Cue integration in spatial search for jointly learned landmarks but not for separately learned landmarks

Yu Du, Neil McMillan, Christopher R. Madan, Marcia L. Spetch, & Weimin Mou

Address correspondence to:

Marcia L. Spetch

Department of Psychology

University of Alberta

P-217, Biological Science Building

Edmonton, Alberta

Canada T6G 2E9

Email: mspetch@ualberta.ca

Telephone: (+1)780-4927548

Abstract

We investigated how humans use multiple landmarks to locate a goal. Participants searched for a hidden goal location along a line between two distinct landmarks on a computer screen. On baseline trials, the location of the landmarks and goal varied, but the distance between each of the landmarks and the goal was held constant, with one landmark always closer to the goal. In Experiment 1, some baseline trials provided both landmarks, and some provided only one landmark. On probe trials, both landmarks were shifted apart relative to the previously-learned goal location. Participants searched between the locations specified by the two landmarks and their search locations were shifted more toward the nearer landmark, suggesting a weighted integration of the conflicting landmarks. Moreover, the observed variance in search responses when both cues were presented in their normal locations was reduced compared to the variance on tests with single landmarks. However, the variance reduction and the weightings of the landmarks did not always show Bayesian optimality. In Experiment 2, some participants were trained only with each of the single landmarks. On subsequent tests with the two cues in conflict, searching did not shift toward the nearer landmark and the variance of search responses of these single-cue trained participants was larger than their variance on single-landmark tests, and even larger than the variance predicted by using the two landmarks alternatively on different trials. Taken together, these results indicate that cue combination occurs only when the landmarks are presented together during the initial learning experience.

Keywords: spatial memory; cue integration; cue conflict; landmark; Bayesian integration

The ability to localize goals is essential for humans, as it is for other animals. While searching for something from memory, information from multiple cues in the environment may be used to improve search precision. For example, if we have to find our way to work from a different location than we normally start from (e.g., after staying over at a friend’s house), we may attempt to search for and combine in memory the nearby roads, buildings, and other landmarks that allow navigation to that particular location; likewise, we sometimes may need to adjust which landmarks we rely on, if for example there has been road construction, if multiple buildings look similar, or if some landmarks are too far away to be useful for pinpointing location. To what extent do people actually integrate, or average, spatial information about landmarks to estimate the location of a goal? And does combination of landmark information always occur, or must training include instances in which multiple cues are present? If combination occurs, does the weighting given to different cues depends on their reliability? If so, is the combination optimal as indicated by reduced variance (e.g., Cheng, Shettleworth, Huttenlocher, & Rieser, 2007)? In the current research we addressed these questions in two experiments using a simple spatial task.

Combination of information from multiple cues could take different forms. One possibility is that the information might be fully integrated into a single unified representation that includes the spatial information between each cue and the goal as well as the spatial information between the cues themselves (e.g., Ishikawa & Montello, 2006; Mou & Spetch, 2013; Pantelides, Kelly, & Avraamides, 2016; Yamamoto & Shelton, 2008). A second possibility is that the information from each cue might be represented separately, but the information is averaged in some way at the time of searching such that the estimated goal location is jointly determined by each cue. In either case, to the extent that each cue adds useful information (e.g., as in the multiple bearing hypothesis of Kamil & Cheng, 2001), then searching should be more accurate when multiple cues are present than when only a single cue is present. Moreover, if the cues provide conflicting information, then the search location may show a compromise between the locations specified by each cue. For the purpose of the present research, we refer to both of these possible ways of combining information as the *Integration Model.*

On the other hand, information from multiple cues could instead be used separately rather than being combined to find the goal. There are at least two models of multiple cue use that assume no integration of information. First, the *Hierarchical Model* assumes that the most reliable or salient source of information will be used when it is present; other cues will be used only when the preferred one is not available. This model is also known as “take the best” or “winner-takes-all” (Gigerenzer & Brighton, 2009; Lea et al., 2009; Legge, Madan, Spetch, & Ludvig, 2016; Legge, Spetch, & Batty, 2009; Spetch & Edwards, 1988). Second, the *Alternation Model* suggests that multiple cues may be used across a number of search attempts but the information will not be averaged into a single estimate of the goal location. For example, if conflict exists, the person may first try the location specified by the preferred cue and then try the location specified by the non-preferred cue.

Many studies have shown that humans can combine information from different types of cues (for a review, see Cheng, Shettleworth, Huttenlocher, & Rieser, 2007), such as path integration and landmarks (Chen & McNamara, 2014; Nardini, Jones, Bedford, & Braddick, 2008; Sjolund, 2014; Zhao & Warren, 2015b), directional information from beacons and dead reckoning (Bodily, Daniel, & Sturz, 2012), egocentric and allocentric cues (Byrne & Crawford, 2010), boundaries and landmarks (Doeller & Burgess, 2008), or spatial categories (e.g. left and right halves in a rectangle) and fine-grained spatial information (Huttenlocher, Hedges, Corrigan,

& Crawford, 2004; Huttenlocher, Hedges, & Vevea, 2000; Sampaio & Wang, 2009; Sandberg, Huttenlocher, & Newcombe, 1996). Frequently, however, multiple cues of the same type could be used to localize a goal, such as multiple objects that could each serve as discrete visual landmarks. Some studies have suggested that non-human animals can combine information from multiple landmarks (Legge et al., 2016), but the manner and extent to which humans combine information from multiple discrete landmarks to locate a goal is less clear.

Since cue combination has been observed with other types of spatial cues, it would be reasonable to speculate that people will show combination of information from multiple landmarks. However, very few studies have directly tested this, and the evidence to date does not provide a clear answer. There is some evidence that humans, like other animals (Legge et al., 2016; Spetch & Mondloch, 1993), will differentially weight multiple landmarks in spatial tasks (Byrne & Crawford, 2010; Sturz & Bodily, 2010; see Cheng et al., 2007 for a review). However, these studies did not provide direct tests of whether the information from the different landmarks was combined or used independently in determining where to search. Some evidence that humans integrate information from multiple landmarks was also provided by a recent detection task study (Mou & Spetch, 2013) in which participants viewed a layout of objects in a virtual reality environment and then were asked to detect whether a target object had moved. In this case, the target objects can be seen as the goal and other objects can be seen as landmarks. The test results suggested that multiple inter-object vectors were encoded in an integrative fashion. In particular, when four landmark objects were present during encoding, detection accuracy with all four landmarks present at test was better than the optimal sum of the performance with the two closest landmarks and performance with the two farthest landmarks. However, this study did not assess searching behavior with continuous measures (e.g., distance error to the original target location) and therefore did not provide direct evidence about whether humans combine information from multiple landmarks in determining where to search. This study also did not directly examine the combination of single landmarks because only either of the two nearest or two farthest landmarks were presented together during encoding.

On the other hand, some findings suggest that humans may not combine the information from multiple landmarks. Baguley, Lansdale, Lines, and Parkin (2006) used two single landmarks to investigate the combination of spatial memories. Participants first learned a target's position along a horizontal line with two individually-presented landmarks on a computer screen and later were asked to locate the target when both landmarks were presented at the same time. Surprisingly, participants did not show higher search accuracy when both landmarks were presented than when only a single landmark was presented, suggesting that they did not combine the information from the two landmarks in determining where to search. Even when the participants learned two landmarks together during training, their estimation of the target's position when both cues were shown was similar to their estimations when each cue was tested alone. A subsequent study with a similar paradigm reported similar findings (Clark, Dunn, & Baguley, 2013).

Thus, the evidence to date does not provide a clear answer to the question of whether humans combine information from multiple landmarks to locate a goal. It is possible that combination is most likely to occur when spatial cues are processed by different systems (e.g., the path integration system and visual landmarks, Nardini et al., 2008; self-to-object and object-to-object systems, Mou & Spetch, 2013) or are not perceptually comparable (e.g., boundary and landmarks, Doeller & Burgess, 2008). When cues come from different systems for spatial localization, combination of information may be common because there may be less competition or interference between the cues. The question of whether, and under what conditions, people combine information from cues that come from the same system and are of comparable perceptual salience remains to be answered.

Based on previous studies, we speculated that if humans combine information from multiple landmarks, the information from each landmark would be weighted based on the landmark's "certainty" in specifying the goal location. Ruprecht, Wolf, Quintana and Leising (2014) have suggested that spatial accuracy is a function of a landmark's proximity to the goal, its stability (in terms of the variance of the landmark-goal vector across trials), as well as the reward probability signalled by the landmark. In other words, proximity, stability and reward probability can all be seen as determinants of the certainty of the landmarks with respect to finding the goal. For example, Zhao and Warren (2015a) found that the stability of landmarks can largely influence participants' weighting of the landmarks and path integration in a navigation task. The reliability of a cue can be mathematically defined in terms of the inverse of the response variance (Chen & McNamara, 2014; Cheng et al., 2007; Nardini et al., 2008; Zhao & Warren, 2015a); a more reliable cue results in less variant responses and greater certainty about the goal. A more proximal landmark to the goal should provide greater certainty about the goal's location and should result in a smaller variance in response locations when this landmark is available alone compared to a landmark that is farther from the goal. When multiple landmarks are present, information from the more proximal landmark should therefore be weighted more heavily.

Our research addressed two related questions about the goal localization process in humans: (1) whether humans combine information from multiple landmarks, and (2) under what conditions and in what manner the information is combined. Our study used a single spatial dimension and manipulated the proximity of two landmarks to the goal location. We positioned the two landmarks and a goal on a line similar to that of Baguley et al. (2006) and Clark et al. (2013). One of the two landmarks was always closer to the goal. We assume that the closer landmark will result in lower response variance, and hence we refer to the closer landmark as the reliable cue (*Reliable Cue*, R) and the farther landmark as the unreliable cue (*Unreliable Cue*, U). The visual salience and presentation frequency of the landmarks were controlled such that distance to the goal was the only difference in the landmark's certainty.

Our procedure differed from two previous studies (Baguley et al., 2006; Clark et al., 2013), which found that people did not combine information from multiple landmarks even when the cues were learned together. First, those studies used multiple sets of landmarks and target locations across trials which required participants to remember considerably more information than in the current study. Second, in those two studies, the landmarks were always presented at the left and right edges of the screen (respectively) during training, and each pair of landmarks was associated with one specific target. Therefore it is possible that the target was always encoded with respect to the edges of the screen and the landmarks were used only as contextual cues to specify the identity of the target. We used only one pair of landmarks and one goal, which reduced the cognitive demands of the task, and we varied the absolute location of our stimuli (landmarks, goal) on the computer screen while keeping the relative distance between them constant (see Method).

