**Supplementary Figure S1**

****

****

In a “fixed dot” task after the main experiment, we presented participants with ­dot-clouds of a fixed sample size (Ndots) and asked them to aim for the middle of the dot-clouds (as in the main experiment). The aim was to measure, on an individual basis, how the probability of hitting the target increased with Ndots. For Blue bars in Supplementary Figure 1a show mean hit rates for each group/condition. Adults were presented with all possible Ndots in the main task (1 to 20, 25 trials per Ndots condition). However, to keep the task child-friendly (i.e., to limit test duration), children were only presented with the Ndots = 2,3,7 and 15 conditions. To interpolate smoothly across remaining Ndots conditions we fitted spline-functions to the data, constrained to 3 knots, concave and increasing in shape, with a minimum of 0 and maximum of 1. The resulting black dotted curves indicate the interpolated hit-probabilities. These curves were used in main analysis as direct measure of hit probability (Fig 2, Main text).

**Theoretical background and rationale:**

The red dashed curves in Figure S1 show predicted hit probabilities for an ideal observer who estimates the mean of the dot-cloud perfectly. The ideal observer’s estimate is an unbiased estimate of the location of the center of the target, whose variance decreases linearly as a function of Ndots:

|  |  |
| --- | --- |
| . | ( **Eq S1** ) |

As detailed previously by Juni, Gureckis & Maloney (2016), the predicted probability that the aiming point will land within the target circle *T* can then be computed by integrating a bivariate Gaussian, centered on the target andwith variance ,across the target circle *T*:

|  |  |
| --- | --- |
| . | ( **Eq S2** ) |

In reality, however, any sensory, cognitive or motor error in the participant’s estimate of the mean of the dot-cloud will increase response error: Eq S2 would then overestimate the participant’s true hit rates. This can be accounted for in the model by adding an additional zero-mean error term, , which represents the additive sum of all possible sources of internal noise, thus:

|  |  |
| --- | --- |
| Screen Clipping, | ( **Eq S3A** ) |
| where: |  |
| Screen Clipping. | ( **Eq S3B** ) |

In principle, one could attempt to measure explicitly (e.g., see *Jones et al, JASA, 2013*). In practice, however, such measurements are non-trivial, and often require the experimenter to make a number of questionable assumptions (e.g., independence between Ndots and ). It was also unnecessary for the present study, as we were only interested in the final, overall amount of error, irrespective of its source. We therefore quantified total response error empirically, by presenting participants with fixed numbers of dot-cues, and computing the variance in observed response error, thus:

|  |  |
| --- | --- |
| , | ( **Eq S4** ) |

where εx and εy are response errors in the horizontal and vertical directions, respectively. Values of σobs were estimated independently for different values of Ndots (i.e., as some sources of internal noise may vary with Ndots), and were estimated independently for each participant (i.e., as the magnitude of internal noise may differ between observers).

Note that σobs incorporates all possible sources of response error, including both internal noise (e.g., motor error, suboptimal integration, etc.), and external noise:

|  |  |
| --- | --- |
| Screen Clipping, | ( **Eq S4** ) |

and by combining this total standard covariance matrix with Eq 3A yields a predicted hit function of:

|  |  |
| --- | --- |
| Screen Clipping. | ( **Eq S5** ) |

In this way, expected hit rates, *p*[hit|Ndots], were adjusted to reflect the performance/abilities of each individual observer. This resulted in more realistic predictions (black dashed line), versus if participants were assumed to be ideal observers (red dashed line).

This analyses allowed us to estimate sampling choices independent of any age differences in σinternal – whilst there were substantial individual differences (see Supplementary Figure 2), age differences were small, as (see Table 1, and the comparable heights of the blue bars plotting hit probability in Supplementary Figure 1) though (marginally) significant (σinternal, high reliability x Age: F4,69=2.17, p= 0.08, σinternal, low reliability x Age: F4,71=3.97, p= 0.006.), demonstrating the importance of measuring and correcting for this factor.

