for the probability of uniform response, u, it was r = .460 [.080, .854]. As in the accuracy score model, these correlations suggest that individual differences, rather than randomness, might account for differences in performance between individuals.

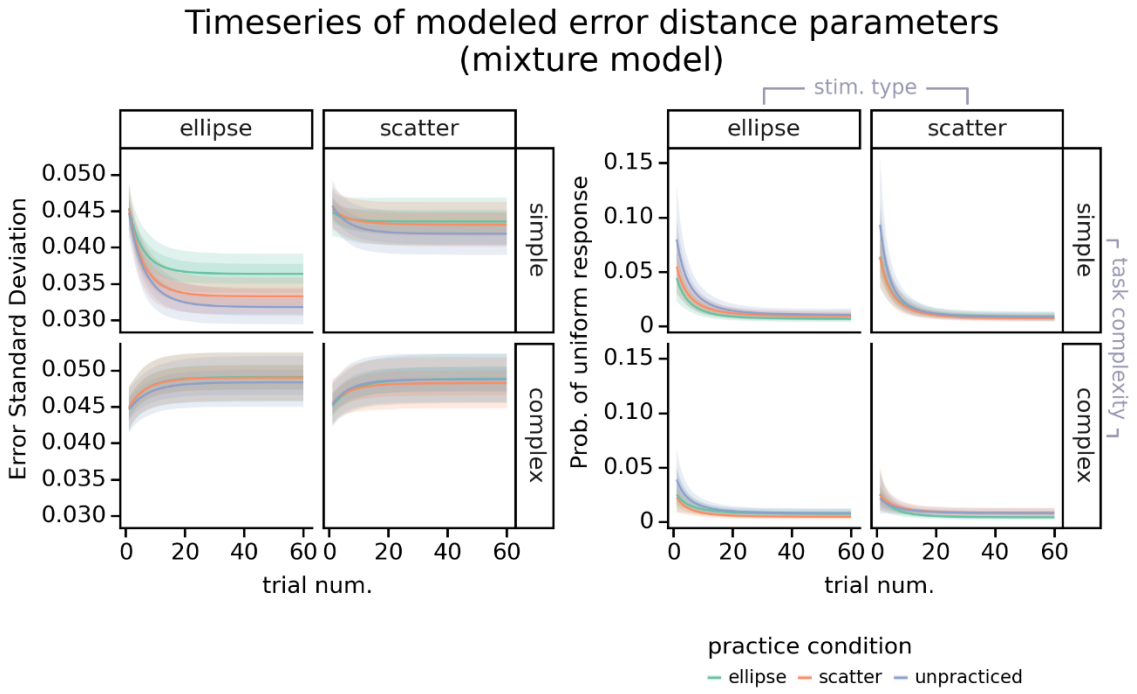

Figure 8. Posterior timeseries of the response error mixture model by stimulus complexity (rows) and stimulus type (columns). The parameters of this model are the standard deviation of the normal error probability distribution σ, modeling participants' typical accuracy in terms of error distance (left) and the probability of a response coming from the uniform distribution instead of the normal probability distribution u, modeling the probability of an extremely erroneous response (right). Time series show mean, central 66% (dark ribbons) and 95% (light ribbons) credible intervals. Trial time is the proportion of the 60-trial block elapsed.

3.3 Response Time

In contrast to the models of response accuracy, the model of response time showed a dramatic difference between those who practiced and those who did not (Figure 9). Averaging over the course of blocks, in the simple 2-estimate task with ellipse stimuli, unpracticed participants responded slower by 1.78s [1.66s 1.90s] and 1.80s [1.68s, 1.92s] than those who practiced with scatter or ellipse stimuli, respectively. In the simple 2-estimate task with scatter stimuli, unpracticed participants responded slower by 2.04s [1.91s, 2.16s] and 2.42s [2.30s, 2.55s] compared to those who practiced with scatter and ellipse stimuli, respectively. In the complex 4-estimate task with ellipse stimuli, unpracticed participants responded slower by 1.37s [1.25s, 1.48s] and by 1.29s [1.17s, 1.40s] compared to those who practiced with scatter or ellipse stimuli, respectively. In the complex 4-estimate task with scatter stimuli, unpracticed participants responded slower by 1.00s [0.88s, 1.11s] and by 1.53s [1.43s, 1.64s] compared to those who practiced with scatter or ellipse stimuli, respectively. Less dramatically, with scatter stimuli, participants who practiced with scatter stimuli responded slower by 0.39s [0.28, 0.49s] and 0.54s, [0.44s, 0.63s] compared to those who practiced with ellipse stimuli in the simple 2-estimate task and the complex 4-estimate task, respectively.

