Ideal observer analysis for continuous tracking experiments

Pierfrancesco Ambrosi
Department of Neuroscience, Psychology, Pharmacology and Child Health, University of Florence, Florence, Italy

David Charles Burr
Department of Neuroscience, Psychology, Pharmacology and Child Health, University of Florence, Florence, Italy
Institute of Neuroscience, National Research Council, Pisa, Italy
School of Psychology, University of Sydney, Sydney, New South Wales, Australia

Guido Marco Cicchini
Institute of Neuroscience, National Research Council, Pisa, Italy

Continuous tracking is a newly developed technique that allows fast and efficient data acquisition by asking participants to "track" a stimulus varying in some property (usually position in space). Tracking is a promising paradigm for the investigation of dynamic features of perception and could be particularly well suited for testing ecologically relevant situations difficult to study with classical psychophysical paradigms. The high rate of data collection may be useful in studies on clinical populations and children, who are unable to undergo long testing sessions. In this study, we designed tracking experiments with two novel stimulus features, numerosity and size, proving the feasibility of the technique outside standard object tracking. We went on to develop an ideal observer model that characterizes the results in terms of efficiency of conversion of stimulus strength into responses, and identification of early and late noise sources. Our ideal observer closely modeled results from human participants, providing a generalized framework for the interpretation of tracking data. The proposed model allows to use the tracking paradigm in various perceptual domains, and to study the divergence of human participants from ideal behavior.

Introduction

Experiments in psychophysics are generally time-demanding and mostly concerned with estimation of perceptual thresholds from discrete trial presentations. These methods can be difficult to apply in ecologically relevant situations (Huk, Bonnen, & He, 2018), which are important to address neural processes underlying behavior, and for understanding inferences from image statistics for natural perceptual function (Geisler & Ringach, 2009; Kersten, Mamassian, & Yuille, 2004). Continuous paradigms may provide an important tool for the investigation of behavior in real-life situations and to assess dynamic features of the perceptual system under study. One such example is "continuous tracking," where participants are asked to track a target with a cursor, and cross-correlogram analysis between responses and stimuli captures participants' ability to localize a target. Performance improves with the signal-to-noise ratio (SNR) of the stimulus (Bhat, Cicchini, & Burr, 2018; Bonnen, Burge, Yates, Pillow, & Cormack, 2015; Bonnen, Huk, & Cormack, 2017; Li, Sweet, & Stone, 2005; Mulligan, 2002; Mulligan, Stevenson, & Cormack, 2013), consistent with the idea that if a stimulus is sufficiently perceivable for participants to answer psychophysical questions about it, then they can accurately point to its position. The strength of correlation between target position and its tracking by participants is a predictor of standard psychophysical thresholds (Bonnen et al., 2015). More importantly, this technique allows for the collection of a large amount of data in a short time.

Data from tracking experiments, however, can be difficult to interpret. Typically, cross correlograms spread over long time windows, and have a temporal lag of a few hundred milliseconds. This means that the target and response positions cannot be directly compared without introducing spurious error estimates due to the lagged response. Such lag cannot be merely included as a temporal shift, because the spread of the

Citation: Ambrosi, P., Burr, D. C., & Cicchini, G. M. (2022). Ideal observer analysis for continuous tracking experiments. *Journal of Vision*, *22*(2):3, 1–16, https://doi.org/10.1167/jov.22.2.3.

https://doi.org/10.1167/jov.22.2.3 Received June 4, 2021; published February 1, 2022 ISSN 1534-7362

cross-correlogram through time suggests that subjects integrate across various frames. These limitations need to be addressed to make tracking a reliable technique in perceptual studies.

In this work, we address quantitatively the question of how much of the tracking performance is due to sensory uncertainty or to noise occurring in motor systems. We did so by measuring two perceptual attributes that have not been addressed with tracking experiments to date, yet have a long history of psychophysical studies – numerosity and size. Numerosity perception refers to the ability to estimate the number of objects in a scene without serial counting, using what has been termed the "approximate number system" (ANS; Dehaene, 2011). This ability is shared with many animal species (Nieder, 2019), indicating that it may give an ecological and evolutionary advantage, such as choosing zones with more food, or quickly determining which group of competitors is more numerous. Many works have investigated this system using the many paradigms offered by psychophysical research, which have highlighted important features of this neural system, such as its susceptibility to adaptation (Arrighi, Togoli, & Burr, 2014; Burr & Ross, 2008; Castaldi, Aagten-Murphy, Tosetti, Burr, & Morrone, 2016), its partial independency from continuous magnitudes (Anobile, Cicchini, & Burr, 2014, Anobile, Cicchini, & Burr, 2016; Cicchini, Anobile, & Burr, 2016; Ross & Burr, 2010), and that it acts on segmented objects (Anobile, Cicchini, Pomè, & Burr, 2017; Franconeri, Bemis, & Alvarez, 2009; He, Zhang, Zhou, & Chen, 2009) rather than on texture. Size is also a primary perceptual attribute that guides human and animal behavior. Studies on perception of object size have shown that it has a topographical representation in the parietal cortex (Harvey, Fracasso, Petridou, & Dumoulin, 2015), whose perceived magnitude modulates neural activity in the primary visual cortex (Pooresmaeili, Arrighi, Biagi, & Morrone, 2013). It is susceptible to serial dependence (Kristensen, Fracasso, Dumoulin, Almeida, & Harvey, 2021), adaptation, independently from spatial frequency (Baker & Meese, 2012), and attention (Tonelli, Pooresmaeili, & Arrighi, 2020). The two perceptual features, which share a Weber-law behavior but arise from independent perceptual mechanisms (Anobile et al., 2014; Ganel, Chajut, & Algom, 2008), allowed us to design equivalent experimental paradigms relying on different perceptual systems.

Participants were presented with a cloud of dots changing in numerosity or area (in separate sessions) and asked to compensate for the changes with mouse movements, to keep the numerosity or area of the stimulus constant. This annulling paradigm provides dynamic information on participants' responses, which can be extracted from the cross-correlogram: its peak, temporal lag, and width (Bhat et al., 2018; Bonnen et al., 2015; Bonnen et al., 2017; Mulligan, 2002; Mulligan et al., 2013). In addition, it was possible to measure the effect of increased visibility of changes in the stimuli as the increase in the square root of the explained variance between participants' responses and those predicted by a linear virtual observer, and we labeled this measure efficiency (Barlow, 1962; Pelli, 1991a; Pelli & Farell, 1999). In other words, our design allowed us to compare participants' performances to those of a noiseless observer, which shares the participants' motor response to a single change in the stimulus, and to measure the overall visibility as the correlation between the real and ideal response. This provides evidence for the existence of a linearly scaling range of performance, where participants' behavior in tracking can be better understood in terms of signal strength.

In order to test whether deviations from ideal behavior arise from random fluctuations in mouse movements or from how stimuli are converted into responses, we designed a *Psychometric Observer* that combines dynamic information from the annulling task and discrimination thresholds estimated from a standard two-alternative forced choice (2AFC) experiment. The result is a virtual observer that responds probabilistically to incoming changes in the stimulus, with a probability determined by participants' perceptual sensitivity. Importantly, our Psychometric Observer closely replicated the improvements of performance with stimulus visibility, indicating how visual characteristics of the stimuli may affect human performance in the tested range. This suggests that results obtained through continuous tracking are mostly related to how information is processed by the perceptual system, independently from the perceptual feature under study, making it a valuable technique to apply to various perceptual domains, yielding temporal and dynamic information in a fast and spontaneous way.

Methods

Participants

Nine voluntary participants (ages 24–35 years, 5 women) were recruited. All had normal or corrected-to-normal vision. All participants had prior experience in psychophysical experiments, only one of them had prior knowledge about the details of the experiment (one of the authors). All were right-handed and used their right hand for tracking. Experimental procedures are in line with the Declaration of Helsinki and were approved by the regional ethics committee (Comitato Etico Pediatrico Regionale — Azienda Ospedaliero-Universitaria Meyer — Firenze, FI). Written informed consent was obtained from each

participant, which included consent to process and preserve the data and publish them in anonymous form.