**Supplementary Figure S2**

****



The ideal observer model predicts the expected score for each sampling strategy based on measures obtained in the fixed dot condition (see Supplementary Materials 1). To assess how well this model captures participant’s actual task performance we have plotted in the top graph, the predicted average score per trial (x-axis) against the scores actually obtained (y-axis) on the task. The expected value EV of the participants choice of Ndots is the probability of hitting the target after sampling Ndots times the value remaining:

where M is the initial value of the target.

Data points are color-coded according to variability in the sampling strategy (by the standard deviation of error around the mean Ndots sampled), across all 100 trials, with warmer colors corresponding to more variable sampling / higher standard error. The data points clearly follow the identity line, suggesting that the ideal observer prediction captures task performance well. Scores were slightly lower than predicted (data points fall below the identify line) for individuals who sampled more variably across the experiment (depicted in warmer colors), in line with the prediction for a suboptimal sampler indicated by a triangle with error-bar in Figure 2 of the main text.

The bottom two plots show σinternal, the mean standard deviation of aiming locations from the center of the dot cloud measured in the fixed Ndots condition (x-axis) against the same measure obtained in the free Ndots condition (y-axis), averaging the measures across Ndots sample sizes between 2 and 15. The data closely follow the identity line. This suggests that participants were using consistent sampling strategies across the condition on which we based the ideal observer model, and the main sampling task in which we then used this model to predict performance.