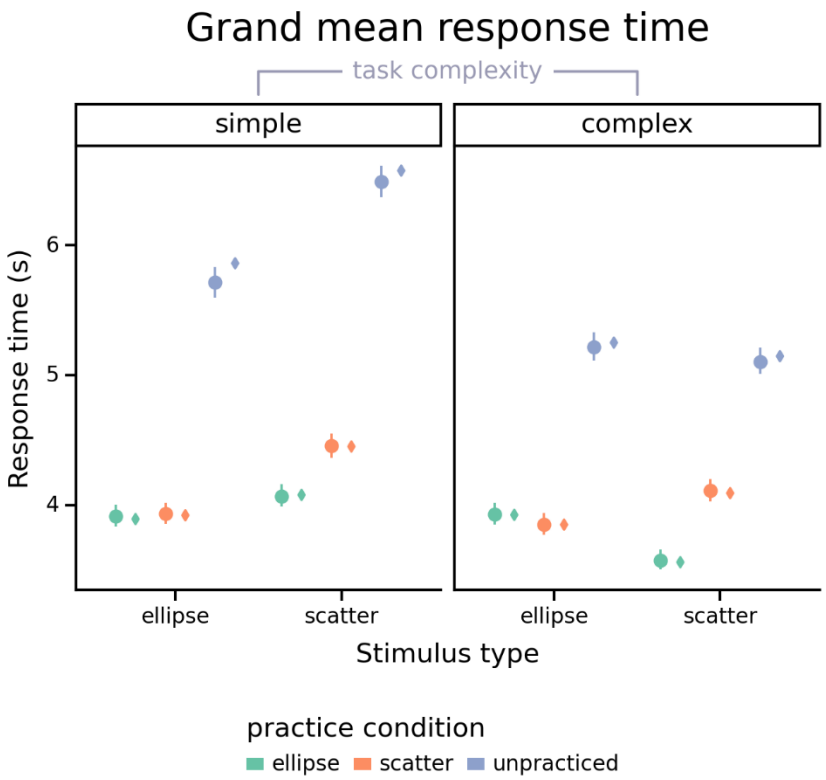

Figure 9. Posterior predicted response time (lower is faster) averaged over time is indicated by circles with a 95% credible interval. Diamonds show the experimental sample mean.

The time course of RT means and posterior predictive intervals are shown in Figure 10. The initial and final RT estimates are also shown in Figure 11 with expanded scales to visualize differences more clearly. There was a mismatch between observed mean response times on the first one or two trials and the model's posterior predictive distribution over those trials in the simple 2-estimate task for some practice conditions. This is likely due to extreme response times on initial trials as participants used the 3D interface for the first time in the main experiment. A more detailed model could account for this brief, initial burn-in, but since posterior predictive fit was good for subsequent trials, we instead focus on trials 3+. Participants who did not practice responded more slowly throughout the experiment than those who did. For example, unpracticed participants' posterior predictive mean initial response time for scatter stimuli in the simple 2-estimate block was 10.14 s [9.26, 11.06] compared to 6.15 s [5.46, 6.91] for those who practiced with scatter stimuli and 6.18 s [5.49, 6.99] for those who practiced with ellipse stimuli. The type of stimuli used in practice did not affect response times very much, with the exception at the end of the simple 2-estimate task block with scatter stimuli: participants who practiced with scatter stimuli responded 0.50s [0.01s, 0.99s] slower than those who practiced with ellipse stimuli. Response times in the complex 4-estimate block were similar to corresponding response times in the simple 2-estimate block, with the exception that unpracticed participants using scatter stimuli responded slower by 0.84s [0.23s, 1.46s] at the end of the simple 2-estimate task compared to at the end of the complex 4-estimate task. This suggests that the added complexity did not substantially slow down task performance.

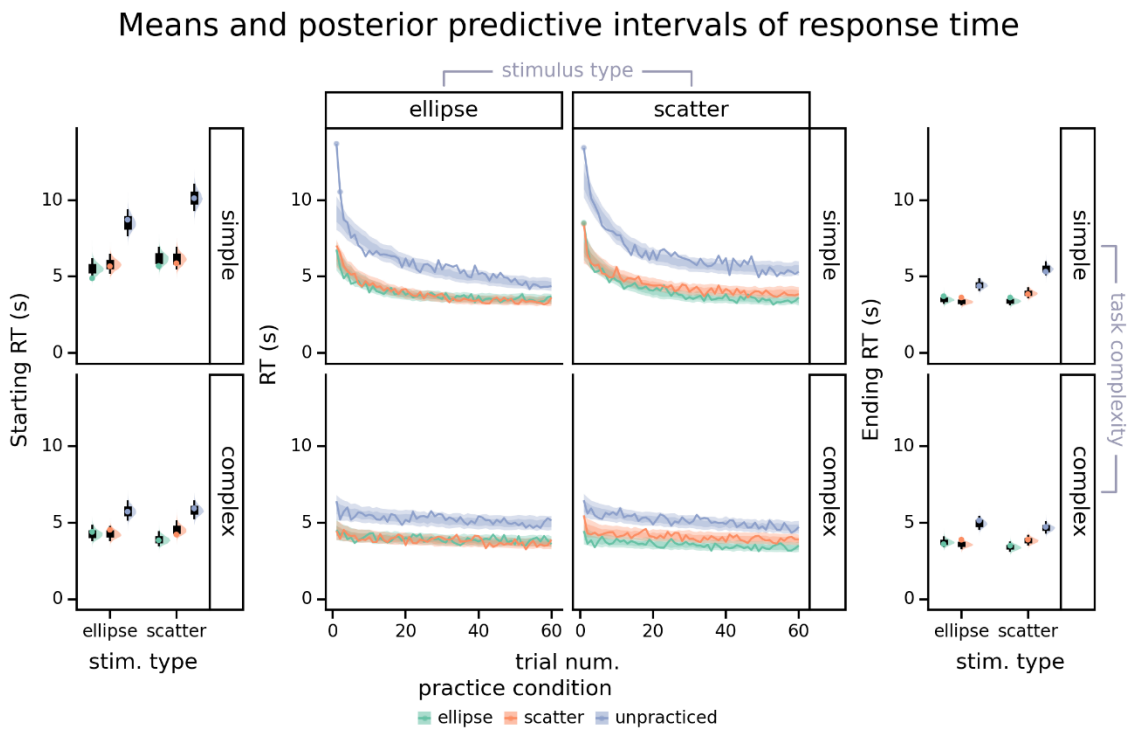

Figure 10. Posterior predictive summaries of response time. Posterior summaries include the central 66% and 95% credible intervals of the starting response time (response time on trial 3 of the block, left) and ending response time (response time on the last trial of the block, right). Dots show the experimental sample mean. Trial 3 is illustrated, because of very long response times on trials 1 and 2 that fell far outside the posterior predictive distribution. The central panel shows how the response time changed over time, with light ribbons covering a 95% credible interval and darker ribbons covering a 66% credible interval. On the central panel, dots indicate cases in which response times fell outside the 95% credible interval on the first two trials. Lines show the experimental sample mean.