Setup

Stimuli were displayed on a 70 × 40 cm Display++ LCD Monitor (Cambridge Research Systems, Rochester, UK) with resolution = 1920 × 1080 pixels and refresh rate = 120 Hz. A regular USB mouse was used to collect responses in the tracking experiment, and a regular USB keyboard in the 2AFC experiment. In all experiments, participants were placed at a 57 cm distance from the screen.

Stimuli were generated by the Psychophysics Toolbox (Brainard, 1997; Kleiner, Brainard, Pelli, & Ingling, 2007; Pelli, 1997). The target was a cloud of dots of random sizes (from 0.3 degrees to 0.5 degrees diameter) on a uniform gray background. Each dot had a maximum presentation time (lifetime) of 100 ms. A limited lifetime paradigm was adopted to avoid participants performing the task by ignoring the dots always present on the screen and responding merely to the appearing or disappearing of a few dots. In the numerosity tracking task, the number of dots forming the cloud changed. The starting patch contained 20 dots arranged in a cloud of 200 cm^2 (radius approximately 8 degrees). Numerosity of the patch was free to change from numerosity 2 to 100. When pooled across all trials, the resulting numerosities were Gamma distributed, with $k = 2.2$, $\theta = 23.4$. The area task was analogous to the numerosity task, except that the area (convex hull) of the cloud of dots was varied, whereas the numerosity was kept constant. Stimulus size was free to vary between approximately 35 cm^2 and approximately 800 cm^2 (radius between approximately 3 degrees and approximately 16 degrees). The sizes of the patch of dots presented on screen were also Gamma distributed, with $k = 2.0$ and $\theta = 203.0$ cm^2.

Tracking

The numerosity of the stimulus changed at random time intervals and they followed a Gamma distribution with $k = 1.1$ and $\theta = 0.228$ seconds. Five different experimental conditions with increasing relative changes were tested. In each condition, changes had two possible strength values in a ratio 1 to 2.5: "weak" were 0.050, 0.075, 0.100, 0.150, and 0.200 octaves, and "strong" 0.125, 0.187, 0.250, 0.375, and 0.500 octaves. We will refer to each condition with a label corresponding to the average change undergone by the stimulus in that condition. Condition properties are summarized in Table 1. Movie 1 gives a demonstration of the stimuli in absence of mouse movements, and movie 2 shows an example of the effect of mouse movements on stimuli.

Condition label	Weak change (octaves)	Strong change (octaves)	Average change (octaves)
1	0.050	0.125	0.087
2	0.075	0.187	0.131
3	0.100	0.250	0.175
4	0.150	0.375	0.262
5	0.200	0.500	0.350

Table 1. Stimulus changes in each condition. Stimuli changed at random instants, with equally distributed positive and negative changes, and equally distributed weak and strong changes. Changes are expressed as the ratio between the new stimulus value (numerosity or area) and the previous, expressed in octaves. Each condition is identified by the average between these values or by its label.

Each of the five conditions comprised 18 blocks of 20 seconds each (6 minutes). The number of changes in the positive and negative directions was equal in every trial. The task of participants was to move the mouse leftward or rightward to counteract stimulus changes: rightward movements increased numerosity, whereas leftward movements decreased it. Movements of the mouse were rescaled to compensate for the greater changes undergone by the stimulus in different conditions, so that participants would perform the task with the same motions. This was done to prevent the enhancement of motor errors, which increase in variance with larger movements (Fitts, 1954; Harris & Wolpert, 1998). The same holds for the area task.

Alternative forced choice

To measure psychophysical sensitivity within a standard 2AFC framework we ran a 2AFC experiment slightly differently from the standard implementation: trials were made from clips from the tracking experiment, with dots of limited lifetime, each present on the screen for a maximum of 100 ms, as in the tracking experiment. In separate experiments, the cloud of dots changed in either area or numerosity. Each observer was presented with 200 clips, each 1000 ms long: for the first 500 ms numerosity (or area) was constant (the reference), for the second 500 ms either lower or higher (the probe). Participants indicated by keypress whether the numerosity appeared to increase or decrease. The color of the dots changed from blue to yellow to flag the two intervals. Participants were allowed 2 seconds to respond before the presentation of the new clip. Each clip showed changes ranging from −1 to 0.6 octaves (10 to 30 dots). The area task had the

same architecture, also showing changes ranging from −1 to 0.6 octaves, with the probe comprising 20 dots of approximately 200 cm^2 (approximately 8 degrees in radius). These differences from the standard 2AFC paradigm were necessary to have a realistic estimate of participants' discrimination thresholds in the tracking conditions.

Data analysis

Data analysis was implemented on Matlab 2019a (MathWorks, Natick, MA, USA) and on JASP (version 0.14.1, JASP Team, 2020; jasp-stats.org).

Tracking

We measured the normalized cross-correlation between changes in mouse position (output) and computer-driven changes in the stimuli (input). Because of the annulling strategy, mouse movements were reversed to match the direction of changes on screen to yield positive rather than negative correlations. The maximum correlation value (peak), the full-width-half-max (width), and delay at maximum value (lag) are taken as parameters for participants' performance in each condition.

Psychophysics

Data from each participant in the 2AFC tasks were analyzed separately. For each task (area or numerosity), occurrence of reporting an increase in the stimulus was plotted as a function of changes in the stimulus, expressed in octaves, and fitted with a cumulative Gaussian distribution. The width of the underlying Gaussian divided by its mean gives the participant's Weber Fraction (WF).

Efficiency

The cross-correlation between participant response and input stimuli in the tracking tasks was used to estimate the transfer function kernel from input stimuli to output response (see Appendix A for details). This was used to generate an ideal response in each condition, which was interpreted as an ideal observer responding to every change in the stimulus, limited by the motor implementation of the participant. In this sense, comparison of human and ideal performance yielded an estimate of the total amount of signal converted into a response by the subject. We termed this measure of the explained variance between the two time-series efficiency (Barlow, 1962; Pelli, 1991a; Pelli & Farell, 1999). Efficiency was measured as the correlation between the real and ideal response, which is an estimate of the explained variance that preserves the information on the timing and direction of stimulus variation (see Appendix B). The kernel was computed from the whole dataset for each participant (see Discussion below). We also implemented a cross-validation analysis of efficiency, where the kernel was computed using half the data (3 minutes per condition, for a total of 15 minutes), and results were compared to the other half of the data. In addition, a bootstrap analysis was implemented combining different trial blocks to build several different 1-minute-long sessions. The session length was chosen as it provided the best trade-off between the number of possible combinations without repetition and statistical significance of the session. We then evaluated the average sensitivity index (d') across different conditions for efficiency and peak of the cross-correlogram, which is generally the best predictor for the quality of participants' responses among the cross-correlogram parameters (Bhat et al., 2018).

Psychometric observer

We implemented a Psychometric Observer to relate tracking and standard psychophysics, and to address possible sources of noise explaining the discrepancies between real tracking and the Psychometric Observer's tracking. The Psychometric Observer is an ideal observer that first interprets incoming sensory information probabilistically and then implements responses via the participant's kernel. The probabilistic behavior in the first front-end stage of processing has been inferred from the subjective responses in the 2AFC discrimination task: the psychometric curve in this task reveals how much a given stimulus change is likely to be interpreted as an increase or a decrease in the target feature, representing the best possible guess of the observer under conditions of no time pressure. Performance of this Psychometric Observer can thus be used as a reference for participants' behavior, as it incorporates the front-end stage of early sensory processing and constitutes a reasonable benchmark for actual observer performance. First inspection of the data revealed how the real observers have performances (i.e. efficiencies) below those of the ideal Psychometric Observer. We thus added noise in various stages to determine which noise source is mostly accountable for the resulting behavior of real observers. According to our modeling, there are two possible stages where noise may arise: a late stage affecting the motor implementation of mouse movements and an early stage that hinders the perception of stimulus changes (Pelli, 1991b). The performances of the Psychometric Observer were averaged across 30 independent

simulations, to reduce fluctuations on the effects of noise.