In our studies the landmarks were far enough from the goal that they could not serve as beacons (i.e. beacons are cues located near enough to the goal that distance and direction from the cues does not need to be encoded). Moreover, in contrast to some other studies that used multiple cues (Ruprecht et al., 2014; Sturz & Bodily, 2010), we used a continuous search space (along a single dimension) rather than discrete response locations; this provided more spatial response resolution so that we could detect evidence of compromise during conflict situations.

In the current studies, participants learned a goal location relative to two landmarks (R and U) along a line. In the three main comparison conditions of testing, the participants were then asked to search for the goal with: (1) one of the landmarks, (2) two landmarks that were the same distance apart as in training, or (3) two landmarks that were shifted farther apart than in training—thus providing conflicting information. To investigate whether presenting the compound cues in learning influenced cue combination, we provided different types of training trials to participants. In Experiment 1, training provided both single cue trials as well as trials in which both cues were presented together. In Experiment 2, some participants were trained with single cues only, to investigate whether participants would combine information when the two cues were learned separately, compared with groups trained with both cues simultaneously.

Two types of evidence may be used to show that cues are combined for spatial localization. First, if the cues are combined, variance should be reduced when the cues are presented together compared to when they are presented singly. If the combination is optimal, the variance would reduce to the Bayesian prediction in which the weighting given to different cues depends on their reliability. Second, when cue locations are shifted so that the cues are in conflict, information from more than one cue is used to determine where to search and the search location will therefore show a compromise between the locations specified by each cue individually. We used a model comparison approach to test these predictions.

Model Predictions

By examining conflict trials (i.e. when the cues were shifted), we investigated whether people combined the information provided by the two cues or used the information independently. On shift trials, the reliable landmark (R) indicated one goal location (g_R) and the unreliable landmark (U) indicated a different goal location (g_U).

For modeling, we followed the Bayesian formulas specified by Cheng et al. (2007) and Nardini et al. (2008). The degree to which participants relied on each landmark is given by the relative proximity (*rp*) of their mean search location to these different goal locations (Doeller, King, & Burgess, 2008; Nardini et al., 2008). In the current study, this is the index of weightings of landmarks. As per Nardini et al. (2008), if the distance to the R-defined goal location (g_R) is d_R and distance to the U-defined goal location (g_U) is d_U, the relative proximity to the R-defined location (rp_R) is:

$$rp_R = \frac{\frac{1}{d_R}}{\frac{1}{d_R} + \frac{1}{d_U}} = \frac{d_U}{d_R + d_U} \qquad (1)$$

A larger *rp* value means heavier weighting on the respective landmark. The relative proximity to the other landmark's goal location, i.e., rp_U, is equal to 1 – rp_R. If the two landmarks are equally weighted, then $rp_U = rp_R = 0.5$.

If the information from multiple cues is combined in a weighted average, this can be expressed mathematically by an *Integration Model*. First, the predicted mean search location on shift trials should be constrained to be between the R-defined location (g_R) and the U-defined location (g_U) and the distribution of search responses should be unimodal. Second, compared with single cue trials, the predicted variance on shift trials (σ^2_{Integ}) should be:

$$\sigma^2_{Integ} = rp_R^2\sigma_R^2 + rp_U^2\sigma_U^2 \qquad (2)$$

where $\sigma^2{}_R$ and $\sigma^2{}_U$ are the variance in the locations of the search responses when only the respective single landmark is available (Nardini et al., 2008).

If participants combine the information from two landmarks, another question is whether the combination is *optimal*. The optimal combination is to follow a Bayesian combination (Cheng et al., 2007), where the predicted variance reaches its minimum value, and the weightings of the cues reach the optimal weightings. As per Cheng et al., 2007, the optimal weightings on each landmark (W_R, W_U) can be can be calculated as follows:

$$W_R = \frac{\sigma_U^2}{\sigma_R^2 + \sigma_U^2} \tag{3}$$

$$W_U = \frac{\sigma_R^2}{\sigma_R^2 + \sigma_U^2} \tag{4}$$

Consequently $W_U = 1 – W_R$. By Equation 2, 3 and 4, the optimal predicted variance should be:

$$\sigma_{optimal}^2 = W_R^2 \sigma_R^2 + W_U^2 \sigma_U^2 \tag{5}$$

If the observed variance and weightings do not meet the criteria of optimal (Bayesian) combination, then the combination is called as “sub-optimal” or non-Bayesian combination (Mou & Spetch, 2013).

We compared the predictions of the Integration Model to that of two models that do not assume information is combined. In the *Alternation Model*, only one cue is used on a given trial, but across time, either cue could be used. In other words, participants do not combine the cues, but rather alternate between them. The frequencies of using each cue is positively related to its weighting, such that cues with higher weighting are used more often. Therefore, with two cues, the overall distribution of responses should be a bimodal distribution, or otherwise seen as a mixture of two distributions with variances of $\sigma_R{}^2$ and $\sigma_U{}^2$ and means of R-defined location (μ_R) and U-defined location (μ_U). Variance of the mixture distribution cannot be reduced relative to the single cues, but it is predicted to increase slightly as both cues are used, owing to their separation (Nardini et al., 2008). Thus the predicted variance (σ^2_{Alt}) for this mixture distribution can be calculated as:

$$\sigma_{Alt}^2 = rp_R\left(\mu_R^2 + \sigma_R^2\right) + rp_U\left(\mu_U^2 + \sigma_U^2\right) - \left(rp_R\mu_R + rp_U\mu_U\right)^2 \qquad (6)$$

In the current study, we used rp_U and rp_R, the observed weightings of the cues, to indicate the probabilities of following either cue, as done by Nardini et al. (2008).

In the third model, referred to as the *Hierarchical Model,* only the most dominant cue is used and there is a separate model for each of the two cues; some researchers describe this with the term “exclusivity” because two representations of the same location are mutually exclusive (Baguley et al., 2006; Clark et al., 2013). In this model, the predicted variance is the same as the variance of responses when only the respective single landmark is available (σ^2_R and σ^2_U, respectively). Also, the predicted mean search location on shift trials should be either at the R-defined location (g_R) or the U-defined location (g_U).

Research Design

Each of the present experiments had multiple participant groups (see Table 1). In Experiment 1, training included trials with the single landmarks as well as trials with both landmarks together (indicated by “T” in the group name) and the only difference between the two groups was the line orientation: in Group T-H, the line was horizontal across the screen; in Group T-V the computer monitor was rotated 90°, and participants instead searched along the vertical dimension. In Experiment 2, Group T-many was identical to Group T-H in Experiment 1 with two exceptions: some no-feedback trials were added in training, and 10-px shift trials were replaced by 40-px shift trials. “T” denotes that participants were given training with both cues and single cues; “many” denotes that participants in this group had 252 training trials in total.

The task given to the participants in the other two groups (S-many and S-few) differed from Group T-many in up to two ways: (1) participants did not receive training trials with both cues presented simultaneously, and instead only had single-cue trials (“S”); and (2) participants in Group S-few had fewer training trials, 198 in total (see Table 1).

[Table 1]

Experiment 1

Experiment 1 had two purposes. The first purpose was to examine whether participants combined the information from two vertically or horizontally aligned landmarks when both landmarks were presented singly as well as together in training. If combination occurred, the second purpose was to examine whether the information was optimally combined (in a Bayesian manner).

Method

Participants. A total of 95 university students participated in Experiment 1 and were assigned into two independent groups. After exclusion (see Data Recording section), 91 participants remained. Group T-H (*N* = 47; Age (years): Range = 18-23, Mean = 19.11; 26 females) and Group T-V (*N* = 44; Age (years): Range = 17-22, Mean = 18.91; 25 females) were tested sequentially. Participants provided written consent and the experimental protocol was approved by a University of Alberta ethics board. All participants received credit in an introductory psychology course for their participation.

Apparatus. Participants were tested separately. Stimuli were presented on a 21” computer monitor (resolution: 1920 × 1080 pixels; refresh rate: 60 Hz) using E-Prime (v. 2.0.8.22; Psychology Software Tools Inc., Sharpsburg, PA). For Group T-H, the principal axis of the screen was aligned horizontally. For Group T-V, the computer monitor was rotated 90°

counterclockwise so the principal axis of the screen was aligned vertically. Participants used the mouse to click on screen as their search response. In the following sections, the stimuli are specified in pixels (abbreviated as px; 1 px = 0.26 mm).

Design and procedure.

General method. On each trial, participants were shown a black line (2 px wide) on a computer screen with white background (see Figure 1A), along with one or two colored shapes as landmarks: a red circle with a radius of 15 px (3.97 mm) and a blue square with a width of 30 px (7.94 mm). The black line was presented at the center of the secondary axis of the screen throughout the session. For Group T-H, the cursor always started at the horizontal middle of the screen, 180 px above the line. For Group T-V, the cursor started at the vertical middle of the screen, 180 px to the left of the line. Participants were instructed to search for a goal (a rectangle, 20 px × 40 px, i.e., 5.29 mm × 10.58 mm) on the line by clicking on it with the mouse. The correct response area (i.e., the goal) was defined relative to the landmarks and was not fixed in absolute space. In absolute space, on each training or baseline trial, the center of the correct response area was set at locations with a distance to the left (for Group T-H) or upper (for Group T-V) boundary of the screen on the line, which was pseudorandomly chosen from the nine following values: 480, 600, 720, 840, 960, 1080, 1200, 1320 and 1440 px. On each probe trial, the location was pseudorandomly chosen from the five following values: 480, 720, 960, 1200 and 1440 px. These values were determined as proportions (multiples of 1/16) of the width of the principal axis (i.e., 1920 px), which ensured equal frequencies of appearance at different sections of the line. The position of the landmark-goal array thus varied across trials in absolute space, but was always centered on the line, and array elements were always maintained at the same relative distance to each other.

[Figure 1]

The experiment included three training phases and a testing phase (see Table 1). The participants were not told which phase they were currently in. On each trial during the training phases, participants always received feedback. When participants clicked within the perimeter of the correct response area, the goal appeared as a black rectangle and the phrase “Found the goal!” and the points earned in this trial (e.g., “Score for this trial: 350”) were printed on screen. The calculation of the earned points for each trial was based on the equation:

$$Points \ = \frac{6000 - 1000 \times (number\ of\ clicks) - reaction\ time}{10} \qquad (7)$$

where reaction time (in milliseconds) is for the click within the goal area. On each trial, up to three responses were permitted until the goal was found and negative values were converted to zero so that the points earned on each trial could range from 0 to just under 500. Participants were instructed to be as fast and accurate as possible to maximize the amount of points earned. If participants missed the goal (i.e., all of their three clicks fell outside the goal area), the goal still appeared and the phrase “Oops! You missed the goal!” was printed on the screen. Feedback was shown for 1.5 s, after which the text disappeared along with the stimuli. At this point, the cursor’s position was reset and a 2-s inter-trial interval (ITI) began with the line remaining on the screen until the next trial. At the end of each phase, the accumulated total score (e.g., “Your total score so far: 1200”) as well as the message: “The next trial will begin shortly.” appeared on a blank screen and lasted for 5s until the beginning of the next trial.