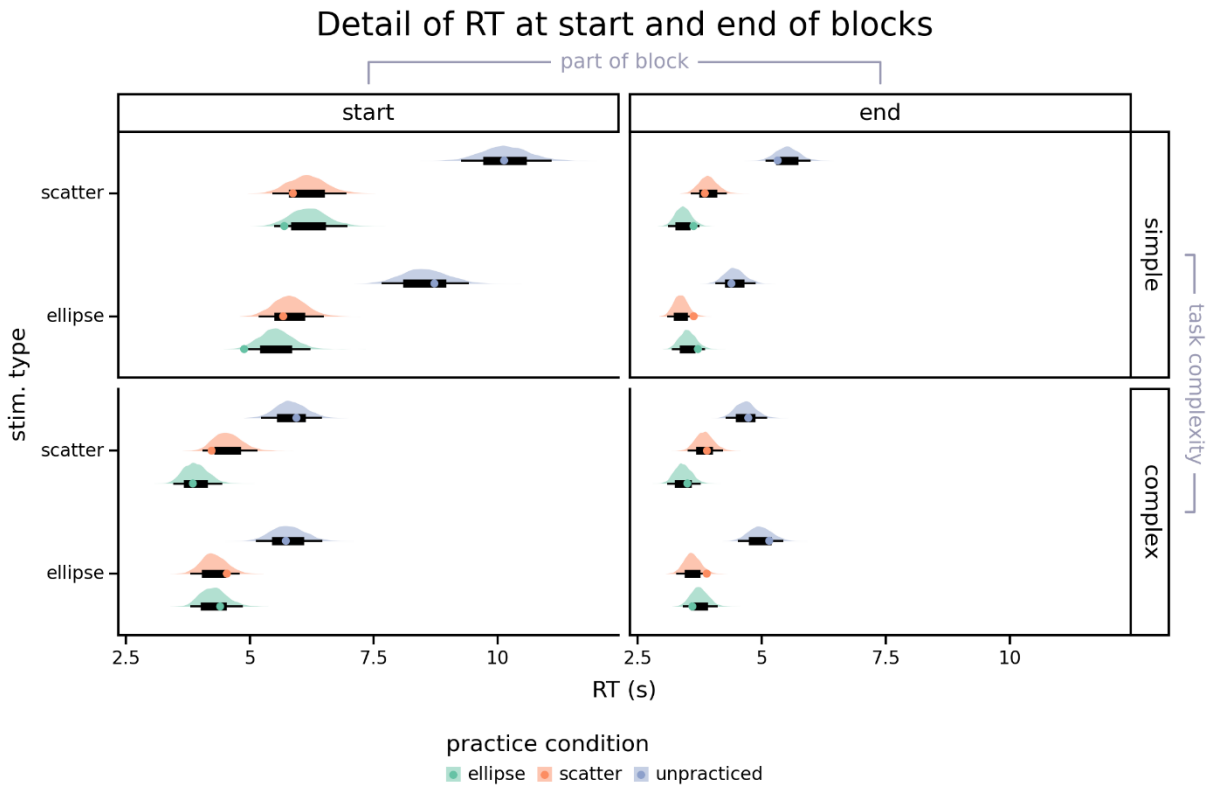

Figure 11. Detail of response time distributions at start (trial 3) and end (trial 60) of blocks. Posterior summaries include the central 66% and 95% credible intervals of the response time on trial 3 (left) and final response time (right). Dots show the experimental sample mean.

The parameters of the response time model were the mean, μ, and standard deviation, σ, of the normal distribution over the natural logarithm of response time (Figure 12). Both parameters showed a separation of unpracticed participants from those who practiced with either ellipse or scatter stimuli. In all cases, the mean parameter, μ, decreased more slowly over time, reflecting faster responses, compared to the standard deviation parameter, σ, which decayed relatively faster. This initially high but rapidly dropping standard deviation parameter indicates more unusually slow responses at the beginning of both simple 2-estimate and complex 4-estimate blocks. The slow responses at the beginning of the simple 2-estimate block might be partially explained by general unfamiliarity with the virtual environment and its interface (although all participants experienced the same environment and interface briefly in the qualifying session). However, in the complex 4-estimate block, this likely reflects adaptation to the much more complex task.