Results

We designed two tracking experiments, where participants had to continuously counteract variations in the numerosity or size of a cloud of dots shown on screen. First, we verified that tracking performance in the two tasks increased when the stimulus underwent larger changes, that is, when SNR was increased, and compared performance to a standard psychophysical task. Then, we attempted to express tracking performances as a comparison with an ideal observer, that is, a noise-free virtual observer. Finally, we modified the ideal observer into a "psychometric observer" in order to address whether performance was more affected by random movements in the mouse, or by noise in stimulus detection.

Tracking

Participants viewed a dynamic dot display whose numerosity changed smoothly, following a random walk, and annulled the changes in numerosity by moving a computer mouse appropriately (see Methods). An example of the resulting tracks is shown in Figure 1A. The continuous black trace shows the random-walk in numerosity attempted by the computer, and the red trace shows the mouse movements of the participant attempting to annul the changes in numerosity. The actual numerosity on screen (given by the combination of the other two traces) is shown by the dotted trace.

Figure 1B shows the cross-correlation between stimulus changes (input) and mouse movements (output) for five different levels of signal to noise. Because the experiment required annulling of changes in the stimulus, mouse movements were reversed in sign to yield positive correlation when movements were made in the correct direction. For all levels of signal strength, the cross-correlograms have a clear positive peak, with a delay of 600 to 1000 ms. In both tasks, the amplitudes and latencies of the peaks clearly depend on SNR. Figure 1C plots the average peak, lag, and width across participants as a function of the average change. Like Bonnen et al. (2015), we find a significant dependence on the average change for the peak, width, and lag of the cross-correlogram. Results are summarized in Table 2. Peak of the cross-correlogram results the best predictor for the SNR of the stimulus.

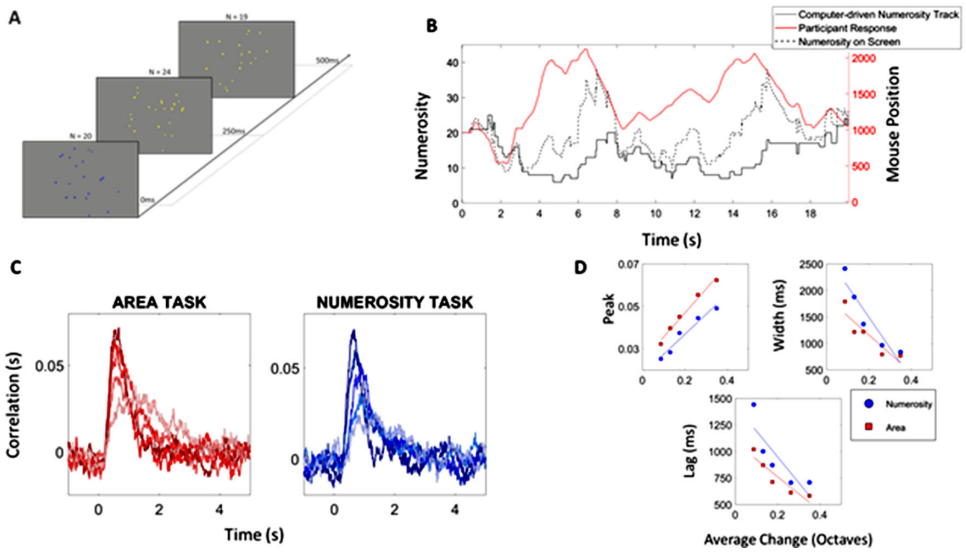

Figure 1. **Results of the tracking experiments.** (**A**) Example stimuli for the numerosity task. (**B**) Example of a 20 second trial of numerosity tracking. The black solid line represents the evolution of the stimulus numerosity if no mouse movements were made by the participant, and the red line the mouse position. The black dashed line is the actual numerosity presented on screen, which is the result of the sum of mouse movements and computer driven changes in numerosity. The numerosity changes independent of mouse movements were generated according to a pseudo random walk (see Methods) hence the changes were uncorrelated from time to time and served as a basis for deriving cross-correlations. (**C**) Cross-correlations in Area and Numerosity annulling tasks for various levels of signal strength. The left panel shows results for the area task for a representative participant, right panel results for the numerosity task for the same participant. The color scale represents the different conditions: lighter colors higher SNR, larger changes in the stimuli. (**D**) Linear trend of the average across participants for the three parameters (peak, width, and lag) for the two tasks (red for area and blue for numerosity) as a function of changes in the stimuli. The three measures are all significantly correlated with average changes. Results are shown in Table 2.

		Average change		Weber fraction	
Feature	Task	Correlation	Log_{10} Bayes factor	Correlation	Log_{10} Bayes factor
Peak	**Area**	$r = 0.99$ $p < 10^{-3}$	1.3	$r = -0.75$ $p < 10^{-3}$	2.0
	Numerosity	$r = 0.97$ $p = 0.005$	0.9		
Width	**Area**	$r = -0.90$ $p = 0.003$	0.5	$r = 0.49$ $p = 0.04$	0.3
	Numerosity	$r = -0.93$ $p = 0.02$	0.6		
Lag	**Area**	$r = -0.92$ $p = 0.02$	0.6	$r = 0.20$ $p = 0.41$	−0.4
	Numerosity	$r = -0.84$ $p = 0.07$	0.3		
Efficiency	**Area**	$r = 0.97$ $p = 0.007$	0.9	$r = -0.65$ $p = 0.003$	1.2
	Numerosity	$r = 0.93$ $p = 0.02$	0.6		

Table 2. Summary of dependencies of cross-correlogram parameters (peak, lag, and width) and efficiency as a function of the average change in the stimulus, separately for the two tasks, and as a function of Weber Fraction.

Two alternative forced choice

To compare the results of the annulling paradigm with more standard psychophysics, participants performed two 2AFC discrimination tasks, one for area and one for numerosity (see Methods for details). Points of subjective equality (PSEs) are close to the expected null value (mean $PSE_{AREA} = -0.02$ octaves; mean $PSE_{NUM} = -0.11$ octaves). The WFs, given by the slope of the psychometric functions, are of more interest to compare with the tracking results. Figure 2B shows the WFs for all participants in the two tasks, showing that WFs for the area task were significantly lower than those for the numerosity task, confirmed by a two-sample *t*-test ($t = 3.4$, $p < 0.01$).

Figure 2C shows participant-by-participant averages of the cross correlogram parameters as a function of the WF. The height of the peak correlated strongly and significantly with WFs ($r = -0.75$, $p < 10^{-3}$, and $\log_{10}BF = 2.0$), suggesting that it is the parameter most closely related to numerosity sensitivity, as measured by standard psychophysical techniques. The width of the correlograms is also correlated to the WFs, but the correlation strength was weaker, and the Bayes factor was low ($r = 0.49$, $p = 0.04$; and $\log_{10}BF = 0.3$). The lag of the correlograms did not correlate with WF ($r = 0.2$, $p = 0.4$, and $\log_{10}BF = -0.4$).

Comparison with ideal observer

Given that the two paradigms yielded results consistent with each other, we examined how well participants can be modeled by a linear observer. We developed the Linear Virtual Observers, which behave as linear operators with a filtering kernel given by participant impulse response functions, and compared responses of this ideal, noise-free observer with human participants. The correlation between the two response sequences yields an index of similarity of the actual observer to the noise free observer and thus acts as an index of efficiency (Barlow, 1962; Pelli, 1991a; Pelli & Farell, 1999).

Figure 3 shows how the virtual observer is built and how it is compared to the real observer. The human observer is approximated by a linear model, which receives changes in the stimulus as input (S) and converts these inputs into an output (R), the changes in mouse position. Cross correlating the input and the output across many different trials provides an estimate of the transfer function (K) of the Linear Observer (see Appendix A for details). The resulting transfer function can be used to implement an ideal observer that responds to each change in the stimulus, generating an ideal response (R') for the same inputs received by the human observer.