Training Phase 1. On each trial in Training Phase 1, participants could see the goal (a white box with a black border) along with a single landmark (either a red circle or a blue square) and were instructed to click on the goal to learn its location. One of the landmarks, referred to as the *Reliable* landmark, was always presented closer to the goal than the other landmark. The center of the reliable landmark was 80 px (i.e., 21.17 mm) from the center of the goal. The other landmark, referred to as the *Unreliable* landmark, was farther from the goal; its center was 120 px (i.e., 31.75 mm) from the center of the goal. The location (on which side of the goal) and identity (circle or square) of the reliable and unreliable landmarks were counterbalanced between participants but remained consistent throughout the experiment for each participant. Training Phase 1 included 9 trials with each landmark (18 trials total) presented in a randomized order.

Training Phase 2. In this phase, all settings were the same as Training Phase 1 except the goal was no longer shown at the beginning of each trial. The goal appeared with the appropriate feedback either (a) when participants found the goal (i.e., clicked within the perimeter of the goal) within three attempts or (b) after they missed the goal on all three clicks. The goal, stimuli and the feedback disappeared during the ITI.

Training Phase 3. This phase introduced trials in which both landmarks were presented together (referred to as *Both-cue* trials); on these trials, the two landmarks were presented at a 200-px distance to one another and in their normal location relative to the goal (80 px from the center of the reliable cue and 120 px from the center of the unreliable cue: see Figure 1A). In this phase, participants were presented with 18 trials identical to the Reliable or Unreliable trials in Training Phase 2, as well as 9 Both-cue trials. All other experimental parameters were identical to Training Phase 2.

Testing phase. Testing included 135 baseline trials and 30 probe trials, which were mixed and randomly presented. Baseline trials were identical to trials in Training Phase 3, with all three trial types (Both-cue, Reliable, and Unreliable) presented for 45 trials each.

On probe trials, participants did not receive feedback (i.e., the goal’s location and how many points they earned was not displayed) even if they clicked within the perimeter of the goal.

Participants were not informed whether a trial was a probe trial when the stimuli were shown at the beginning of each trial. After three clicks, the message: “No feedback on this trial.” appeared instead of the feedback on training trials. Participants had been informed prior to beginning the experiment that there would be trials without feedback on which they would still earn points and that these points would be added to their total score. Participants were given feedback on their accumulated total score after every six probe trials. The points earned on each probe trial were calculated as the current average score per trial.

There were 6 types of probe trials, each presented 5 times. Three types of probe trials were the same as the three baseline trial types (Both-cue, Reliable, and Unreliable) except that no feedback was given on these trials. The other three types of probe trials were new tests: 10-px shift, 20-px shift and Reversal. On 10-px shift trials, both landmarks were shifted 10 px farther from the trained goal location, making them a total of 220 px away from each other. On 20-px shift trials, both landmarks were shifted by 20 px, making them a total of 240 px away from each other. On Reversal tests, the landmark locations were switched relative to the Both-cue trials; after switching, the center of one landmark was at the location of the previous center of the other landmark. The two landmarks maintained the same distance from each other as on baseline trials.

Data recording. On each trial, the X and Y coordinates of the participants’ clicks were recorded relative to the location of the goal. Because of the visible line, responses were very accurate on the axis orthogonal to the landmark/goal array. Therefore, only the principal axis (i.e., horizontal dimension for Group T-H, vertical dimension for Group T-V) was used for data analysis. The signed deviations from the center of the original goal location along the principal axis for each of the three clicks were used to assess each participant’s estimate of the goal location, where a positive value of signed deviation meant a shift from the original goal location towards the reliable cue (see Figure 1B). Probe test data were based on each participant's 15 clicks (3 per trial across the 5 trials). Individual responses were excluded if they were more than three standard deviations from the mean in the respective test. If more than 5 out of 15 responses were excluded for any of the probe tests, the participant was excluded in further analyses. Following these exclusion criteria, data from one participant from Group T-H and three from Group T-V were excluded.

We also recorded the reaction time (RT) in ms of the first click participant made on each trial.

Results

For each probe trial, we averaged the signed deviations across their three click locations. We then calculated the mean and the standard deviation (SD) of the signed deviation across the five trials for each probe trial type and each participant. Preliminary analyses revealed that sex, the identity of the reliable cue, and the location of the reliable cue had no effect on the mean signed deviations [all *p*s > .05]. As a result, data were collapsed across these factors in subsequent analyses. The signed deviations of the Reversal test were analyzed separately and the results are shown in Supplementary Materials.

We examined both the mean of the signed deviation and the variance of the signed deviation for evidence of cue integration. Integration would be indicated by (a) searching on the shift tests at a location that is intermediate to the locations defined by single landmarks and (b) variance reduction when both landmarks were presented compared to when they are presented alone.

In the following analyses, Bonferroni's correction for multiple comparisons *t* tests (corrected α = .05/number of *t* tests) and Greenhouse-Geisser correction for violation of sphericity in ANOVA were applied where applicable. Cohen's *d*s were calculated as $\sqrt{2}d_z = \sqrt{2}\left(\frac{M_D}{SD_D}\right)$ where D denotes the difference between the two groups.

Mean of signed deviations. Signed deviation as a function of probe trial type (Reliable, Unreliable, Both-cue, 10-px shift, 20-px shift) and group (T-H, T-V) is plotted in Figure 2. In general, mean signed deviations were small and in most cases within the range of the 20 pixel wide goal area. On single cue and Both-cue tests the signed deviations were close to zero or very slightly negative, whereas on shift tests, signed deviations were most often positive, indicating that the mean searching location was shifted toward the reliable landmark.

Deviation from original goal location. The mean of signed deviations from the original goal location on both the 10-px and 20-px shift tests were significantly above 0 for both Group T-H and T-V according to one-sample *t* tests [all *t*s > 7.35, *p*s < .001, Cohen's *d*s > 1.52]. To ensure that the positive deviations reflected the assignment of more weight to the reliable cue rather than a more general response bias, we also tested whether the signed deviations in the single cue (Reliable and Unreliable) and Both-cue tests differed from 0. Signed deviations on these other trial types were not significantly different from 0 [$p >$ corrected $\alpha = .01$] with one exception: the signed deviation on the Reliable probe trials in Group T-H was significantly below 0 [$t(46) = 3.28$, $p = .002$, Cohen's $d = 0.68$], indicating that participants searched slightly, but significantly, farther away from the reliable landmark than they should. Thus, the deviation toward the reliable cue on shift tests appears to reflect greater weight being given to the reliable cue when there is a conflict, rather than a general tendency to respond closer to that cue.

[Figure 2]

Comparing shift tests to the Both-cue test. A 5 × 2 mixed ANOVA (Trial Type [Reliable, Unreliable, Both-cue, 10-px shift, 20-px shift] × Group [T-H, T-V]) showed significantly different signed deviations across the five types of trials [$F(2.35, 208.94) = 35.07$, $p < .001$, $\eta_p^2 = .28$]. Neither the main effect of group [$F(1, 89) = 0.42$, $p = .517$, $\eta_p^2 < .01$], nor the interaction of Trial Type and Group [$F(2.35, 208.94) = 1.58$, $p = .203$, $\eta_p^2 = .02$] was significant. We therefore collapsed across Group in subsequent analyses. Planned comparisons indicated that the mean of signed deviation was greater on both shift tests than on Both-cue trials [10-px shift: $t(90) = 7.86$, $p < .001$, Cohen's $d = 1.17$; 20-px shift: $t(90) = 11.22$, $p < .001$, Cohen's $d = 1.66$]. Thus, when the landmarks were shifted apart, participants deviated their searches towards the more reliable landmark.

Reaction times. Figure 3 shows the mean reaction time (RT) of the first click on the probe trials. A 5 × 2 mixed ANOVA (Trial Type [Reliable, Unreliable, Both-cue, 10-px shift, 20-px shift] × Group [T-H, T-V]) showed significantly different RT across the five types of trials [$F(3.10, 275.56) = 8.93$, $p < .001$, $\eta_p^2 = .09$]. Neither the main effect of group [$F(1, 89) = 0.78$, $p = .380$, $\eta_p^2 < .01$], nor the interaction of Trial Type and Group [$F(3.10, 275.56) = 1.55$, $p = .201$, $\eta_p^2 = .02$] was significant. We collapsed across Group for subsequent planned comparisons of the Both-cue test with the single cue and shift tests as well as between the two shift tests. These planned comparisons revealed only two significant differences: participants responded faster on Both-cue tests than on Unreliable tests [$t(90) = 3.72$, $p < .001$, Cohen's $d = 0.55$] and responded faster on 10-px shift tests than on 20-px shift tests [$t(90) = 2.60$, $p = .011$, Cohen's $d = 0.38$]. Importantly, there was no significant difference between Both-cue tests and either type of shift tests [both ps > .049 > corrected $\alpha = .017$], indicating that when the landmarks were shifted, participants did not take longer to make a response.

[Figure 3]

Variance reduction and model fitness. The SD of signed deviation in each test for each participant was used as the dependent variable. The observed SD of individual participants on the shift tests are shown in the Supplementary Materials.

Comparisons between Both-cue and single cue tests. Groups T-H and T-V both showed a reduction in variance on the Both-cue test (lower SD in their signed deviation) compared to both single cue tests [all *t*s > 2.60, *p*s < .05 < corrected α = .025, Cohen's *d*s > 0.55; see Figure 4]. As mentioned in the section of model predictions, the SDs in single cue tests are mathematically the same as the predicted SD by Hierarchical Models. Therefore Hierarchical Models would predict that the SD in the Both-cue tests would be the same as the SD in the single cue tests. Our results were not consistent with the Hierarchical Models. This suggests that when both landmarks were presented, the participants used both of them instead of using only one of them, resulting in a smaller variance than would otherwise be predicted.