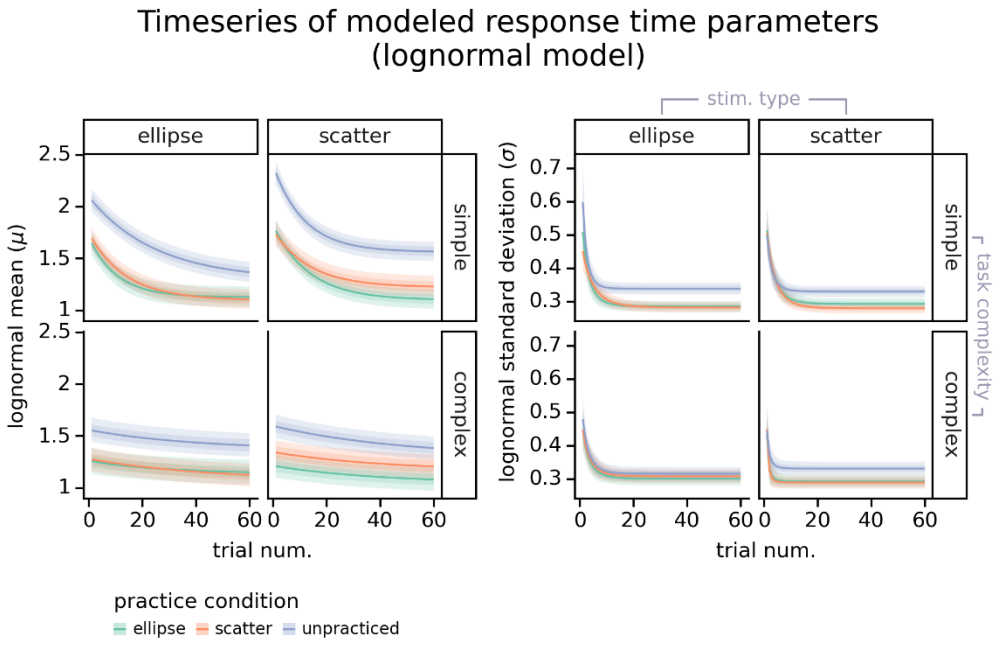

Figure 12. Posterior timeseries of the response time model by stimulus complexity (rows) and stimulus type (columns). The parameters of this model are the mean parameter, μ (left) and standard deviation parameter, σ (right) of the normal distribution of log(RT). Time series show mean, central 66% (dark ribbons) and 95% (light ribbons) credible intervals. Trial time is the proportion of the 60-trial block elapsed.

Random effects of participants for the simple 2-estimate and complex 4-estimate blocks were again highly correlated, especially the mean parameter, μ: $r = 0.850$ [.821, 0.875] and less so the standard deviation parameter, σ: $r = .629$ [.540, .709]. Individual differences are thus likely to influence performance not only in terms of accuracy but also in terms of response time.

3.4 Results Summary

Practice with 2D abstract stimuli led to faster responses in the 3D sand table transfer task, even for a different visualization type (ellipse vs. scatter) and with greater data complexity. However, the practice did not yield improvements in accuracy on the transfer task. Responses were much slower for participants who received no practice compared to participants who practiced in abstract 2D with either scatter or ellipse stimuli. Despite this difference in response times, accuracy score and error distance were both similar across practice conditions, with the exception that accuracy averaged over the simple 2-estimate task block with ellipse stimuli was higher for unpracticed participants than those who practiced, although this difference was small and only detected when averaging over the entire task block.

Participants were less accurate when responding in the complex 4-estimate condition compared to the simple 2-estimate condition. The comparison between performance with simple 2-estimate and with complex 4-estimate stimuli reveals a drop in accuracy score from simple to complex. The same comparison for error distance showed an increase in error distance associated with ellipse stimuli, but for scatter stimuli error distance was similar between the simple 2-estimate

and complex 4-estimate task blocks. Response times were similar or faster in the complex 4-estimate task block compared to the simple 2-estimate task block.

Participants generally performed better, or similarly, with the 3D ellipse visualization than with the 3D scatter visualization. Within the simple 2-estimate task block, responses had higher accuracy scores for ellipse stimuli than scatter stimuli, although the error distances were more similar. Response times for ellipse stimuli in the simple 2-estimate task block were faster than response times for scatter stimuli in the simple 2-estimate task block except for the group who practiced with ellipse stimuli.

4 DISCUSSION

The main finding of this experiment is that practice with an abstract, 2D uncertainty practice task results in substantially faster responses on a transfer task situated within a 3D virtual environment, and this increase in speed is not associated with a loss in accuracy. This speed increase was demonstrated in a task that requires participants to combine multiple, uncertain estimates visualized with either contour ellipses or scatter plots to identify a most likely spatial location. This result is good news, as it suggests that practice in an abstract task can produce speed improvements that transfer to a new context and platform, as well as to different visualizations and across difficulty levels. In what follows, we discuss additional details of the results as well as their implications for designs of visualizations and other tools to help people to make better, faster probabilistic decisions.

4.1 Practice improved speed across visualization types

Our results replicate, in a new 3D context and platform, that practice with one visualization type (ellipse or scatter) improves speed of performance with the other type of visualization, without sacrificing accuracy. This supports H1 and is similar to what was observed in Kusumastuti et al. (2022), who found that practice in abstract 2D with one visualization type led to improvements in speed as well as accuracy when tested on different visualization types. Kusumastuti et al. (2022) used the same platform and context for practice and for testing, and they tested participants with fairly low data complexity (two or three independent data sources visualized at a time). This similarity of performance suggests that learning with one visualization type transfers to performance with the other. A theoretical account of reasoning with visualized uncertainty grounded in cognitive science (Padilla et al., 2018) describes the process of converting a visual array to a conceptual message and then using that conceptual message to answer questions and make decisions. Practice with a given visualization seems unlikely to improve the process of converting an entirely different visual array to a conceptual message, because those processes are specific to the visualization type. However, it is consistent with the Padilla et al. account that practice could improve the reasoning processes that are common to both the practiced and un-practiced visualization, i.e., the underlying uncertainty integration and probabilistic inference task.

Although practice with one stimulus type had similar effects on performance using the other, there were some subtle differences. For simple 2-estimate ellipse stimuli in the 3D task, practice with ellipses in the 2D practice task was associated with lower accuracy and lower precision (i.e., lower μ and κ estimates, the accuracy score beta distribution mean and sample size, respectively) as well as higher error standard deviation estimates. Along similar lines, for both simple 2-estimate and complex 4-estimate 3D scatter stimuli, those who practiced with scatter stimuli responded more slowly than those who practiced with ellipses. Together, these findings suggest some minor advantages to cross-practice, although it is not clear why that might be the case.