**Supplementary Figure S3**

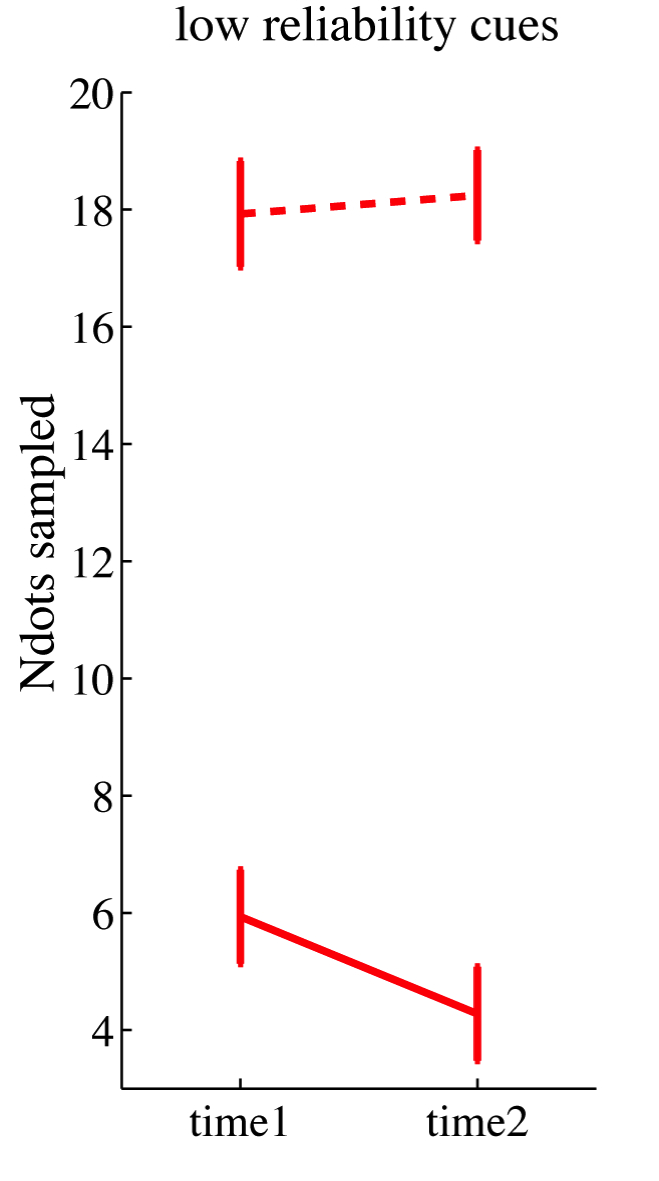


Figure S3. Number of dots sampled during the first (time1) and last (time2) 15 trials of the low reliability cue condition. Solid lines are from the main experiment where dot-cues cost 1 point each (for gain-maximizing number of dots for each age group, see Table 1 of Main Manuscript). The dotted line reflects the numbers of dots 6 to 7-year-olds sampled when additional cues came at zero cost. In the zero cost condition, 6 to 7-year-olds sampled significantly more information, made many button presses, endured longer trials, and did not exhibit signs of fatigue over the course of the experiment. It follows, therefore, that the sampling choices of young children in the main-experiment cannot be explained simply by (i) implicit sampling costs, such as fatigue or boredom, (ii) an unwillingness to sample more than 4 or 5 dots, or (iii) misunderstanding that sampling more dots increases the probability of successfully touching the hidden target.

**Supplementary Figure S4**



Figure S4 shows the mean number of dots sampled following a string of correct/incorrect responses on the last N trials (1-3 respectively), compared to the mean Ndots sampled in the K trials preceding these N trials (we plot K=5, but other values give similar results). The top panel shows data for High Reliability Cues, the lower panel shows data for Low Reliability Cues. The change in the number of dots sampled after N (1-3) consecutive hits (green) or misses (red), relative to the mean Ndots in the preceding 5 trials, is plotted per age group. Error bars indicate bootstrapped 95%CIs. Delta N*dots*=1 means that on average 1 more dot was sampled on a trial following N correct/incorrect responses As can be seen in the Figure, adults and older children tended to increase their sampling after a series of misses (>1) and decrease their sampling after a series of hits, showing that even though their sampling strategies were closer to the ideal location than in childhood from early on in the task, trial-to-trial feedback did inform their sampling choices. The youngest children, in contrast, did not adjust their sampling significantly based on previous hits and misses.

**Supplementary Materials S5**



Through simulations, we tested if participants’ sampling decisions at any age would yield the maximum expected score given a noisy estimate of the spread of aiming points around the target (σobs, SD > 0). To model the ideal sampling strategy under this scenario, we assumed that the observer’s estimate of σobs took the form of a distribution with an unbiased mean σobs, mu and normally distributed error σobs, SD (range: 0, 2, 4, or 6 for reliable cues, 0, 4, 8, or 12 for unreliable cues). To constrain σobs, SD to be positive and symmetrical, we truncated the distribution at 2 x σobs, SD. As in the main task, σdots was set to 12.5mm or 27.5mm for reliable vs. unreliable cue conditions respectively, and σobs was set to 4mm, approximating empirical values measured in the fixed dot condition (Table 1 of Main Manuscript). We then computed the average ideal strategy, by identifying the Ndots with the highest expected gain on average across 10000 trials, with σobs drawn randomly from its truncated normal distribution (mu=σobs, mu, sigma=σobs, SD) on each trial. Results of this analysis are shown in the figure below. The dotted line indicates the sampling strategy that would maximizes score across a large number of trials, given error in the estimate of σobs. Green and red curves indicate the distribution of probability curves and expected gain-landscapes across trials given a σobs with 13.1 (left) and 27.7 (right) drawn from a truncated normal distribution with SD = 0, 4, 8, 12, (left) and SD = 0, 2, 4, 6 (right). Thick black dotted lines indicate the average gain-landscape and its peak. The graphs show that an observer who optimally accounts for error around the estimate of σobs (graphs in bottom three rows) should sample fewer dots than an ideal observer with a perfect estimate of σobs to maximise score (shown in top row). Importantly, however, the reduction is only small and does not approach childhood sampling behaviour of sampling (indicated by the pink dotted lines), even given a very large amount of noise in the estimate of σobs (up to 0.5 x σdots).

**Supplementary Materials S6**



Observers taking part in the fish-catching task may estimate their overall response uncertainty (σobs) in a straightforward manner, by keeping track of deviations between their location estimates and the target (feedback about both are provided simultaneously), and computing their standard deviation across a number of trials . When computing SD across a sufficiently large number of trials, the estimated value of will approach the true underlying value, σobs. However, if only a few previous trials are considered (i.e., due to limited memory capacity), σobs will tend to be systematically underestimated.