[Figure 4]

Comparisons between observed SD and predicted SD on shift tests. To test the fit of the observed values to the predictions based on the Integration Model and the Alternation Model, the rp_R values were calculated for each participant (Equation 1). The observed mean rp_R on the 10-px and 20-px shift tests for each group are shown in Figure 5 and these observed rp_R values were used to calculate the SDs predicted by each model (Equation 2 and 6). For each group, we compared the observed SD on the 10-px and 20-px shift tests with the predicted SDs using both paired *t* tests and the Bayesian Information Criterion (*BIC*) model comparison approach (see Tables 2 and 3). Using a Generalized Linear Mixed Model (GLMM), we calculated a *BIC* for each model based on the -2 log likelihood of the fit of the observed data to the predicted model. This calculation was performed through IBM SPSS (Version 23), which uses number of participants (rather than data points) as *n* in the *BIC* calculation. By convention, if the difference between two model fits is larger than two, then the model with a smaller *BIC* is significantly better than the other (Burnham & Anderson, 2004). The BIC values of the two models are summarized in Table 3.

[Table 2]

[Table 3]

For Group T-H, the mean SD on the 10-px shift test [7.78 ± 5.47 px] was significantly smaller than the prediction of the Alternation Model [13.55 ± 7.31 px; *t*(46) = 4.77, *p* < .001, Cohen's *d* = 0.98] but not different from the prediction of the Integration Model [7.58 ± 3.50 px; *t*(46) = 0.26, *p* = .794, Cohen's *d* = 0.05]. This evidence in favor of the Integration Model is further supported the results of the by model comparison, which showed a difference in *BIC* of 31 in favor of the Integration Model. On the 20-px shift test, the mean SD [9.06 ± 4.39 px] was significantly lower than the prediction of the Alternation Model [13.64 ± 6.39 px; *t*(46) = 5.00, *p* < .001, Cohen's *d* = 1.03] and greater than the prediction of the Integration Model [7.42 ± 2.99 px; *t*(46) = 3.31, *p* = .002, Cohen's *d* = 0.68]. Although the *t* tests results failed to show clear support for either model, the model comparison showed a difference in *BIC* of 28.17, which is strong evidence in favor of the Integration Model. Thus, the observed data for Group T-H on both the 10 px and 20 px shift tests were better fit by the Integration Model than the Alternation Model.

[Figure 5]

For Group T-V, on the 10-px shift test, the mean SD [7.34 ± 3.81 px] was significantly lower than the prediction of the Alternation Model [18.79 ± 18.35 px; *t*(43) = 4.16, *p* < .001, Cohen's *d* = 0.89], but did not differ from the prediction of the Integration Model [8.38 ± 6.00

px; *t*(43) = 1.23, *p* = .224, Cohen's *d* = 0.26]. The model comparison showed a difference of 22.50 in *BIC*, which is very strong evidence in favor of the Integration Model. Consistency with the Integration Model was also found on the 20-px shift test, where the mean SD [10.09 ± 6.33 px] was significantly lower than the prediction of the Alternation Model [19.72 ± 19.05 px; *t*(43) = 3.31, *p* = .002, Cohen's *d* = 0.71] but did not differ from the prediction of the Integration Model [8.69 ± 6.54 px; *t*(43) = 1.34, *p* = .188, Cohen's *d* = 0.28]. The model comparison showed a difference in *BIC* of 14.87, which is strong evidence in favor of the Integration Model. Thus, both groups showed evidence of integration on both shift tests.

The optimal integration of information from two landmarks. To examine whether the information from two landmarks was optimally combined, we compared the observed SD in the Both-cue test with the optimal SD predicted by optimal integration weightings (Equation 5; see Figure 5). Additionally, for shift tests, we compared the rp_R based on observed data with the predicted optimal integration weightings (Equation 3 and 4) and compared the observed SD on shift tests with the optimal SD. The results of the following analyses were summarized in Table 4.

[Table 4]

Both-cue test. The observed SD on Both-cues tests was not significantly different from the predicted optimal SD for either Group T-H [6.11 ± 2.51 px; *t*(46) = 1.00, *p* = .323, Cohen's *d* = 0.21] or Group T-V [6.60 ± 5.37 px; *t*(43) = 0.25, *p* = .803, Cohen's *d* = 0.05]. The search behavior of both groups was consistent with the prediction of optimal integration.

Shift tests. For Group T-H, the observed rp_R was greater than the optimal weighting [0.59 ± 0.23] on both the 10-px shift test [0.74 ± 0.22; *t*(46) = 2.90, *p* = .006, Cohen's *d* = 0.59] and the 20-px shift test [0.73 ± 0.16; *t*(46) = 3.00, *p* = .004, Cohen's *d* = 0.62]. The observed SD was different significantly from the predicted optimal SD on the 20-px shift test [$t(46) = 4.88$, $p < .001$, Cohen's $d = 1.00$] but not on the 10-px shift test [$t(46) = 2.03$, $p = .048 >$ corrected $\alpha = .025$, Cohen's $d = 0.42$].

For Group T-V, similarly, the observed rp_R on the 10-px shift test [0.75 ± 0.20] was greater than the optimal weighting [0.60 ± 0.28; $t(43) = 2.61$, $p = .013$, Cohen's $d = 0.56$]. However, the observed rp_R on the 20-px shift test [0.71 ± 0.17] did not differ from the optimal weighting [$t(43) = 2.01$, $p = .050 >$ corrected $\alpha = .025$, Cohen's $d = 0.43$]. Again, the observed SD was significantly different from the predicted optimal SD [6.60 ± 5.37 px] on the 20-px shift [$t(43) = 3.27$, $p = .002$, Cohen's $d = 0.70$] but not on the 10-px shift [$t(43) = 0.92$, $p = .364$, Cohen's $d = 0.20$].

Considering that for both of the shift tests and for both groups, at least one of the measures deviated significantly from the optimal prediction, these results suggest that integration was sub-optimal on shift tests.

Discussion

Based on the evidence of intermediate searching between single-cue-defined-locations, variance reduction, and model fit, Experiment 1 indicated that when two landmarks were shifted, people combined the information from the two landmarks with more weighting given to the more reliable landmark. The results were similar for Group T-H and Group T-V suggesting that weighted integration occurred similarly in both dimensions of space. Therefore, only the horizontal orientation of the stimuli was used in subsequent experiments. The unequal weighting given to the two landmarks supports the suggestion that cue reliability is influenced by proximity to the goal (Ruprecht et al., 2014) and also indicates that our manipulation of the cue reliability was successful. The integration found in Experiment 1 is consistent with other evidence for cue combination in humans and animals (Byrne & Crawford, 2010; Cheng et al., 2007; Legge et al., 2016).

Surprisingly, RT for the Both-cue test did not differ significantly from the shift tests as would be expected if the process of averaging conflicting information required more time. It seems that even when the cues provided conflicting information, the weighted averaging process did not involve additional computation time, at least not as reflected in the observed RT.

Experiment 2

In this experiment we investigated whether experiencing both cues together during the learning phase is an important determinant of whether cue integration occurs. We tested three groups that differed in their training conditions. One of the groups had both cues presented in training just as in Experiment 1, whereas the other two groups were trained without any trials in which both cues were presented together. The last two groups differed from each other only in the number of training trials. If the three groups showed different results in terms of cue integration, this would suggest that the learning experience influences the process of integration.

In Experiment 2, we also tested with larger shifts. Although we did not find consistent differences between the 10-px and 20-px shift tests in terms of whether participants combined the information, both of these shifts were fairly small and may not have been very detectable. Therefore, in Experiment 2 we increased the shift amount between the landmarks to test whether integration would still occur with a larger, and presumable noticeable shift in landmark locations. Finally, we included some no feedback training prior to the test phase to better prepare participants for the test phase.

Method

Participants. A total of 131 university students, who all received course credit for their participation, were assigned into three experimental groups. After exclusion (see Data Recording section in Experiment 1), there were 123 participants. Training for one group (T-many) included trials with both cues together, whereas training for the other groups (S-many and S-few) did not; the two S groups differed only in the number of training trials (see Table 1). Groups T-many and S-few were tested first, with participants randomly assigned into these two groups, and Group S-many was tested subsequently (S-many). Group T-many included 39 people (Age (years): Range = 18-29, Mean = 19.79; 28 females), Group S-few included 42 people (Age (years): Range = 18-24, Mean = 19.24; 30 females, one individual answered prefer-not-say), and Group S-many included 42 people (Age (years): Range = 18-31, Mean = 20.12; 27 females).

The exclusion criteria were the same as in Experiment 1. Four participants from Group T-many, two from Group S-few, and two from Group S-many were excluded from the following analyses.

Apparatus, design, and procedure. The apparatus was the same as for Group T-H in Experiment 1. Each group received four training phases and one testing phase (see Table 1). As the duration of training was longer than Experiment 1, written instructions were added between phases.

For all groups, Training Phases 1 and 2 were identical to Experiment 1 (Group T-H) except that instructions were added to the beginning of Training Phase 2 stating that the goal would be hidden. In Training Phase 3, no-feedback trials were added in order to familiarize participants to trials without feedback (i.e., identical to test trials) and instructions at the beginning of this phase stated that some trials would not give feedback. In this third training phase, all participants were given 9 trials with feedback (identical to Training Phase 2) and 3 no-feedback trials with each landmark (24 trials total) in a randomized order. On no-feedback trials, the stimuli were the same as trials with feedback but no feedback was provided even if participants clicked within the correct response area; instead, after three clicks, participants were presented with the message: “No feedback on this trial.” On each no-feedback trial, the center of the correct response area was pseudo randomly chosen from three distances (in pixels) from the boundary of the screen: 480, 960 and 1440, to match the locations that would be used in the subsequent probe trials.

Training Phase 4 always provided feedback. The main difference between the groups was the type and number of training trials. For Group T-many, participants were presented 27 trials (9 trials each for Reliable, Unreliable, and Both-cue trial types), which were identical to those in Training Phase 3 for Group T-H in Experiment 1. For the other two groups, Training Phase 4 only included single landmark trials (Reliable, Unreliable). For Group S-few, participants received 9 trials for each type (18 trials in total). For Group S-many, participants received 13 trials for each trial type (26 trials in total).

In the Testing Phase, the number and type of baseline trials differed for the 3 groups and were the same as in Training Phase 4, but the probe trials were identical for the three groups. Probe tests included Reliable, Unreliable, Both-cue, 20-px shift, 40-px shift, and Reversal, with 5 trials per test. The Reliable, Unreliable, Both-cue, 20-px shift, and Reversal trials were the same as those described for Group T-H in Experiment 1. On 40-px shift trials, both landmarks were shifted by 40 px, making them a total of 280 px apart. For Group T-many, there were 135 baseline trials (45 trials of each type) and for Group S-few there were 90 baseline trials (45 trials of each type, thus matching Group T-many). For Group S-many, there were 136 baseline trials (68 trials of each type) such that the total number of training, baseline and probe trials in Group

S-many was equivalent to that in Group T-many (252 trials in total; see Table 1). All other aspects of the procedure were identical to Group T-H in Experiment 1.