4.2 Practice in 2D improved speed in 3D

It is possible that speed improvements observed by Kusumastuti et al. (2022) may have arisen due to participants becoming more comfortable with using the particular display platform and context, namely, a static 2D display of abstract, context-free data. Our current study found that participants who practiced in abstract 2D transferred their speed gains to a more immersive 3D platform with a more concrete visual context and modest narrative context for the data. This supports H2 and suggests that participants did not simply increase their facility with the display platform or abstract context. However, there are some caveats to consider, given the many differences between 2D and 3D displays.

One explanation for the faster responses of those who practiced is that the practice increased speed and fluency in the reasoning underlying this uncertainty integration task. However, another possible explanation is that experience with the more stringent 2D task timer led to faster responses even when working with the less strict 3D task timer. The 3D task timer was less strict by necessity, as participants would need more time to rotate and navigate the interactive 3D display. The participants who practiced experienced the same 3D task as those who did not, but trials in the abstract task ended after 10 seconds (rather than 20 seconds in the 3D task). Participants might have internalized the 10-second deadline from the practice and tried to stick to it even in the 3D task. Additionally, participants who did the 2D practice task experienced a longer experiment overall, so they might have been in more of a hurry to get done compared to the participants who did not practice. Despite differences in speed, posterior estimates of accuracy (in terms of accuracy score and error distance) were similar or better among the participants who practiced compared to those who did not, with the exception of a small accuracy advantage for unpracticed participants using the ellipse stimuli averaged over the simple 2-estimate task block In relative terms, unpracticed participants were 3.9% more accurate than those who practiced with ellipse stimuli, while the difference in response time was 46.0%. The absence of an accuracy decrement elsewhere and a relatively small accuracy decrement in the simple 2-estimate task with ellipse stimuli casts doubt on explanations that involve participants rushing.

Although a dramatic speed increase with no or very minor accuracy decrement could be considered very advantageous in time-limited environments, it might be surprising that accuracy was not also improved with practice. One possibility is that the longer procedure for participants who practiced led to those participants being fatigued relative to those who did not practice. The average total study duration was 69 minutes for unpracticed participants and 89 minutes for those who practiced. We might imagine that accuracy would have been higher if practiced participants had not experienced an additional 20 minutes of procedure or that unpracticed participants' accuracy would be worse if they had done an extra 20 minutes of procedure that was not practice. Future work could examine this possibility by ensuring equivalent procedure duration by adding some equally fatiguing but unrelated task for the unpracticed group.

Prior studies that have compared performance in 2D vs 3D displays found that response times are slower in a 3D display as compared to a 2D display, and attributed this to 3D displays having visual clutter that distracts people from making inferences from helpful cues that may have been more salient in a 2D display (Liao, Dong, Peng, and Liu, 2016; Hegarty et al., 2009). We did not compare response times between the 2D and 3D tasks in this study, since differences between the tasks would make such a comparison difficult to interpret, although one possibility is that practice in the 2D task enabled people to ignore the task-irrelevant details in the 3D task, leading to faster performance compared to the unpracticed participants. Additionally, there was a substantial difference in visual clutter between the 3D ellipse stimuli and the 3D scatter stimuli. This could potentially account for the small performance advantage of ellipse stimuli over scatter stimuli observed in the simple 2-estimate task blocks. In complex 4-estimate task blocks, performance using ellipse and scatter stimuli were similar. If clutter accounts for the performance difference in the simple 2-estimate task blocks, perhaps even the ellipse stimuli resulted in too much clutter on the complex 4-estimate blocks.

As we did not vary the 3D environment, it is possible our findings could be somehow specific to the properties of the 3D terrain we used. For example, the mountainous terrain introduced substantial verticality into the display that distorted the contours of the ellipse stimuli, but the terrain was smooth enough that occlusion was not a major concern. A more jagged or more urban terrain would likely introduce challenges to the visualizations we used in this study. Similarly, a more realistic terrain might introduce substantially more clutter and exacerbate any problems clutter introduces. Our 3D environment primarily provided a different task context with different visualization details to assess the effects of practice in 2D, but the extent to which any details of our findings extend to other 3D environments would need to be verified.

Regardless of previous experience in the 2D practice task, both accuracy and speed improved over the course of the 3D task. Accuracy improvements were associated both with more precise and accurate responses as well as with reductions in the frequency of very incorrect and very slow responses. These improvements primarily happened over the first 20 or so trials in the simple 2-estimate block. Practice with feedback enables people to improve their performance, whereas repeating a task without feedback (or any other way to evaluate the degree of success a given approach to the task) makes improvement much less likely (Ericsson et al., 1993). Without feedback, a person cannot usefully try out different strategies or shape their behavior incrementally to approach their goal. Although this finding might not be surprising, it underscores the value in providing immediate feedback in training programs.

4.3 Practice improved speed with more complex data

Improvements in performance may derive from improvements in the underlying probabilistic inference that participants must perform to succeed at the tasks. However, it is also possible that participants simply discovered or settled upon rule-of-thumb heuristics that, while not optimal, may have nonetheless yielded improved performance in simple data contexts. The current study examined this by requiring participants to engage with the visualization and decision-making task using four independent data sources. This more complex data would render many simplistic heuristics ineffective. In support of H3, we found that participants trained on simpler data were able to transfer their speed gains to the more complex data tasks.

4.4 What practice reveals

Some uncertainty visualization research does not include performance-related feedback (for examples, see Nguyen et al., 2020; Padilla, Powell, et al., 2021; Ruginski et al., 2016; Tak et al., 2014, 2015). Understanding how someone without training or previous experience uses a visualization and what errors they might make is critical in communicating uncertainty to the general public. However, this approach obscures any distinction between inherent limitations of the particular visualization itself and limitations due to unfamiliarity or inexperience on the part of the user. When a visualization is intended to be used by someone without specific familiarity or experience, then this distinction is not particularly relevant or important; the goal is to empower that naïve user. On the other hand, when the goal is to produce a tool to help someone who has interest and opportunity to invest in some practice or training, then the distinction between a flawed visualization and a less intuitive one becomes highly relevant.