Figure 3B shows an example of the results for a single 20 second trial. The input S (in black) is a series of instantaneous changes occurring at random intervals, the blue curve the response R, given by the velocity of mouse the movements, and the red curve the response of the ideal observer R'. The two curves are inverted in the figure because the paradigm requires annulling of changes, so positive changes in the stimulus should induce negative changes in mouse position, and the latter were inverted for both the

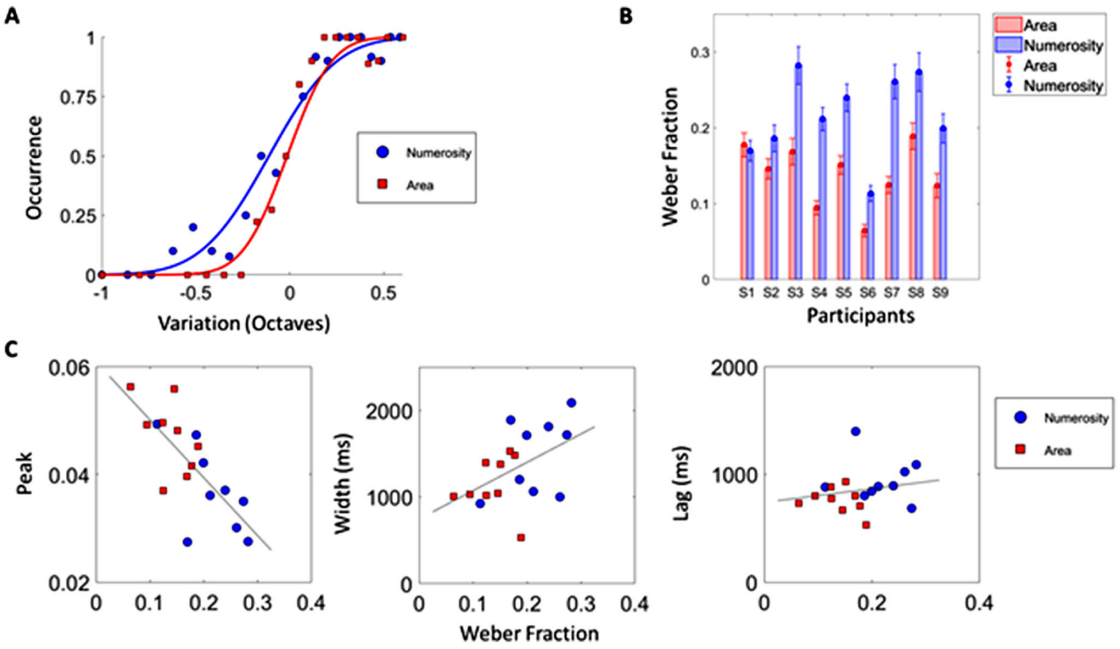

Figure 2. **Correlation between tracking and traditional 2AFC psychophysics.** (**A**) Psychometric functions for the two 2AFC tasks, combining data from all participants. Proportion of time the participant reported an increase of the stimulus plotted as a function of the variation in the presented stimulus, on logarithmic scale. Red squares show responses in the area task and blue circles the numerosity task. Color-coded lines show the best cumulative gaussian fit. (**B**) WFs in the two 2AFC tasks (red area and blue numerosity) computed separately for different participants. Errors are the standard errors of the WFs. (**C**) Participant-by-participant average values of peak, width, and lag of the cross correlogram plotted as a function of the Weber Fraction. The grey lines are the result of linear fits of the data.

human and ideal observers to ease the visualization. Correlating the ideal response R' and the real response R yields an estimate of the efficiency of human observer in transforming inputs into outputs (see Appendix B). Note that the two tracks are generally similar but can diverge when the human observer processes the inputs incorrectly.

Figure 3C shows the numerosities implied by the previous data (integration of the curves in Figure 3B). The black track is the random walk attempted by the computer, which is closely reproduced by the red curve representing the ideal observer, save for a delay introduced by the estimated transfer function. The blue curve represents the actual mouse movements of the participant. The ideal observer is a good predictor of mouse movements but reveals that participants did not always follow the target track correctly. The discrepancy between the two observers shows how accurately the participant detected changes in the stimulus on that particular trial.

Figure 4A shows efficiency as a function of the average change in the two tasks. As for peak correlation values, a clear trend is visible as the SNR increases. This holds for both tasks (area: r = 0.97, $p = 0.007$, $\log_{10}$BF = 0.9 and numerosity: r = 0.93, $p = 0.02$, $\log_{10}$BF = 0.6).

Average change (octaves)	Weber fraction (p value)	Area versus numerosity (p value)
0.087	−0.54 ± 0.01 (0.02)	0.48 (0.2)
0.13	−0.69 ± 0.01 (0.001)	0.55 (0.1)
0.175	−0.61 ± 0.01 (0.007)	0.78 (0.01)
0.26	−0.64 ± 0.01 (0.004)	0.73 (0.03)
0.35	−0.48 ± 0.01 (0.04)	0.75 (0.02)

Table 3. Results from correlating the efficiency with Weber Fractions, aggregating the two tasksLast column shows correlation between the efficiencies in the two tasks in different conditions.

Figure 4B shows the efficiency across conditions as a function of the WF. The two measures are significantly correlated in all conditions (Table 3). The results shown in Figures 4A and 4B can be interpreted as efficiency being an effective measure of tracking performances, because it is related to both perceptual sensitivity and stimulus SNR.

Figure 4C shows in blue how these correlations progress across conditions (expressed in absolute value to help visualization). Correlations between efficiency and WFs follow and inverted U-shape pattern with

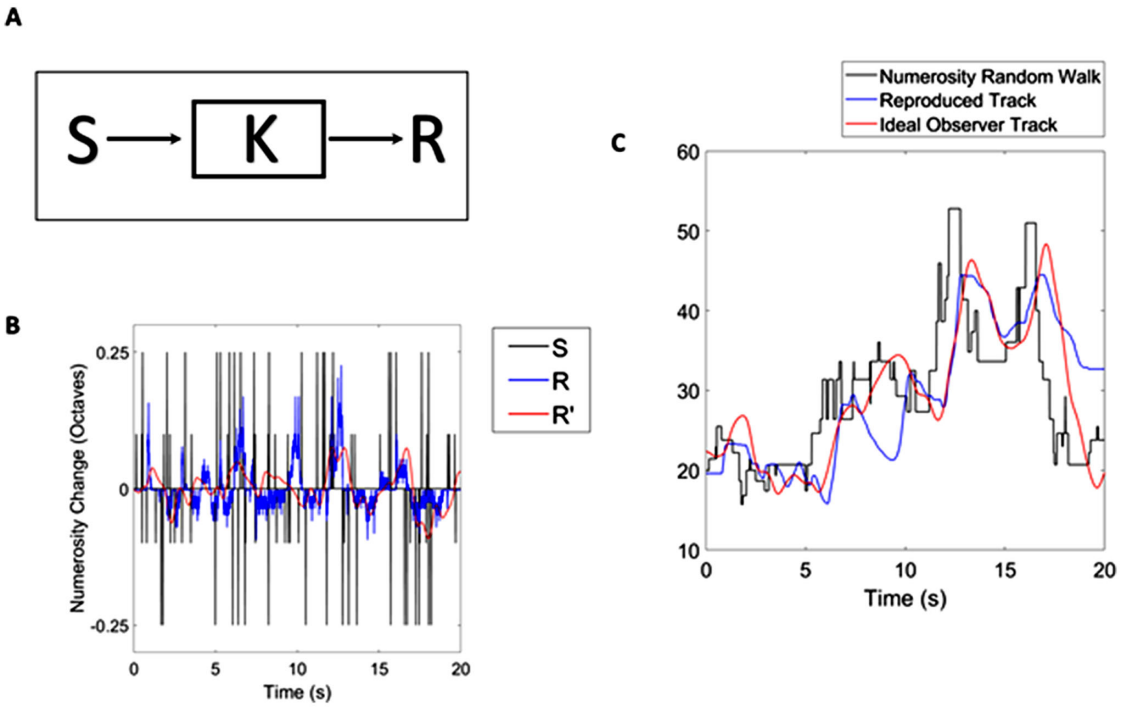

Figure 3. **Ideal Observer Model.** (**A**) Participants are modeled as a linear observer that receives changes in the stimulus (S) as input and produces responses (R) through a transfer function (K). (**B**) Example of the comparison between responses by a human participant and the ideal observer. The black curve represents changes in the stimulus, which were received equally by the human participant and the ideal observer. The blue and red curve represent respectively the participant's and ideal observer responses inverted in sign, because the used paradigm required annulling of changes in the stimulus. The two tracks were rescaled to be compared with the input by normalizing their standard deviations to the standard deviation of the input. Correlating the blue and red curve gives an estimate of the similarity between the two tracks. (**C**) Tracks resulting by the integration of the curves in panel **B**. The ideal observer (red) follows the stimulus random walk more closely than the human observer (blue).

lower values at the extreme SNR conditions (with average change in the stimulus of 0.087 and 0.35 octaves, respectively).