Results

As in Experiment 1, the mean and the SD of signed deviation across the five trials for each probe trial type and each participant were calculated. As sex, the identity of reliable cue, and the location of reliable cue had no effect on signed deviation [all *p*s > .10], we collapsed data across these factors in subsequent analyses. The analysis of signed deviations in the Reversal test is reported in the Supplementary Materials. Again, we examined the mean and the variance of signed deviation separately.

Mean of signed deviations. As illustrated in Figure 2, mean signed deviations in general were small and in most cases within the range of the goal area. For Group T-many (trained with both cues), on single-cue and Both-cue tests the signed deviations were close to zero or very slightly negative, whereas on shift tests, signed deviations were most often positive, indicating a shift toward the reliable landmark. For the two groups trained with only single cues (Groups S-few and S-many), on all types of tests, the signed deviations were close to zero, indicating that the searching was not biased toward either of the two landmarks.

Deviation from original goal location. For Group T-many, when the landmarks were shifted, the signed deviation was significantly above 0 [both *t*s > 4.43, *p*s < .001, Cohen's *d*s > 1.33]. However, for Group S-few and Group S-many, the signed deviations were not different from 0 on the shift tests [all *t*s < 1.09, *p*s > .285, Cohen's *d*s < 0.24; see Figure 2]. Signed deviations on the single-cue or Both-cue tests were not different from 0 for any of the groups [*p* > corrected *α* = .01], indicating that the positive deviation for Group T-many on shift tests was not due to a general response bias.

Comparing shift tests to the Both-cue test. A mixed ANOVA (Trial Type [Reliable, Unreliable, Both-cue, 20-px shift, 40-px shift] × Group [T-many, S-few, S-many]) on the signed deviation scores showed a significant main effect of trial type [$F(3.16, 378.60) = 6.76$, $p < .001$, $\eta_p^2 = .05$]. The main effect of Group [$F(2, 120) = 3.16$, $p = .046$, $\eta_p^2 = .05$] and the interaction between Trial Type and Group [$F(6.31, 378.60) = 2.48$, $p = .021$, $\eta_p^2 = .04$] were also significant. Therefore, the signed deviations of three groups were examined separately. For Group T-many, the mean signed deviations were larger on the 20-px shift and 40-px shift tests than on the Both-cue test [both *t*s > 3.61, *p*s < .002, Cohen's *d*s > 0.81]. After correcting for multiple comparisons, the mean signed deviations on the 20-px shift and 40-px shift tests were not different from that in the Both-cue test for either Group S-few [both *t*s < 1.47, *p*s > .149, Cohen's *d*s < 0.32] or Group S-many [both *t*s < 2.07, *p*s > .044, Cohen's *d*s < 0.45].

The results indicate that when the two landmarks were shifted, participants in Group T-many deviated their searches towards the more reliable landmark. The other two groups did not show significant deviations.

Reaction times. We examined the reaction time (RT) of the first click on each trial in testing phase (see Figure 3). For Group T-many, there were no differences in RT between the Both-cue test, the Reliable test, and the Unreliable test [all *p*s > .05]. However, participants responded faster on the Both-cue test than on the 40-px shift test [$t(38) = 3.08$, $p = .004$, Cohen's $d = 0.70$].

For the two groups with only single cue training, participants responded slower on the Both-cue test than on either the Reliable test [Group S-few: $t(41) = 4.81$, $p < .001$, Cohen's $d = 1.05$; Group S-many: $t(41) = 4.98$, $p < .001$, Cohen's $d = 1.09$] or the Unreliable test [Group S-few: $t(41) = 3.62$, $p = .001$, Cohen's $d = 0.79$; Group S-many: $t(41) = 3.47$, $p = .001$, Cohen's $d =$

0.76]. There was no difference in RT between the Both-cue test and the shift tests, nor between the two shift tests [all *p*s > .12].

Variance reduction and model fitness. As in Experiment 1, the SD of signed deviation in each test for each participant was used as the dependent variable in the analyses on variance (see Figure 4). The observed SD of individual participants on the shift tests are shown in the Supplementary Materials.

Comparisons between Both-cue and single cue tests. Similar to Experiment 1, for Group T-many, the mean SD on the Both-cue test was smaller than the mean SD on the Unreliable test (which is the predicted SD by the Hierarchical-Unreliable Model) [*t*(38) = 3.57, *p* = .001, Cohen's *d* = 0.81]. However, the mean SD on the Both-cue test was not significantly smaller than the mean SD on the Reliable test (which is the predicted SD by the Hierarchical-Reliable Model) [*t*(38) = 1.47, *p* = .151, Cohen's *d* = 0.33]. For Group S-few, the mean SD on the Both-cue test was greater than on either of the single cue tests [both *t*s > 3.23, *p*s < .003, Cohen's *d*s > 0.70]. For Group S-many, the mean SD on the Both-cue test was greater than that on the Reliable test [*t*(41) = 2.73, *p* = .009, Cohen's *d* = 0.60] and similar to that on the Unreliable test [*t*(41) = 2.01, *p* = .051, Cohen's *d* = 0.44].

Comparisons between observed SD and predicted SD on shift tests. The observed rp_R values and predicted SDs by models were calculated for each participant (Equation 1, 2 and 6; see Table 2 and Figure 5). For Group T-many, the observed SD on the 20-px shift test [10.41 ± 7.21 px] was not different from the predicted SD by either the Alternation Model [15.15 ± 12.55 px; *t*(38) = 2.14, *p* = .039 > corrected *α* = .025, Cohen's *d* = 0.48] or the Integration Model [9.26 ± 9.19 px; *t*(38) = 0.69, *p* = .497, Cohen's *d* = 0.16] and the *t* tests did not differentiate between the models. However, the model comparison (Table 3) showed a *BIC* difference of 7.02, which is strong evidence in favor of the Integration Model. Thus, the Integration Model fit the observed data for Group T-many on the 20-px shift test better than the Alternation Model.

On the 40-px shift test, the observed SD [12.84 ± 8.49 px] also did not differ from the prediction of either the Alternation Model [15.46 ± 13.29 px; $t(38) = 1.00$, $p = .325$, Cohen's $d = 0.23$] or the Integration Model [7.42 ± 2.99 px; $t(38) = 2.04$, $p = .048 >$ corrected $\alpha = .025$, Cohen's $d = 0.46$] and the *t* tests did not differentiate between the models. The model comparison showed a *BIC* difference of 2.24, which provides some positive evidence in favor of the Alternation Model. Thus, in contrast to the 20-px shift test, the *BIC* results suggest that Alternation Model fit the observed data for Group T-many on the 40-px shift at least as well, and possibly better than the Integration Model.

As predicted, the SD comparison results for the other two groups differed from Group T-many. For Group S-few, on the 20-px shift test, the mean SD [25.49 ± 28.44 px] was not only significantly greater than the prediction of the Integration Model [7.63 ± 3.21 px; $t(41) = 4.20$, $p < .001$, Cohen's $d = 0.92$] but also significantly greater than the prediction of the Alternation Model [12.15 ± 4.33 px; $t(41) = 3.03$, $p = .004$, Cohen's $d = 0.66$]. On the 40-px shift test, likewise, the mean SD [30.68 ± 24.54 px] was significantly greater than the prediction of the Integration Model [7.35 ± 3.01 px; $t(41) = 6.22$, $p < .001$, Cohen's $d = 1.36$] and the prediction of the Alternation Model [12.36 ± 3.88 px; $t(41) = 4.83$, $p < .001$, Cohen's $d = 1.06$]. For Group S-many on the 20-px shift test, the mean SD [17.47 ± 14.86 px] was significantly greater than the prediction of the Integration Model [7.43 ± 4.98 px; $t(41) = 4.45$, $p < .001$, Cohen's $d = 0.97$] and the prediction of the Alternation Model [11.32 ± 6.49 px; $t(41) = 2.66$, $p = .011$, Cohen's $d = 0.58$]. On the 40-px shift test, the mean SD [29.61 ± 20.53 px] was significantly greater than the prediction of the Integration Model [7.60 ± 5.61 px; $t(41) = 6.82$, $p < .001$, Cohen's $d = 1.49$]

and the prediction of the Alternation Model [11.65 ± 6.53 px; $t(41) = 5.74$, $p < .001$, Cohen's $d = 1.25$]. The results from these t tests were conclusive, making calculating of formal model selection (*BIC*s) for these groups unnecessary: Neither model was supported in any of the shift tests in Group S-few and S-many.

The optimal integration of information from two landmarks. To further examine whether the information from two landmarks was optimally combined, the observed SDs for the Both-cue test were compared with the optimal SD predicted by optimal integration weightings.

Both-cue test. The observed SD for Group T-many in the Both-cue test did not differ significantly from the predicted optimal SD [6.47 ± 3.59 px; $t(38) = 0.32$, $p = .751$, Cohen's $d = 0.07$; see Table 4]. In contrast, the observed SD in the Both-cue test was greater than the predicted optimal SD for both of the groups trained with single cues only [Group S-few: 5.94 ± 2.08; $t(41) = 4.26$, $p < .001$, Cohen's $d = 0.93$; Group S-many: 6.16 ± 4.77; $t(41) = 3.49$, $p = .001$, Cohen's $d = 0.76$]. Therefore, only the group trained with both cues together (Group T-many) showed evidence of optimal integration on the Both-cue test.

Shift tests. The only fit of observed SDs with the integration model was found in Group T-many for the 20-px shift. Therefore, we compared observed rp_R with the predicted optimal integration weightings (Equation 3 and 4) and observed SDs with the optimal SD (see Table 4 and Figure 5) only for this group and condition. This comparison showed that the observed rp_R in the 20-px shift test [0.64 ± 0.20] for Group T-many was consistent with the optimal weighting [0.60 ± 0.30; $t(38) = 0.87$, $p = .390$, Cohen's $d = 0.20$]. However, the observed SD was significantly greater than the predicted optimal SD [6.47 ± 3.59 px], in the 20-px shift test [$t(38) = 3.36$, $p = .002$, Cohen's $d = 0.76$], suggesting that the integration on the 20-px shift test was sub-optimal.

Discussion

We had two novel findings in Experiment 2. First, the shift distance influenced the occurrence of integration. For the group trained with both landmarks together (Group T-many), in the 20-px shift test, the observed data were best explained by the Integration Model, consistent with the results in Experiment 1. However, in the 40-px shift test, the observed data were at least as well, and nominally better, explained by the Alternation Model. The finding that integration is less likely to occur when there is a large discrepancy between the locations specified by each cue is consistent with previous suggestions: for example Cheng et al. (2007) suggested that it may be beneficial to combine cues when the subjective discrepancy is small, but to rely on a single source of information if the cues are too discrepant. Second, integration occurred only in the group that was trained with both cues together (Group T-many), not in either of the two groups that received no training with the two cues together (Groups S-few and S-many).