Confidence intervals have been shown to be ineffective at conveying uncertainty accurately (Belia et al., 2005; Correll & Gleicher, 2014; Joslyn & LeClerc, 2012; Padilla et al., 2017) and similar concerns might apply to the ellipse stimuli in the present study. Moreover, these displays have the drawback of implying an apparently discrete boundary in what is better understood as a continuous space (Padilla, Castro, et al., 2021). The ellipse stimuli in the present study enclose the central 50% and 95% of the underlying probability mass, but the distribution itself is continuous and there is no special distinction between a point just inside a contour and just outside a contour. Simultaneously, the ellipse stimuli also have

the advantage of consisting of salient, continuous contours that can readily be perceived as shapes, and this may have contributed to the slightly better performance observed with ellipses than with scatter stimuli in our study.

The scatter stimuli we used have in common with other ensemble visualizations the benefits of emphasizing that a probability distribution defines the probabilities of different outcomes. Ensemble visualizations have been shown to improve understanding of forecasts as probabilistic (Ruginski et al., 2016) and to avoid deterministic construal errors (Joslyn & LeClerc, 2012). Other ensemble displays include quantile dot plots (Kay et al., 2016) and animated hypothetical outcome plots (Hullman et al., 2015). Ensemble displays might also activate frequency framing, which leads people to better understand probabilities (Gigerenzer, 1996). In addition to these advantages, the ensemble visualization we used might have some drawbacks; the scatter stimuli used many distinct points that were grouped by color, and they were positioned randomly, so perceiving the underlying distinct distributions might have been harder.

Another concern with the scatter stimuli we used is the overlapping of two or four different distributions. This led to regions of the display with a very high density of marks in some stimuli. Several approaches have been developed to mitigate overplotting in a 2D context (Ellis & Dix, 2007). We selected the number of samples per distribution in a compromise between having enough points to characterize the broader distributions while not suffering too much from overplotting with denser distributions. A more sophisticated approach might have led to better performance from participants using scatter stimuli. We might have expected the scatter stimuli here to perform better based on the superiority of ensemble displays in other contexts, and perhaps a better-designed scatter stimulus would have led to better performance. However, the performance advantage of the ellipse stimuli over the scatter stimuli was only apparent in the simple task depicting only two distributions, where presumably overlap would be less of an issue.

In the current experiment, performance with the ellipse stimuli was initially similar to that of the scatter stimuli, but by the end of the simple 2-estimate task block, performance with ellipse was better than with scatter. These differences in performance after practice could be due to the particularities of our implementations of these two visualizations. More research is needed to know if this pattern is generally true for other forms of summary and ensemble displays. Nonetheless, our results suggest that practice might provide an opportunity for an otherwise less intuitive visualization to become more effective.

4.5 Individual differences

In addition to the effects of practice, we observed evidence for stable inter-individual differences in performance. Each model included random effects of participant to characterize how a participant's performance differed from others in the same condition. We used a potentially correlated random effect in the simple 2-estimate task block and the complex 4-estimate task block. We found high correlations between the random effects in simple 2-estimate and complex 4-estimate task blocks. These high correlations indicate that participants' performance relative to those in their condition was stable across these blocks. Understanding measurable correlates of these stable individual differences in probabilistic decision-making (Grounds & Joslyn, 2018) and use of visualized uncertainty (Padilla, Castro, et al., 2021), such as underlying skills or dispositional traits, could help us predict who might perform tasks like this better or worse, and it could help us identify additional practice or other interventions that might help improve performance.

4.6 Instructions vs. practice

The importance of effective instructions and explanations for how to use a visualization has been demonstrated (Boone et al., 2019; Fiore et al., 2019; Song et al., 2019). However, in some studies, instructions increased understanding of a visualization (as confirmed with a comprehension test) but still did not eliminate errors associated with misinterpretation

of that visualization (Boone et al., 2019; Joslyn & LeClerc, 2012). Moreover, trained scientists who receive statistical education are shown to misinterpret uncertainty visualizations that are in common use in scientific publications (Belia et al., 2005). Taken together, these findings support the view that instruction alone is insufficient to enable people to make error-free decisions based on visualized uncertainty.

The present results point to practice with feedback as a key enabler of effective use of visualized uncertainty. In the current experiment, all participants received instructions that explained how to use both the scatter and the ellipse visualization types. Despite not having experience using all visualization types, the instructions seemed to be sufficient to enable use of the unpracticed visualization type. However, we found that the group who practiced with either visualization type showed much faster response times than those who received instructions only.

5 CONCLUSION

Practice is an important element of attaining expertise in a task or field, and this is no less the case for uncertainty integration tasks. For probabilistic reasoning with visualized bivariate Gaussian distributions, we found that practice-induced expertise transfers to different visualization types, to different difficulty levels, and to qualitatively different contexts and platforms. This transfer was evident in much faster response times with no substantial loss of accuracy. Although the particulars of the visualization and of the user interface changed across our experiment, the underlying inference task did not. This suggests that practice, with feedback, improved participants' skill with the underlying probabilistic reasoning involved in these tasks, which were limited to a spatial task using ellipse and scatter visualizations. It remains an empirical question whether benefits of this practice would extend to other tasks of probabilistic reasoning and whether similar practice effects might be observed with other kinds of probabilistic reasoning tasks.