This suggests that despite interindividual differences, there is an optimal range where participant tracking is mostly related to perceptual abilities. This range appears to be in the proximity of the discrimination threshold, whose average across participants in the two tasks is shown as a vertical black dashed line, and the shaded area represents the standard error of the mean. Figure 4C also plots the correlation between the two measures (grey diamonds). Correlations between the two efficiencies follow monotonically the SNR of the stimuli. For conditions one and two, these correlations are low, and then become stronger and stronger. One possibility is that, at low SNRs, the probability of detecting the stimulus is highly dependent on the perceptual channel involved, so the two measures are not correlated. Indeed, the correlation between just noticeable differences (JNDs) from the 2AFC experiments is low ($r = 0.5$, $p = 0.1$, $\log_{10}BF = -0.12$). On the other hand, the fact that these measures become strongly correlated at SNRs above threshold (peaking at 0.8) suggests that performance is limited by a mechanism common to both tracking tasks, possibly due to the motor decision stage which implements mouse movements. The fact that these correlations are high, however, reassures that at high SNR the data are still capable of yielding a graded set of responses across observers and reassures that the drop in correlation between WF and efficiency is not due to poor data quality.

Efficiency can be interpreted as an indication of how predictable, or less noisy, participants' responses are as the SNR increases. To check if this is the case, we implemented a cross-validation analysis: we estimated the participant's kernel from a subset of the data and used it to produce ideal observe responses for the data subset that was left out. In our case, we used half data (3 minutes for each condition and 15 minutes total for blocks chosen randomly) to estimate the kernel and computed the efficiency of the remaining half of data, without overlapping. Results are displayed in Figure 4D. The two samples were significantly correlated in both tasks (area: $r = 0.95$, $p < 10^{-3}$, $\log_{10}BF = 20$ and numerosity: $r = 0.97$, $p < 10^{-3}$, $\log_{10}BF = 24$).

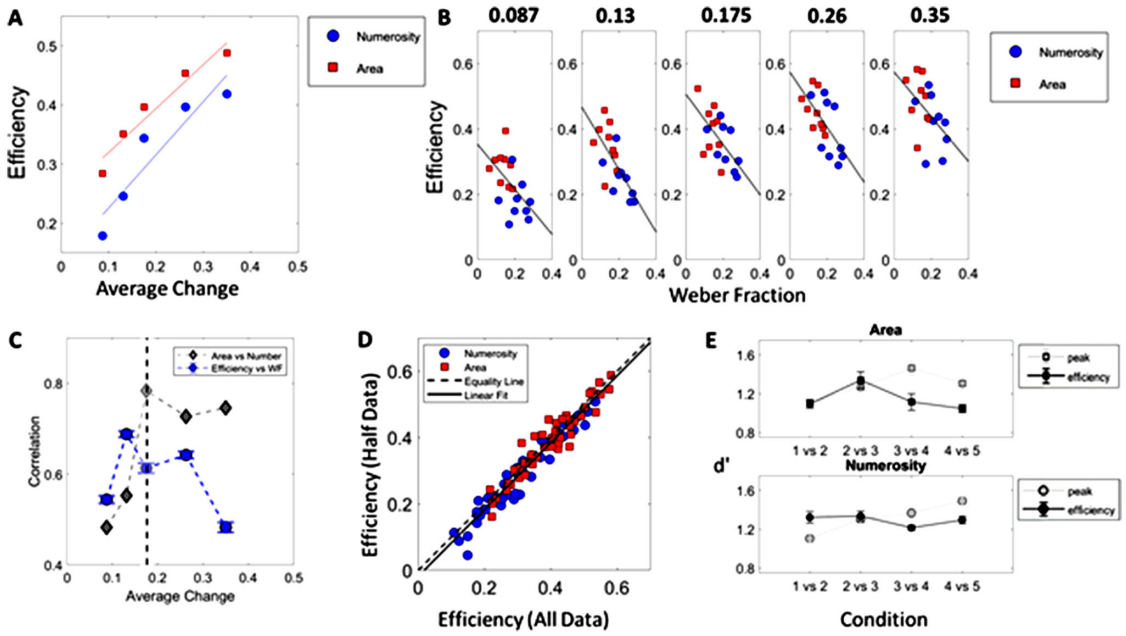

Figure 4. **Efficiency as a measure of noisiness in the annulling paradigm.** (**A**) Mean efficiency across participants plotted as a function of average change (red for area and blue for numerosity). Lines are the result of linear fit (area: m = 0.90 and q = 0.14 and numerosity: m = 0.74 and q = 0.24). (**B**) Efficiency as a function of Weber Fractions in different conditions. The two measures are significantly correlated in each condition (results in Table 3). Black lines are linear fits of the data. (**C**) Correlation between efficiencies in the two tasks (gray diamonds) and between efficiency and WFs, plotted separately for each condition, in absolute value. Error bars represent the standard error of the mean obtained by bootstrapping. The black dashed vertical line is the mean discrimination threshold across all participants and tasks. Shaded area is the S.E.M. (**D**) Results of cross-validation analysis of efficiency: efficiency from the half dataset (y axis) plotted against efficiency with whole dataset (red for area and blue for numerosity). The solid line is the linear fit of the data, which yields a slope of 1.01 and an offset −0.02. The black dashed line is the equality line. (**E**) Mean d' across conditions computed via bootstrapping, for the area task in the upper panel and for the numerosity task in the lower one.

As efficiency is significantly correlated with average changes in the signal, we asked if efficiencies could work as a proxy for stimulus strength in a realistic scenario. We compared couples of adjacent conditions (e.g. the first and second conditions) and measured the d-prime of the efficiencies and of the peaks. Figure 4E shows the mean *d'* computed via bootstrapping, averaged across participants, for the peak of the cross-correlogram (gray) and efficiency (black). Trial blocks were combined to form 1-minute-long sessions for each condition, and then estimated the *d'* of the resulting peaks of correlation and efficiencies across conditions. This procedure was repeated 500 times to obtain an estimate of the variability of the *d'*. Efficiency results in a better predictor when SNR is very low, whereas peak discriminates better at high SNR, with the two measures being equivalent in between. Participants have a lower WF in the area task than in the numerosity one, so the subjective SNR is higher in the area task, and the two *d'* result similar even in the condition with lowest SNR.

Psychometric observer

Efficiency allowed us to compare participants' performance with those of a virtual observer implemented as a linear observer model. In this way, it was possible to relate results in the tracking task to standard psychophysics, connecting inter-participant differences with their discrimination abilities (i.e. with their WFs). However, the factors limiting performance (leading to efficiencies lower than unity) are still not clear. In addition, results from efficiency indicate that, in the tested range, the assumption of linear behavior of human participants is effective.