The two groups that did not see both cues simultaneously in training (Groups S-few and S-many) showed longer reaction times on the Both-cue test and shift tests compared to the group trained with both cues (Group T-many). It is possible that the groups trained with single cues treated the tests with two landmarks as a new situation, and the novelty of this stimulus would be expected to give rise to longer reaction times. The finding that variability in search location was not reduced when the two landmarks were presented together at their trained distances from the goal suggest that the learned information from the two landmarks was not combined to better locate the goal. Even additional learning experience with single cues (Group S-many) did not facilitate the combination of landmark information.

General Discussion

In the current study, we compared observed data to model predictions to examine whether humans combined information from multiple landmarks to localize a goal. We found that the hierarchical model, which assumes that only one cue is used, never fit the observed data in any of the groups. These results differ from previous studies with a comparable one-dimensional search task (Baguley et al., 2006; Clark et al., 2013) because in those studies, the results supported an exclusive model (in our terminology, the Hierarchical Model). This might be due to a difference in the stimuli used in the experiments. In those studies, the landmarks always appeared at the left and right edges of the screen and the participants may have used the edges to encode the location of the target. However, in the current study, the landmark locations and corresponding goal location on the screen varied from trial to trial. Our procedure may have encouraged learning the spatial relationship between each landmark and the goal, as well as between the two landmarks when they were seen together during training.

Whether people showed cue combination depended on how the information is learned, and specifically whether both cues were experienced together during training. In Experiment 2, participants who were trained with both landmarks showed integration on 20-px shift trials, whereas participants who were trained only with single cues did not show this combination. In fact, the response variability on shift tests for participants trained with single cues was even greater than predicted by the Alternation Model, suggesting random noise, possibly due to the novelty of seeing the two cues together. Our findings in Experiment 2 were not consistent with the findings of Baguley et al. (2006) that learning two cues together did not facilitate the combination of two sources of information, as Group T-many did combine the information. This inconsistency may be due to differences in paradigms; whereas the current study used a single pair of landmarks in a continuous response space, Baguley et al. used multiple pairs of landmarks and discrete response locations. Furthermore, our task required participants to memorize much less information about the identity or perceptual properties of the landmarks and the goal.

We speculate that training with both cues in our study provided participants with information that the two landmarks are spatially related, and that this information may be an important determinant of whether spatial integration occurs. For example, when trained with more than one landmark, humans (Spetch, Cheng, & MacDonald, 1996) and, under some circumstances, birds (Kamil & Jones, 1997; Spetch, Rust, Kamil, & Jones, 2003; Sturz & Katz, 2009) will learn the relative bearings or distance between the landmarks, and will search in a way to maintain a relational rule when the landmarks are shifted. It is possible that integration may depend on whether the location is encoded, not only in terms of absolute metrics, but also in terms of these relative metrics. Moreover, previous studies on using a boundary to memorize locations suggest that knowing the link between different spatial reference points is important for accurate localization performance (Zhou & Mou, 2016). Our results suggest that knowing this link may also facilitate combining these sources of information, provided that conflict between them is not too large.

The degree of incongruence from previous encoded information (i.e., the shift amount of landmarks) also influenced the occurrence of integration. Integration was seen when the landmarks were shifted by a small amount but not when they were shifted by a large amount. Both groups in Experiment 1 showed integration in response to the small (10-px and 20-px) shifts of the landmarks. In Experiment 2, Group T-many also showed integration in response to the 20-px shift but did not show integration when the landmarks were shifted apart by 40 px. On the 40-px shift tests, the modeling results indicated a nominally better fit to the Alternation model, suggesting that participants instead may have alternated their responses between the two possible goal locations defined by single landmarks (the R-defined location g_R and the U-defined location g_U). Furthermore, the participants showed longer reaction times on the 40-px shift test than in the Both-cue test, suggesting that they noticed the large shift, and may have abandoned any attempt to combine the information. This difference based on degree of discrepancy is consistent with previous suggestions and findings (Cheng et al., 2007; Körding et al., 2007; Pfuhl, Tjelmeland, & Biegler, 2011). As discussed by Pfuhl et al., a small discrepancy may indicate measurement error, but a large discrepancy may indicate that the identity or location of at least one of the landmarks may have changed. However, the observed SD on our 40-px shift test was not higher than the model predictions, which suggests that participants did not represent the training experience as being completely irrelevant to the large shift test.

Why might use of an alternation strategy take more time than using an integration strategy? One possibility is that there is a deliberation process that gets activated when participants notice a shift. In the Both-cue and small shift tests, we found integration in all groups trained with both cues. However, in the larger 40-px shift test the participants in Group T-many appeared to be more likely to use an alternation strategy than in the other manipulations. Thus, if the conflict amount is too large to disregard, people may use the cues in a weighted alternation fashion. This may or may not be a conscious process, but it appears to increase reaction times.

Previous studies have also found that the amount of incongruence between different cues can affect whether integration occurs (Sjolund, 2014; Zhao & Warren, 2015b). For example, Zhao and Warren (2015b) examined the use of path integration and landmarks for homing direction and found that when the incongruence was large, people did not combine the information but rather used only one cue at a time. Our results are consistent with their findings. The effect of shift amount could provide another plausible reason for the discrepancy between our results and those of previous studies with a comparable one-dimensional search task (Baguley et al., 2006; Clark et al., 2013). The landmark shifts in these previous studies were larger than those used in the current study, and we only found integration for small shift amounts.

Although the manipulations in the current study resemble real life experiences, in that people may see single or multiple landmarks simultaneously, our findings are based on experiments performed in a small-scale, non-navigable environment. Therefore, further research is needed to determine whether integration of information from multiple landmarks also depends on how the information is encoded (simultaneously or on different occasions) and on the degree of conflict between the cues in large-scale navigable environments.

Acknowledgements

This research was funded by Natural Sciences and Engineering Research Council of Canada to Marcia L. Spetch. We thank the members of Spetch lab for their contribution in data collection. We have no conflict of interest to declare.

References

Baguley, T., Lansdale, M. W., Lines, L. K., & Parkin, J. K. (2006). Two spatial memories are not better than one: Evidence of exclusivity in memory for object location. *Cognitive Psychology*, *52*, 243–289. http://doi.org/10.1016/j.cogpsych.2005.08.001

Bodily, K. D., Daniel, T. A., & Sturz, B. R. (2012). The roles of beaconing and dead reckoning in human virtual navigation. *Learning and Motivation*, *43*, 14–23. http://doi.org/10.1016/j.lmot.2012.01.002

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–304. http://doi.org/10.1177/0049124104268644

Byrne, P. A., & Crawford, J. D. (2010). Cue reliability and a landmark stability heuristic determine relative weighting between egocentric and allocentric visual information in memory-guided reach. *Journal of Neurophysiology*, *103*, 3054–3069. http://doi.org/10.1152/jn.01008.2009

Chen, X., & McNamara, T. P. (2014). Bayesian cue interaction in human spatial navigation. In *Spatial Cognition Conference 2014* (pp. 147–160).

Cheng, K., Shettleworth, S. J., Huttenlocher, J., & Rieser, J. J. (2007). Bayesian integration of spatial information. *Psychological Bulletin*, *133*, 625–637. http://doi.org/10.1037/0033-2909.133.4.625

Clark, D. P. A., Dunn, A. K., & Baguley, T. (2013). Testing the exclusivity effect in location memory. *Memory*, *21*, 512–523. http://doi.org/10.1080/09658211.2012.744421

Doeller, C. F., & Burgess, N. (2008). Distinct error-correcting and incidental learning of location relative to landmarks and boundaries. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 5909–5914. http://doi.org/10.1073/pnas.0711433105

Doeller, C. F., King, J. A., & Burgess, N. (2008). Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 5915–5920. http://doi.org/10.1073/pnas.0801489105

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*, 107–143. http://doi.org/10.1111/j.1756-8765.2008.01006.x

Huttenlocher, J., Hedges, L. V., Corrigan, B., & Crawford, L. E. (2004). Spatial categories and the estimation of location. *Cognition*, *93*, 75–97. http://doi.org/10.1016/j.cognition.2003.10.006

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, *129*, 220–241. http://doi.org/10.1037/0096-3445.129.2.220

Ishikawa, T., & Montello, D. R. (2006). Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive Psychology*, *52*(2), 93–129. http://doi.org/10.1016/j.cogpsych.2005.08.003

Kamil, A. C., & Cheng, K. (2001). Way-finding and landmarks: the multiple-bearings hypothesis. *The Journal of Experimental Biology*, *204*, 103–113.

Kamil, A. C., & Jones, J. E. (1997). The seed-storing corvid Clark's nutcracker learns geometric relationships among landmarks. *Nature*, *390*(6657), 276–279. http://doi.org/10.1038/36840

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE*, *2*(9), e943. http://doi.org/10.1371/journal.pone.0000943

Lea, S. E. G., Wills, A. J., Leaver, L. a, Ryan, C. M. E., Bryant, C. M. L., & Millar, L. (2009). A comparative analysis of the categorization of multidimensional stimuli: II. Strategic information search in humans (Homo sapiens) but not in pigeons (Columba livia). *Journal of Comparative Psychology*, *123*, 406–420. http://doi.org/10.1037/a0016851

Legge, E. L. G., Madan, C. R., Spetch, M. L., & Ludvig, E. A. (2016). Multiple cue use and integration in pigeons (Columba livia). *Animal Cognition*, *19*, 581–591. http://doi.org/10.1007/s10071-016-0963-8

Legge, E. L. G., Spetch, M. L., & Batty, E. R. (2009). Pigeons' (Columba livia) hierarchical organization of local and global cues in touch screen tasks. *Behavioural Processes*, *80*, 128–139. http://doi.org/10.1016/j.beproc.2008.10.011

Mou, W., & Spetch, M. L. (2013). Object location memory: Integration and competition between multiple context objects but not between observers' body and context objects. *Cognition*, *126*, 181–197. http://doi.org/10.1016/j.cognition.2012.09.018

Nardini, M., Jones, P., Bedford, R., & Braddick, O. (2008). Development of cue integration in human navigation. *Current Biology*, *18*, 689–693. http://doi.org/10.1016/j.cub.2008.04.021

Pantelides, S. N., Kelly, J. W., & Avraamides, M. N. (2016). Integration of spatial information across vision and language. *Journal of Cognitive Psychology*, *28*, 171–185. http://doi.org/10.1080/20445911.2015.1102144

Pfuhl, G., Tjelmeland, H., & Biegler, R. (2011). Precision and reliability in animal navigation. *Bulletin of Mathematical Biology*, *73*(5), 951–977. http://doi.org/10.1007/s11538-010-9547-y

Ruprecht, C. M., Wolf, J. E., Quintana, N. I., & Leising, K. J. (2014). Feature-positive discriminations during a spatial-search task with humans. *Learning & Behavior*, *42*, 215–230. http://doi.org/10.3758/s13420-014-0140-3

Sampaio, C., & Wang, R. F. (2009). Category-based errors and the accessibility of unbiased spatial memories: a retrieval model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1331–1337. http://doi.org/10.1037/a0016377

Sandberg, E. H., Huttenlocher, J., & Newcombe, N. (1996). The development of hierarchical representation of two-dimensional space. *Child Development*, *67*(3), 721–739. http://doi.org/10.1111/j.1467-8624.1996.tb01761.x

Sjolund, L. A. (2014). *Cue integration and competition during navigation*. Iowa State University.