Communicating uncertainty is critical for empowering people to make well-informed decisions. We hope this work inspires designers of tools for conveying uncertainty to consider whether it is possible to provide people an opportunity to practice using that tool before they need to use it to make a decision. The findings we report here suggest that practice could lead people to make faster, better decisions using visualized uncertainty.

ACKNOWLEDGMENTS

This work was supported by DEVCOM Army Research Laboratory's human sciences campaign. The views expressed in this paper are those of the authors and do not reflect the official policy or position of the Department of the Army, Department of Defense, or the U.S. Government. Approved for public release; distribution is unlimited.

REFERENCES

Ashraf, A., Collins, D., Whelan, M., O'Sullivan, R., & Balfe, P. (2015). Three-dimensional (3d) simulation versus two-dimensional (2d) enhances surgical skills acquisition in standardised laparoscopic tasks: A before and after study. *International Journal of Surgery, 14*, 12–16. https://doi.org/10.1016/j.ijsu.2014.12.020

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*(4), 389. http://dx.doi.org/10.1037/1082-989X.10.4.389

Beattie, K. L., Hill, A., Horswill, M. S., Grove, P. M., & Stevenson, A. R. L. (2021). Laparoscopic skills training: The effects of viewing mode (2D vs. 3D) on skill acquisition and transfer. *Surgical Endoscopy, 35*(8), 4332–4344. https://doi.org/10.1007/s00464-020-07923-8

Boone, A. P., Maghen, B., & Hegarty, M. (2019). Instructions matter: Individual differences in navigation strategy and ability. *Memory & Cognition, 47*(7), 1401–1414. http://dx.doi.org/10.3758/s13421-019-00941-5

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The berlin numeracy test. *Judgment & Decision Making, 7*(1), 25–47. https://doi.org/10.1017/s1930297500001819

Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly, 19*(2), 189–211. https://doi.org/10.2307/249688

Correll, M., & Gleicher, M. (2014). Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *IEEE Transactions on Visualization and Computer Graphics, 20*(12), 2142–2151. https://doi.org/10.1109/TVCG.2014.2346298

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving Intelligence Analysis With Decision Science. *Perspectives on Psychological Science, 10*(6), 753–757. https://doi.org/10.1177/1745691615598511

Ellis, G., & Dix, A. (2007). A taxonomy of clutter reduction for information visualisation. *IEEE transactions on visualization and computer graphics*, *13*(6), 1216-1223.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363. http://dx.doi.org/10.1037/0033-295X.100.3.363

Fernandes, M., Walls, L., Munson, S., Hullman, J., & Kay, M. (2018). Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3173574.3173718

Fiore, S. M., Song, J., Newton, O. B., Pittman, C., Warta, S. F., & LaViola, J. J. (2019). Determining the Effect of Training on Uncertainty Visualization Evaluations. In T. Z. Ahram & C. Falcão (Eds.), Advances in Usability, User Experience and Assistive Technology (Vol. 794, pp. 141–152). Springer International Publishing. https://doi.org/10.1007/978-3-319-94947-5_14

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Galesic, M., & Garcia-Retamero, R. (2011). Graph literacy: A cross-cultural comparison. *Medical Decision Making*, *31*(3), 444–457. https://doi.org/10.1177/0272989X10373805

Gigerenzer, G. (1996). The psychology of good judgment: Frequency formats and simple algorithms. *Medical Decision Making, 16*(3), 273–280. http://dx.doi.org/10.1177/0272989X9601600312

Greis, M., Joshi, A., Singer, K., Schmidt, A., & Machulla, T. (2018). Uncertainty Visualization Influences How Humans Aggregate Discrepant Information. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 505:1-505:12. https://doi.org/10.1145/3173574.3174079

Grounds, M. A., & Joslyn, S. L. (2018). Communicating weather forecast uncertainty: Do individual differences matter? *Journal of Experimental Psychology: Applied, 24*, 18–33. https://doi.org/10.1037/xap0000165

Han, P. K. J., Babrow, A., Hillen, M. A., Gulbrandsen, P., Smets, E. M., & Ofstad, E. H. (2019). Uncertainty in health care: Towards a more systematic program of research. *Patient Education and Counseling, 102*(10), 1756–1766. https://doi.org/10.1016/j.pec.2019.06.012

Hegarty, M., Friedman, A., Boone, A. P., & Barrett, T. J. (2016). Where are you? The effect of uncertainty and its visual representation on location judgments in GPS-like displays. *Journal of Experimental Psychology: Applied*, *22*(4), 381. http://dx.doi.org/10.1037/xap0000103

Hegarty, M., Smallman, H. S., Stull, A. T., & Canham, M. S. (2009). Naïve cartography: How intuitions about display configuration can hurt performance. *Cartographica: The International Journal for Geographic Information and Geovisualization, 44*(3), 171–186. https://doi.org/10.3138/carto.44.3.171

Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science, 12*(6), 973–986. http://dx.doi.org/10.1177/1745691617702496

Higgins, E. T., Friedman, R. S., Harlow, R. E., Idson, L. C., Ayduk, O. N., & Taylor, A. (2001). Achievement orientations from subjective histories of success: Promotion pride versus prevention pride. *European Journal of Social Psychology*, *31*(1), 3–23.

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research, 15*(1), 1593–1623. http://jmlr.org/papers/v15/hoffman14a.html

Hullman, J. (2020). Why Authors Don't Visualize Uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 130–139. https://doi.org/10.1109/TVCG.2019.2934287

Hullman, J., Resnick, P., & Adar, E. (2015). Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. *PLOS ONE, 10*(11), e0142444. https://doi.org/10.1371/journal.pone.0142444

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). Guilford Press.