To further investigate these aspects, we leveraged the thresholds for each participant to construct an observer that incorporates both the probability of seeing the change in the stimulus and dynamic information resulting from the annulling experiment. This virtual observer has the same impulse response function as the participant and responds either to the incoming change or to its negative, with probability sampled from the

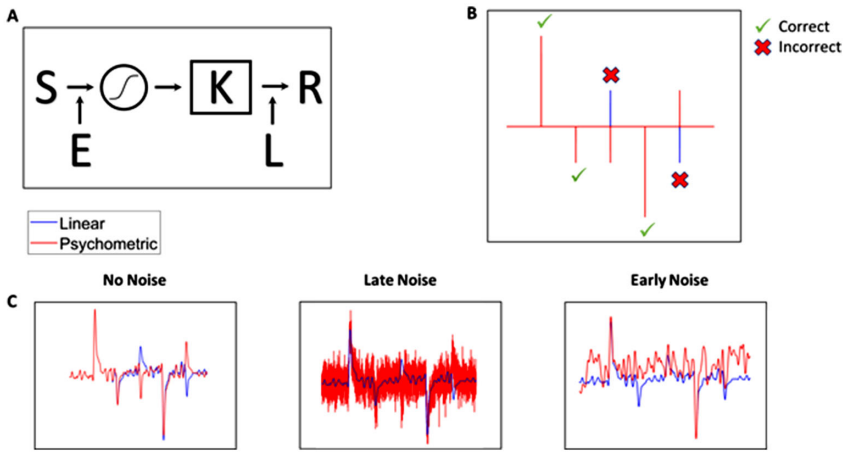

Figure 5. **Description of the Psychometric Observer.** (**A**) The Psychometric Observer receives changes in the stimulus (S) as input, which are corrupted by an early component of noise (E), simulating noise in early stages of perception. The noisy inputs are then weighted according to the psychometric function of the participant, which causes the input to be perceived either correctly or reversed. These inputs are then converted into a response (R) through a transfer function (K), and the responses are corrupted by a late noise component, simulating noise in mouse movements unrelated to the processing stage. (**B**) Example of the weighted stimuli: physical changes in the stimulus (blue) have a probability of being perceived correctly or reversed (red) by the Psychometric Observer, with a probability dependent on the participants' discrimination threshold estimated in the 2AFC experiment. Note that the two larger changes have a higher probability of being interpreted correctly, while smaller ones are more likely to be misperceived. (**C**) Example of the responses of the Psychometric Observer. Left: noiseless ($E = 0$ and $L = 0$) Psychometric Observer, with the inputs that are misperceived in panel **B** clearly resulting in opposite mouse movements in respect to the Linear Observer. Middle: Effect of the late noise L on the Psychometric Observer, resulting in errors in mouse movements completely unrelated to the processing stage; Right: Effect of the early noise E on the Psychometric Observer, also resulting in errors in mouse movements, but related to the misperception of the stimuli.

psychometric function of the participant resulting from the 2AFC experiment (see Methods). In other words, the Psychometric Observer has an additional stage of processing simulating the conversion of physical changes into subjective changes. Adding different sources of noise to this observer results in a virtual observer, which displays various degrees of suboptimal performance, and can be compared to real participants' behavior, which we called a Psychometric Observer.

Figure 5 describes the Psychometric Observer and the effect of noise on its response in simplified conditions for illustrative purposes. Figure 5A schematizes the model from which the Psychometric Observer was built. Physical changes in the stimulus (S) are corrupted by an early component of noise (E), and then the participants' discrimination ability is considered: inputs are interpreted by the Psychometric Observer either correctly or reversed with probability sampled from the psychometric function obtained in the 2AFC experiment. The resulting weighted inputs are then converted into a response (R) through a transfer function (K) estimated from the tracking experiment, and this response is corrupted by a late noise stage (L), simulating noise in the implementation of mouse movements. Figure 5B shows the difference between the linear observer (blue), which interpretates all inputs correctly, and the Psychometric Observer (red), which receives three inputs correctly and the two incorrectly, interpreting them as reversed. Figure 5C compares the responses of the Psychometric Observer (red) and the linear observer (blue) to the inputs in Figure 5B in three different conditions: on the left, the noiseless ($E = 0$ and $L = 0$) Psychometric Observer responds to the first stimuli as the linear observer, but responds in the opposite way to those that were interpreted incorrectly; in the middle, the same Psychometric Observer is corrupted by a late noise component ($E = 0$ and $L > 0$); on the right, the Psychometric Observer is corrupted by an early noise component ($E > 0$ and $L = 0$). When a late noise component is added to the response of the Psychometric Observer, simulating involuntary movements of the mouse unrelated to the stimuli, the resulting response shows random displacements from the linear observer that do not change the overall shape of the response. On the other hand, the effect of the early noise component reflects responses that are driven from misperceptions of the stimuli. In this simplified situation, these effects were exaggerated to highlight the differences between the two noise sources.

Figure 6 shows the Psychometric Observer built from the psychometric functions from two different participants. The efficiency is plotted against the logarithm of the ratio between average changes in the stimuli and the participant's JND, which we labeled

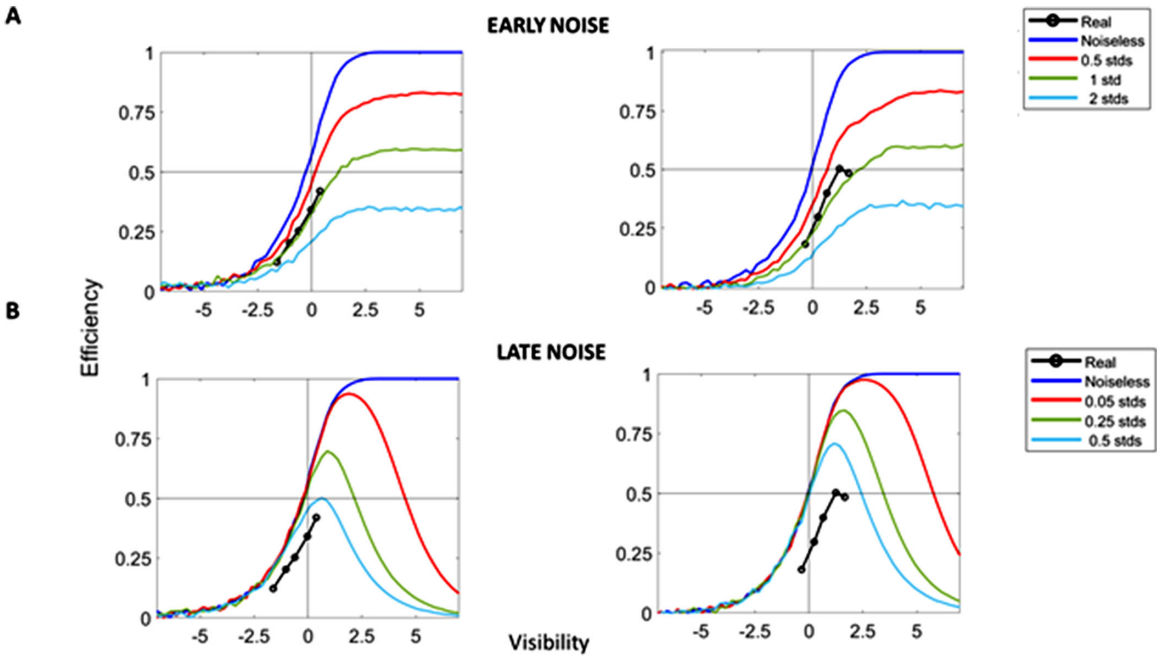

Figure 6. **Psychometric Observers.** Panels on the left show results for the participant with highest Weber Fraction in the 2AFC task, the ones on the right for the participant with the lowest Weber Fraction. The efficiency is plotted against visibility defined as the logarithm of the ratio between the average change and the JND. Negative values of visibility represent conditions where the physical average change in the stimulus is lower than the participant's JND, positive values represent above-threshold conditions, and visibility is zero when changes match the participant's JND. In all panels, blue lines are the results for the noiseless Psychometric Observer and black circles are results of the participant from which the Psychometric Observer was built. (**A**) Gaussian noise with increasing standard deviation was added to the input of the Psychometric Observer, simulating noise in perception. At high visibilities, the efficiency saturates. At low visibilities, the Psychometric Observer's performances are worsened in a way resembling real participants' results. (**B**) Gaussian noise with increasing standard deviation was added to the output of the Psychometric Observer, simulating noise in mouse movements. Such noise has very little effects at low SNRs, and a much bigger effect at high SNRs.

visibility: according to this definition, negative values of visibility represent below threshold condition, positive values above threshold conditions, and visibility is zero at threshold (average change equal to a JND). In this latter case, only half of the stimuli are interpreted correctly, yielding an efficiency of 50%. Figure 5 displays the results from the Psychometric Observer in the numerosity task for two representative subjects, the ones with the highest (left) and lowest (right) WF. Note that each participant was tested on the same physical condition, but because JNDs are different the resulting subjective condition is different for the two cases, with the participant on the right panels being almost always over the threshold and the participant on the left panels almost always under the threshold.