Spetch, M. L., Cheng, K., & MacDonald, S. E. (1996). Learning the configuration of a landmark array: I. Touch-screen studies with pigeons and humans. *Journal of Comparative Psychology*, *110*, 55–68. http://doi.org/10.1037/0735-7036.110.1.55

Spetch, M. L., & Edwards, C. A. (1988). Pigeons’, Columba livia, use of global and local cues for spatial memory. *Animal Behaviour*, *36*(1), 293–296. http://doi.org/10.1016/S0003-3472(88)80274-4

Spetch, M. L., & Mondloch, M. V. (1993). Control of pigeons’ spatial search by graphic landmarks in a touch-screen task. *Journal of Experimental Psychology: Animal Behavior*

Processes, *19*, 353–372. http://doi.org/10.1037/0097-7403.19.4.353

Spetch, M. L., Rust, T. B., Kamil, A. C., & Jones, J. E. (2003). Searching by rules: pigeons' (Columba livia) landmark-based search according to constant bearing or constant distance. *Journal of Comparative Psychology*, *117*, 123–132. http://doi.org/10.1037/0735-7036.117.2.123

Sturz, B. R., & Bodily, K. D. (2010). Encoding of variability of landmark-based spatial information. *Psychological Research*, *74*, 560–567. http://doi.org/10.1007/s00426-010-0277-4

Sturz, B. R., & Katz, J. S. (2009). Learning of absolute and relative distance and direction from discrete visual landmarks by pigeons (Columba livia). *Journal of Comparative Psychology*, *123*, 90–113. http://doi.org/10.1037/a0012905

Yamamoto, N., & Shelton, A. (2008). Integrating object locations in the memory representation of a spatial layout. *Visual Cognition*, *16*, 140–143. http://doi.org/10.1080/13506280701692097

Zhao, M., & Warren, W. H. (2015a). Environmental stability modulates the role of path integration in human navigation. *Cognition*, *142*, 96–109. http://doi.org/10.1016/j.cognition.2015.05.008

Zhao, M., & Warren, W. H. (2015b). How you get there from here: Interaction of visual landmarks and path integration in human navigation. *Psychological Science*, *26*, 1–10.

Zhou, R., & Mou, W. (2016). Superior cognitive mapping through single landmark-related learning than through boundary-related learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1316–1323. http://doi.org/10.1037/xlm0000239

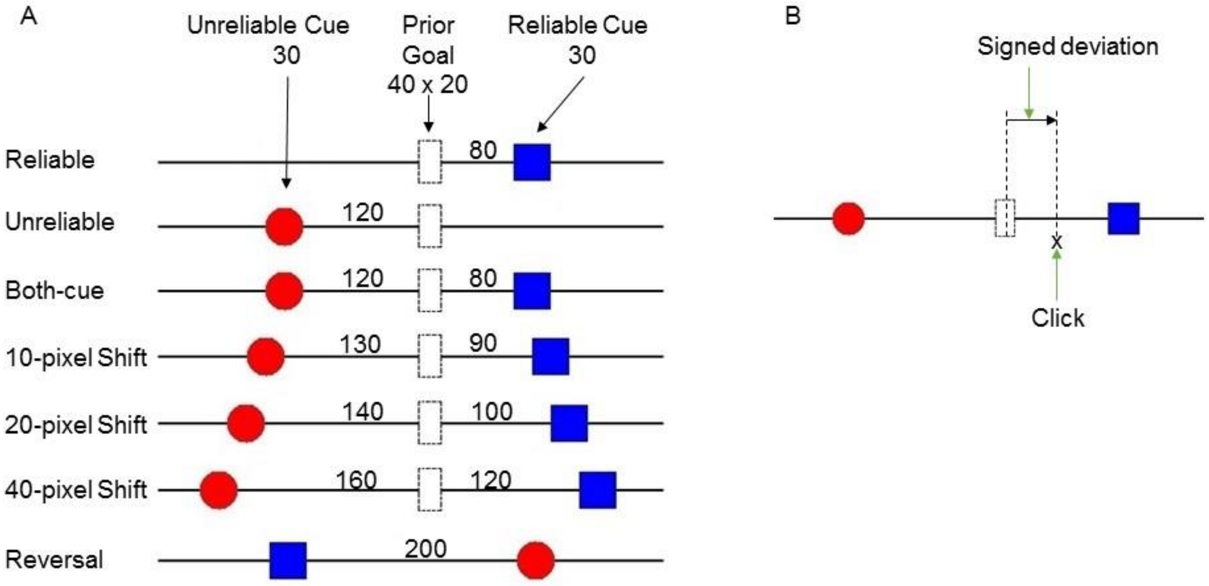

Figure 1. (A) Illustration of the trial conditions in Experiments 1 (Group T-H) and 2, where the square is the reliable cue. The location (on which side of the goal) and identity (circle or square) of the reliable and unreliable landmarks were counterbalanced between participants. Numbers indicate the distance, in pixels, between landmarks or landmark and goal. The rectangle with dashed line shows the original goal location relative to landmarks. The width of the square and the diameter of the circle were 30 px. (B) An example of signed deviation. X represents one click. Signed deviation was the distance from the original goal location to the click's location along the principal axis.

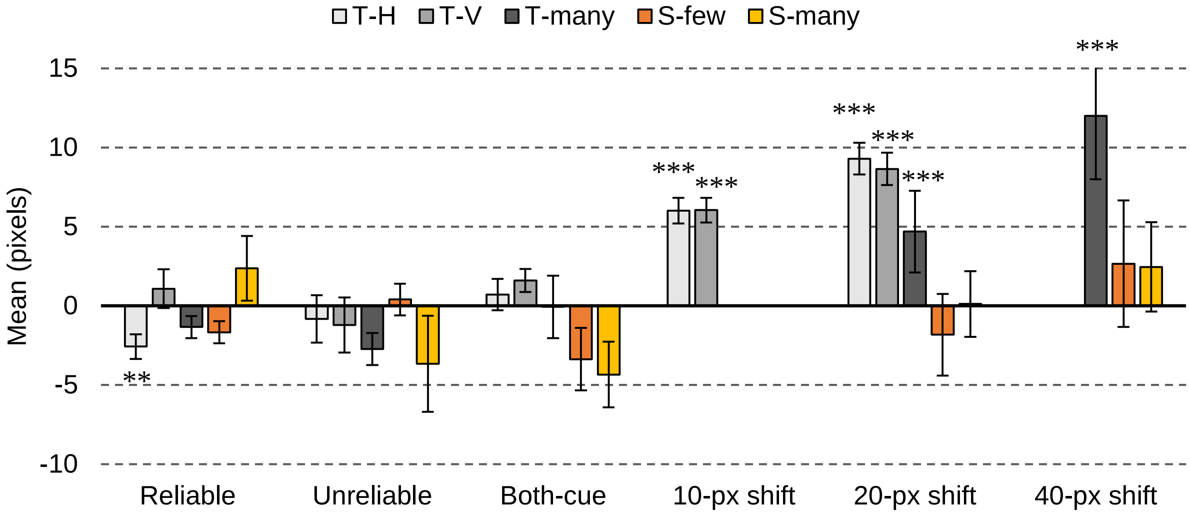

Figure 2. Mean signed deviation for five probe trial types relative to the original goal location (one-sample *t* test: ****p* < .001; ***p* < .01 < corrected *α* = .01). Error bars represent standard errors of the mean. Positive values correspond to searching towards the reliable landmark.

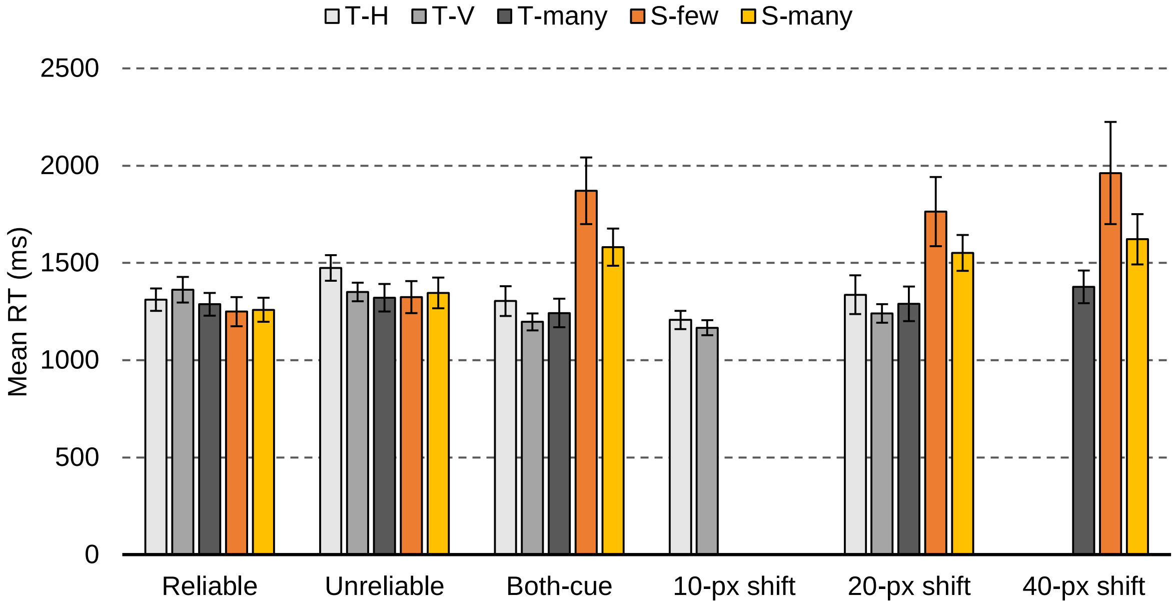

Figure 3. Mean reaction time (RT) for five probe trial types. Error bars represent standard errors of the mean.