Joslyn, S. L., & LeClerc, J. E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied*, *18*(1), 126. http://dx.doi.org/10.1037/a0025185

Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016). When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. *Proceedings of the 2016 Chi Conference on Human Factors in Computing Systems*, 5092–5103. http://dx.doi.org/10.1145/2858036.2858558

Kusumastuti, S. A., Pollard, K. A., Oiknine, A. H., Dalangin, B., Raber, T. R., & Files, B. T. (2022). Practice improves performance of a 2D uncertainty integration task within and across visualizations. *IEEE Transactions on Visualization and Computer Graphics*. http://dx.doi.org/10.1109/TVCG.2022.3173889

Lejarraga, T., & Hertwig, R. (2021). How experimental methods shaped views on human competence and rationality. *Psychological Bulletin*, *147*(6), 535. http://dx.doi.org/10.1037/bul0000324

Liao, H., Dong, W., Peng, C., & Liu, H. (2017). Exploring differences of visual attention in pedestrian navigation when using 2D maps and 3D geo-browsers. *Cartography and Geographic Information Science*, *44*(6), 474–490. https://doi.org/10.1080/15230406.2016.117488

Metcalfe, J. S., Gordon, S. M., Passaro, A. D., Kellihan, B., & Oie, K. S. (2015). Towards a translational method for studying the influence of motivational and affective variables on performance during human-computer interactions. *International Conference on Augmented Cognition*, 63–72.

Nguyen, F., Qiao, X., Heer, J., & Hullman, J. (2020). Exploring the effects of aggregation choices on untrained visualization users' generalizations from data. *Computer Graphics Forum*, *39*(6), 33–48. http://dx.doi.org/10.1111/cgf.13902

Padilla, L. M., Castro, S. C., & Hosseinpour, H. (2021). A review of uncertainty visualization errors: Working memory as an explanatory theory. *Psychology of Learning and Motivation*, 74, 275–315. http://dx.doi.org/10.1016/bs.plm.2021.03.001

Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: A cognitive framework across disciplines. *Cognitive Research: Principles and Implications*, *3*(1), 29. https://doi.org/10.1186/s41235-018-0120-9

Padilla, L. M., Dryhurst, S., Hosseinpour, H., & Kruczkiewicz, A. (2021). Multiple Hazard Uncertainty Visualization Challenges and Paths Forward. *Frontiers in Psychology*, 12. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.579207

Padilla, L. M., Kay, M., & Hullman, J. (2020). Uncertainty Visualization. In Handbook of Computational Statistics and Data Science. *PsyArXiv*. https://doi.org/10.31234/osf.io/ebd6r

Padilla, L. M., Powell, M., Kay, M., & Hullman, J. (2021). Uncertain About Uncertainty: How Qualitative Expressions of Forecaster Confidence Impact Decision-Making With Uncertainty Visualizations. *Frontiers in Psychology*, 11. https://doi.org/10.3389/fpsyg.2020.579267

Padilla, L. M., Ruginski, I. T., & Creem-Regehr, S. H. (2017). Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive Research: Principles and Implications*, *2*(1), 40. https://doi.org/10.1186/s41235-017-0076-1

Pollard K., Siriwardena P.M., Krum D.M., Files B.T. (2022). Volumetric hazard visualization and navigation in simulated augmented reality. Technical report ARL-TR-9572. DEVCOM Army Research Laboratory

Ruginski, I. T., Boone, A. P., Padilla, L. M., Liu, L., Heydari, N., Kramer, H. S., Hegarty, M., Thompson, W. B., House, D. H., & Creem-Regehr, S. H. (2016). Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, *16*(2), 154–172. https://doi.org/10.1080/13875868.2015.1137577

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, 43(3), 450–461. https://doi.org/10.1037/0022-3514.43.3.450

Smallman, H. S., & Cook, M. B. (2011). Naïve realism: Folk fallacies in the design and use of visual displays. *Topics in Cognitive Science*, *3*(3), 579–608. https://doi.org/10.1111/j.1756-8765.2010.01114.x

Song, J., Newton, O. B., Fiore, S. M., Pittman, C., & LaViola, J. J. (2019). Examining Training Comprehension and External Cognition in Evaluations of Uncertainty Visualizations to Support Decision Making. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 1654–1658. https://doi.org/10.1177/1071181319631520

Stan Development Team. (2022). *Stan Modeling Language Users Guide and Reference Manual* (2.29.2) [Computer software]. http://mc-stan.org

Tak, S., Toet, A., & van Erp, J. (2014). The Perception of Visual Uncertainty Representation by Non-Experts. *IEEE Transactions on Visualization and Computer Graphics*, *20*(6), 935–943. https://doi.org/10.1109/TVCG.2013.247

Tak, S., Toet, A., & van Erp, J. (2015). Public Understanding of Visual Representations of Uncertainty in Temperature Forecasts. *Journal of Cognitive Engineering and Decision Making*, *9*(3), 241–262. https://doi.org/10.1177/1555343415591275

6 RESEARCH MATERIAL STATEMENT

Stimuli, stimulus software, anonymized data, and analysis scripts and related code are available online at https://osf.io/5xdsg/

7 AUTHORSHIP

Benjamin T. Files: Conceptualization, Methodology, Software, Formal analysis, Writing – Original Draft, Visualization, Supervision. **Ashley H. Rabin:** Conceptualization, Validation, Investigation, Data Curation, Writing – Review & Editing, Project administration. **Tiffany Raber:** Conceptualization, Resources, Software, Writing – Review & Editing. **Bianca Dalangin:** Conceptualization, Validation, Investigation, Data Curation, Writing – Review & Editing. **Kimberly A. Pollard:** Conceptualization, Methodology, Writing – Review & Editing, Supervision.

8 LICENSE

9 CONFLICTS OF INTEREST

The authors declare that there are no competing interests.