Figure 6A shows the effect of various levels of the early noise component (noise in perception) on the Psychometric Observer, expressed in units of standard deviations of the input signal. Interestingly, this noise affects the Psychometric Observer's performance both at low and high visibilities, with the latter showing saturation effects. This suggests that in the tested range, deviation from ideal behavior can be explained by an early component of noise acting on perception.

Figure 6B shows the effect of the addition of late noise (motor noise), rescaled according to the corresponding condition, as described in the Methods, expressed in units of standard deviations of the noiseless signal. At low visibility levels, the effect of noise is close to zero, whereas at high visibility the efficiency drops. The decrement in efficiency at high visibility is because the late noise component prevents the psychometric observer reaching 100%. As the visibility increases, the discrepancy between the ideal and the psychometric observer increases, reducing the efficiency. When higher levels of noise are added, the difference between the ideal and the psychometric observers arises at lower visibilities, generating the bell-shaped curves shown in Figure 6B. This behavior may explain the saturation effect at high visibility present in the data (black circles) in the left panel, but it is clearly not sufficient to explain deviations from the ideal psychometric observer at lower visibilities, where noisy observers are close to the noiseless. Both noise

sources are required to match participants' behavior, as saturation effects are present for the participant in the right panels, but differences between the ideal observer and the human observer at low visibility can only be explained by an early component of noise.

Discussion

In this paper, we show that it is possible to generalize tracking tasks to different perceptual domains, with results analogous to previous tracking experiments (Bhat et al., 2018; Bonnen et al., 2015; Bonnen et al., 2017; Huk et al., 2018; Mulligan, 2002; Mulligan et al., 2013), notably showing a dependency between SNR and performance. Thus, it is possible to extend the "tracking" paradigm beyond spatial features of the stimuli, potentially to all domains of perception, with an ecological and friendly paradigm. The choice of an annulling paradigm was to avoid possible confounding effects due to participants having to switch gaze between two stimuli, one controlled by the computer and one by the participant. However, this design is conceptually very similar to the traditional tracking technique and has been previously used with similar results to standard tracking (Cormack, 2019; Li et al., 2005). We showed that participant performance in various conditions is related to perceptual discrimination thresholds. Analysis of the results suggests that the main source of noise originates early in the process of analyzing visual stimuli, so the suboptimal efficiency of the responses cannot be ascribed to random variance in the motor plant, but mostly to noise in earlier stages of perceptual processing. This was demonstrated by comparison of participants' behavior to a virtual observer, which we call a Psychometric Observer, which incorporates both the dynamic information from the tracking task and the perceptual discrimination threshold from standard psychophysical measurements. This result is important for general applicability of the tracking paradigm since differences across participants and across conditions unrelated to stimulus presentation would have restricted the relevance of tracking in studies of perception.

Our paradigm and that of Bonnen et al. (2015) implemented task difficulty in two different ways. Bonnen et al. (2015) required position tracking and varied the visibility (the SNR) of the stimulus by changing the variance of the gaussian distribution. Such changes resulted in approximately linearly scaling variations in the cross-correlogram parameters, peak, lag, and width, associated with the altered visibility of the target. In the present work, the stimulus SNR was varied by increasing or decreasing the changes occurring at every frame, presumably in the face of constant internal noise of the observer. Both studies modulated the SNR of the stimulus, in our case modulating the signal, whereas Bonnen et al. (2015) modulated the noise. The results shown in Figure 1 show that augmenting the SNR in this way improves participants' performances in both tasks.

When we correlated the mean parameters of the cross-correlogram across conditions with the corresponding WFs, we found that peak and width of the correlation were related to perceptual sensitivity, but lag was not (see Table 3). We can only speculate on the reasons for this difference. Importantly, our paradigm required annulling the stimulus changes, so the samples presented on the screen were always close to the standard. This is slightly different from a tracking experiment in which the observer has the pressure to follow the target. We also point out that in a previous study (Bhat et al., 2018) where participants tracked the direction of motion, lag did not correlate significantly with other performance measures. However, lag shows a significant reduction with increasing signal strength, presumably because of the increased visibility of stimulus changes (Harris & Wolpert, 1998). This suggests that lag measures may be related to perceptual qualities only loosely, and mostly to the promptness of decision-making processes (Palmer, Huk, & Shadlen, 2005).

With the simple computation described in the appendix, it was possible to estimate the participants' impulse response function, with which we generated a linear observer to compare to participants' responses (Geisler, 1989a). We used the whole dataset to recover the motor kernel, as previous reports have shown that kernels vary subtly between SNR conditions (Bhat et al., 2018; Bonnen et al., 2015). Thus, selecting a particular condition as a gold standard could have artificially decreased the efficiency in the other conditions, as the predicted responses would have been derived by a Kernel that is optimal for another dataset. Thus, we used a Kernel from a comprehensive dataset that was representative of all conditions.

Correlating the responses of the virtual observer with human participants yields a measure of the similarity between the two observers, and therefore about how much of the participants' behavior can be predicted assuming linearity. We term this measure efficiency, because our ideal observer simulates a system that responds to all changes in the stimulus, and the comparison between the two observers yields an estimate of how much of the input signal is converted into a response by the real observer. In this sense, our linear observer can be interpreted as an optimal observer constrained by the participants' motor implant, described by their response function (Geisler, 1989a; Geisler, 1989b; Geisler, 2003; Geisler, 2011). The choice of correlation as a measure of similarity was dictated by the fact that only the shape of the response kernel can be safely estimated, but not its amplitude.

Using correlation, which is invariant for multiplicative factors, allows for a direct comparison of the responses. Our model also assumes that Weber's law holds for both human and ideal observer. This is justified, since it has been shown that Weber's law holds for both numerosity and size perception (Anobile et al., 2014; Ganel et al., 2008). With increasing SNR, participant responses become more predictable, or less noisy, and the correlation between ideal and real responses increases. This is shown in Figure 4A, for both tasks, where a linear trend like that of Figure 1C is present. In addition, the participants' efficiency is significantly correlated with WFs, as were the cross-correlogram parameters. This suggests that efficiency is an effective measure for participant performance in a tracking task, based on the assumption of linearity in the conversion of changes in the stimuli into motor responses in the tested range. The results of Figure 4C suggest there is a linear range where efficiency is best related to perceptual ability: as signal strength increases, efficiencies in the two tasks become more correlated. This is compatible with the interpretation that for higher SNRs participant responses are relatively more corrupted by noise from mouse movements than at lower SNR levels, reducing the contribution of perceptual mechanisms to the differences between the two tasks. This interpretation is reinforced by the fact that correlating the efficiencies with WFs results in higher correlation for the central conditions than for the extremal conditions, where responses are more likely to be corrupted by noise. Contrasting the real observer against an ideal observer with perfect memory also leads to the inclusion of memory drifts into the analysis, which is undesirable. However, because we are basing our comparison on how the ideal and real observer would have changed their responses as a function of stimulus changes, the impact of drift is negligible, as even large drifts would spread over many frames (typically 2400) per session. Indeed, detrending the data to remove drift biases gave near identical results (99.99% correlated, mean difference between efficiencies approximately 10^{-4}).