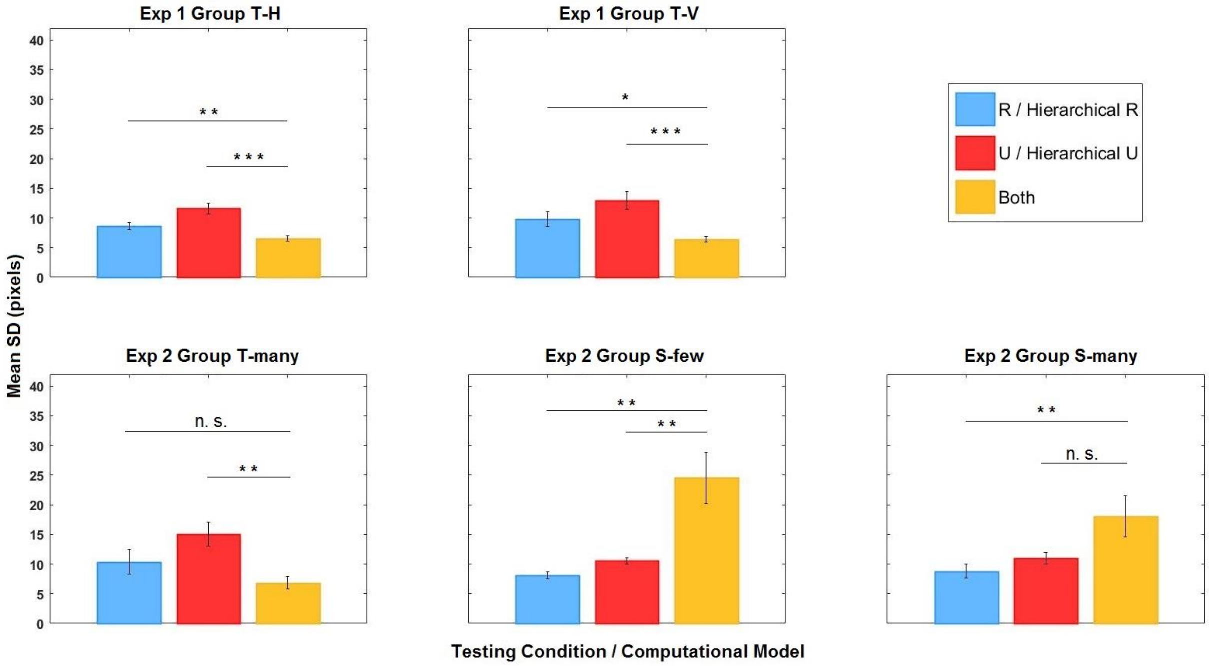

Figure 4. Mean SD of observed signed deviation in single-cue and Both-cue tests. (Paired *t* test: ***p < .001; **p < .01; *p < corrected α = .025; *n.s.* p > corrected α = .025.) Error bars represent standard errors of the mean.

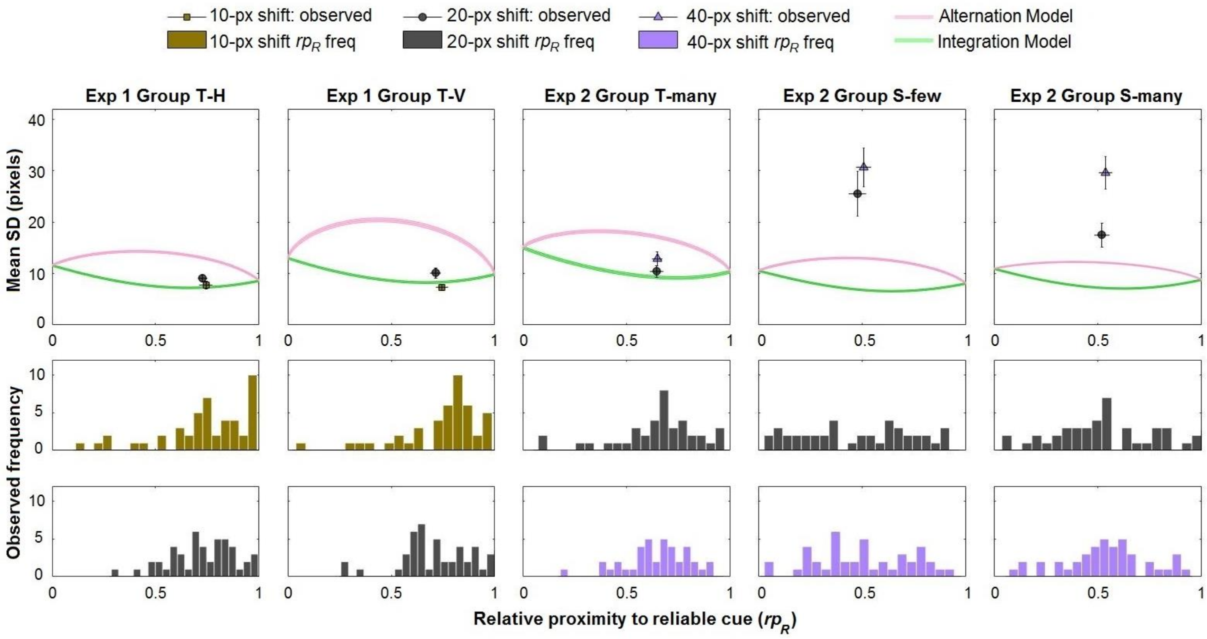

Figure 5. (A) Observed mean SD of signed deviation and predicted SD by computational models on the shift tests. The *x* axes correspond to greater reliance on the reliable landmark, from left to right. Curves represent the means of functions predicting mean SDs (± standard errors of the mean) from different weights of reliable landmark (rp_R) by the Integration (green) or Alternation Models (pink). Points represent observed mean SDs and mean rp_R on shift tests (brown square: 10-px shift; black circle: 20-px shift; purple triangle: 40-px shift). Error bars represent standard errors of the mean. (B) The frequencies of observed rp_R on shift tests (brown: 10-px shift; black: 20-px shift; purple: 40-px shift).

Table 1. *Experimental procedure for each participant group.*

		Training 1	*Training 2*	*Training 3*		*Training 4*	*Testing*		
Group				Feedback	No feedback		Baseline	Probe	Total
Exp 1: T-H and T-V	*Trial type*	*R, U*	*R, U*	*R, U, B*	—	—	*R, U, B*	*R, U, B, 10, 20, r*	
	Goal shown	*yes*	*no*	*no*	—	—	*no*	*no*	
	Feedback	*yes*	*yes*	*yes*	—	—	*yes*	*no*	
	No. of trials	*18*	*18*	*27*	—	—	*135*	*30*	*228*
Exp 2: T-many	*Trial type*	*R, U*	*R, U*	*R, U, B*	*R, U*	*R, U, B*	*R, U, B*	*R, U, B, 20, 40, r*	
	Goal shown	*yes*	*no*	*no*	*no*	*no*	*no*	*no*	
	Feedback	*yes*	*yes*	*yes*	*no*	*yes*	*yes*	*no*	
	No. of trials	*18*	*18*	*18*	*6*	*27*	*135*	*30*	*252*
Exp 2: S-few	*Trial type*	*R, U*	*R, U*	*R, U*	*R, U*	*R, U*	*R, U*	*R, U, B, 20, 40, r*	
	Goal shown	*yes*	*no*	*no*	*no*	*no*	*no*	*no*	
	Feedback	*yes*	*yes*	*yes*	*no*	*yes*	*yes*	*no*	
	No. of trials	*18*	*18*	*18*	*6*	*18*	*90*	*30*	*198*
Exp 2: S-many	*Trial type*	*R, U*	*R, U*	*R, U*	*R, U*	*R, U*	*R, U*	*R, U, B, 20, 40, r*	
	Goal shown	*yes*	*no*	*no*	*no*	*no*	*no*	*no*	
	Feedback	*yes*	*yes*	*yes*	*no*	*yes*	*yes*	*no*	
	No. of trials	*18*	*18*	*18*	*6*	*26*	*136*	*30*	*252*

Note: Trial types are as follows: R, only the reliable cue presented; U, only the unreliable cue presented; B, both cues presented; numbers represent the amount of shift, in pixels; r, both cues presented with positions reversed. In the Experiment 2 group names, “T” refers to training with cues jointly and singly and “S” refers to training with single cues only; “many” and “few” refer to the total number of trials.

Table 2. *Comparisons between observed SD and predicted SD on shift tests according to t tests.*

Model	**Group T-H**		**Group T-V**		**Group T-many**		**Group S-few**		**Group S-many**	
	10	20	10	20	20	40	20	40	20	40
Integration	✓	NC	✓	✓	NC	NC	×	×	×	×
Alternation	×	NC	×	×	NC	NC	×	×	×	×

Note: 10, 20 and 40 denote the 10-, 20- and 40-px shift tests, respectively. ✓ indicates consistency with model prediction; × mark indicates inconsistency with model prediction. NC indicates that the *t* test results were not clear enough to dissociate the two models.

Table 3. *BIC values for each model and shift test for Group T-H, T-V and T-many.*

Model	**Group T-H**		**Group T-V**		**Group T-many**	
	10-px shift	20-px shift	10-px shift	20-px shift	20-px shift	40-px shift
Integration	**879.36**	**825.58**	**1015.85**	**1032.96**	**851.52**	881.03
Alternation	910.36	853.74	1038.39	1047.84	858.54	**878.79**

Note: Bold numbers denote smaller BIC values that support the corresponding model. There is no need to do Bayesian analysis for Group S-few and S-many because the results from *t* test indicate that neither the Integration nor the Alternation Models fit the observed data (see main text for explanation).

Table 4. *Comparisons of SD and weighting* (rp_R) *between observations and optimal (Bayesian) predictions in Both-cue and shift tests.*

Measure	**Group T-H**			**Group T-V**			**Group T-many**			**Group S-few**			**Group S-many**		
	B	10	20	B	10	20	B	20	40	B	20	40	B	20	40
SD	✓	✓	×	✓	✓	×	✓	×	—	×	—	—	×	—	—
Weighting	NA	×	×	NA	×	✓	NA	✓	—	NA	—	—	NA	—	—

Note: B denotes Both-cue test. 10, 20 and 40 denote the 10-, 20- and 40-px shift tests, respectively. ✓ indicates consistency with model prediction; × mark indicates inconsistency with model prediction. — indicates no need to do this comparison based on the results of *t* tests and Bayesian analysis because the observed data did not fit the Integration Model (see main text for explanation). NA, not applicable.