In a next set of simulations, we therefore tested if child sampling behaviour could be described as optimal, assuming that the estimate contained a bias that could arise from forming an estimate of σobs based on a small numbers of previous trials.

For these simulations, we set σdots to 27.5mm (focussing on the unreliable cue condition), and σinternal to 4mm. For each Ndots, we then simulated the distribution of distances between the aiming point and target, given an unbiased pointer (mean = 0) with SD = σobs. We then randomly sampled Ntrials (range 2-40) from this distribution and computed the standard deviation. To obtain the expected we took the average SD across 10.000 of these samples. We then computed the corresponding hit probability and expected score for each Ndots to identify the gain-maximising Ndots.

Figure S5 shows the result of simulations in which σobs was estimated based on 2, 3, 5 or 40 of the previously observed trials. As is clear from the figure, considering only a few trials (2, 3, or 5) to compute σobs leads to a systematic underestimation of uncertainty in the location estimate. An observer who computes the ideal strategy based on this underestimated σobs would sample less than they truly should to maximise their score (namely 7.5 Ndots). However, even for the most extreme case in which σobs is computed across a very small number of trials, the deviation from the ideal observer model with an accurate estimate of σobs is small (for , Ndots, ideal=6.9). In other words, whilst a strategy of tracking spread of aiming points around the target with a limited memory would lead to under-sampling, this factor alone cannot fully explain the observed age differences in sampling.

When estimating error between target and aiming points based on only a small subset of trials (say, the last two observed) the resulting variance estimate may fluctuate extensively from trial to trial due to variability in the sample mean location around the true mean location (the target). Can a combination of bias in the estimate of visual uncertainty and error around this estimate explain child performance? When estimating error around the target based on only two trials on average (top panel of figure), this is equivalent to a scenario in which σdots would be ~21mm (note the true SD was ~27.5 mm). When this “plausible” bias is added to the simulation of error around the hit probability estimate in supplementary Figure 4, the maximum symmetrical error on σobs that we can simulate is SD=10mm. In this most extreme plausible case (σobs= = = 21.4), the ideal dot sample size is significantly lower than for an ideal observer with perfect knowledge of visual uncertainty (σobs= 27.8,SD=0). However, whilst the ideal observer prediction accounting for extreme bias and error around the estimate of hit probability comes close to the Ndots sampled by the youngest age group, it is still higher (optimal dots to sample is ~5.6 dots, while the youngest children sample 4.6 dots on average).

**Supplementary Table S1.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 6-7 vs 8-9 | 8-9 vs 10-11 | 10-11 vs 13-15 | 13-15 vs adult |
| **Absolute Deviation from Ndotsideal - Fig 3a** | | | | |
| High cue Reliability | 0.56 | 0.027\* | 0.45 | 0.79 |
| Low cue  Reliability | 0.43 | 0.037 \* | 0.99 | 0.54 |
| **Score efficiency - Fig 3b** | | | | |
| High cue Reliability | 0.53 | 0.67 | 0.02 | 0.81 |
| **Signed Deviation from Ndotsideal - Fig 5a** | | | | |
| Low cue  Reliability | 0.044\* | 0.69 | 0.082\* | 0.95 |
| **Stand Dev around Ndotsideal - Fig 5b** | | | | |
| High cue Reliability | 0.39 | 0.023\* | 0.18 | 0.26 |

Table S1 reports p-values resulting from post-hoc independent-sample t-tests that compare age differences between consecutive age groups. Black stars indicate p<0.05, red star indicates a non-significant trend of p<0.1. In Figures 3 and 5 in the main manuscript, significant group differences are indicated, and can also be derived from overlap in confidence intervals. Specifically, a CI overlap of less than ~25% indicates a significant difference at p<0.05 (Cumming & Finch, 2005). These post-hoc tests reveal that the efficiency of visual sampling on the current task as quantified by these 4 measures, is adult-like from roughly age 10-11 years onwards.

Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170.