We tested the reliability of efficiency through cross-validation: half of the trial blocks were used to estimate the impulse response function, which then generated ideal responses for the remaining half of the data. We found a strong correlation ($r > 0.9$) between efficiencies computed in the two ways, proving the reliability of efficiency as a measure of performance. This also suggests that robust results would have been found with shorter acquisitions.

We note that the estimation of the impulse response function can be done by aggregating data from different SNR levels, so the amount of data from which the efficiency can be calculated is much higher than for the single parameters of the cross correlogram, which must be computed separately in each condition. This provides a strategy for faster data acquisition, because the Psychometric Observer analysis shows that exists a broad range of informative SNR conditions that can be tested. Tuning algorithms can also be used to adapt the testing conditions until the desired range is reached, using efficiency as a parameter for convergence. This strategy may be particularly useful when the optimal testing range cannot be known a priori, as in absence of knowledge about discrimination threshold. Additionally, we showed that not all SNR conditions are equally informative, so adopting a tuning algorithm will allow to test conditions above or below the perceptual discrimination threshold, depending on the aspects under investigation. In addition, in Figure 4E, we showed that at very low levels of SNR, when the cross-correlogram becomes noisy, efficiency becomes a better discrimination parameter than peak. Therefore, efficiency can be used for noise levels that would make the estimate of the cross-correlogram parameters impractical.

Simulations with the Psychometric Observer also show an important feature of the efficiency measure. Efficiency is particularly dependent on sensory noise especially for variations close to threshold, because perceptual noise is most important in this range and motor noise is not. This is demonstrated by the plots of Figure 6, but also by the fact that the Psychometric Observer captured the rise of efficiency as a function of signal strength. On the other hand, at high visibility levels, where the sensory component has plateaued, the crucial limiting factor for the kernel is the motor plant implementing the action of the observer. In our paradigm, conditions with higher perturbations also entailed a higher mouse-to-screen gain, so that the observer would perform the task identically with the same motions.

These analyses also lead to the speculation that tests at low visibility levels may indeed return a proxy for the quality of sensory representations, whereas the performance at higher SNRs is more dependent on the ability to accumulate and transform a host of highly salient stimuli and thus may reflect secondary processes.

In conclusion, our results support extending the tracking technique beyond pure object-tracking situations, to yield a large amount of psychophysically relevant data with short acquisitions. By testing two different perceptual mechanisms (numerosity and size), we have shown that tracking performances are strictly linked to perceptual abilities, with different discrimination thresholds leading to different tracking performances. Still, in both tasks, the SNR of different stimulus features modulates tracking performance in a similar way. This paradigm may be particularly useful in testing participants who can produce only limited amounts of data, such as children or clinical populations. Additionally, by virtue of its dynamic implementation, the tracking paradigm may provide novel temporal information about perceptual

mechanisms, ideally exploiting realistic stimuli in ecologically relevant circumstances.

Keywords: ideal observer, continuous tracking, numerosity

Acknowledgments

Commercial relationships: none.
Corresponding author: Pierfrancesco Ambrosi.
Email: pfa2804@gmail.com.
Address: Department of Neuroscience, Psychology, Pharmacology, and Child Health, University of Florence, Padiglione 26, Via di San Salvi, 26, 50135, Florence, Italy.

References

Anobile, G., Cicchini, G. M., & Burr, D. C. (2014). Separate mechanisms for perception of numerosity and density. *Psychological Science, 25*(1), 265–270.

Anobile, G., Cicchini, G. M., & Burr, D. C. (2016). Number As a Primary Perceptual Attribute: A Review. *Perception, 45*(1-2), 5–31.

Anobile, G., Cicchini, G. M., Pomè, A., & Burr, D. C. (2017). Connecting visual objects reduces perceived numerosity and density for sparse but not dense patterns. *Journal of Numerical Cognition, 3*(2), 133–146.

Arrighi, R., Togoli, I., & Burr, D. C. (2014). A generalized sense of number. *Proceedings. Biological Sciences, 281*(1797), 20141791.

Baker, D. H., & Meese, T. S. (2012). Size adaptation effects are independent of spatial frequency aftereffects. *Perception, 41*(Suppl. 1), 1.

Barlow, H. B. (1962). A method of determining the overall quantum efficiency of visual discriminations. *Journal of Physiology, 160*, 155–168.

Bhat, A., Cicchini, G. M., & Burr, D. C. (2018). Inhibitory surrounds of motion mechanisms revealed by continuous tracking. *Journal of Vision, 18*(13), 7.

Bonnen, K., Burge, J., Yates, J., Pillow, J., & Cormack, L. K. (2015). Continuous psychophysics: Target-tracking to measure visual sensitivity. *Journal of Vision, 15*(3), 14.

Bonnen, K., Huk, A. C., & Cormack, L. K. (2017). Dynamic mechanisms of visually guided 3D motion tracking. *Journal of Neurophysiology, 118*(3), 1515–1531.

Brainard, D. H. (1997). *The psychophysics toolbox. Spatial Vision, 10*, 433–436.

Burr, D., & Ross, J. (2008). A visual sense of number. *Current Biology, 18*(6), 425–428.

Castaldi, E., Aagten-Murphy, D., Tosetti, M., Burr, D., & Morrone, M. C. (2016). Effects of adaptation on numerosity decoding in the human brain. *Neuroimage, 143*, 364–377.

Cicchini, G. M., Anobile, G., & Burr, D. C. (2016). Spontaneous perception of numerosity in humans. *Nature Communications, 7*, 12536.

Cormack, L. (2019). Dynamics of Motion Induced Position Shifts Revealed by Continuous Tracking. *Vision Sciences Society Annual Meeting Abstract, September 2019, Vol. 19, 294c*.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. New York, NY: Oxford University Press.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology, 47*(6), 381–391.

Franconeri, S. L., Bemis, D. K., & Alvarez, G. A. (2009). Number estimation relies on a set of segmented objects. *Cognition, 113*(1), 1–13.

Ganel, T., Chajut, E., & Algom, D. (2008). Visual coding for action violates fundamental psychophysical principles. *Current Biology, 18*(14), R599–R601.

Geisler, W. S. (1989a). Ideal observer theory in psychophysics and physiology. *Physica Scripta, 39*, 153.

Geisler, W. S. (1989b). Sequential ideal-observer analysis of visual discriminations. *Psychological Review, 96*(2), 267–314.

Geisler, W. S. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science, 27*(3), 379–402.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research, 51*(7), 771–781.

Geisler, W. S., & Ringach, D. (2009). Natural systems analysis. Introduction. *Visual Neuroscience, 26*(1), 1–3.

Harris, C., & Wolpert, D. (1998). Signal-dependent noise determines motor planning. *Nature 394*, 780–784.

Harvey, B. M., Fracasso, A., Petridou, N., & Dumoulin, S. O. (2015). Topographic representations of object size and relationships with numerosity reveal generalized quantity processing in human parietal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 112*(44), 13525–13530.

He, L., Zhang, J., Zhou, T., & Chen, L. (2009). Connectedness affects dot numerosity judgment: implications for configural processing. *Psychonomic Bulletin Review, 16*(3), 509–517.

Huk, A., Bonnen, K., & He, B. J. (2018). Beyond Trial-Based Paradigms: Continuous Behavior, Ongoing Neural Activity, and Natural Stimuli. *Journal of Neuroscience, 38*(35), 7551–7558.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271–304.

Kleiner, M., Brainard, D., Pelli, D., & Ingling, A. (2007). What's new in Psychtoolbox-3. *Perception, 36*, 14.

Kristensen, S., Fracasso, A., Dumoulin, S. O., Almeida, J., & Harvey, B. M. (2021). Size constancy affects the perception and parietal neural representation of object size. *Neuroimage, 232*, 117909.

Li, L., Sweet, B. T., & Stone, L. S. (2005). Effect of contrast on the active control of a moving line. *Journal of Neurophysiology, 93*(5), 2873–2886.

Mulligan, J. B. (2002). Sensory processing delays measured with the eye-movement correlogram. *Annals of the New York Academy of Sciences, 956*, 476–478.

Mulligan, J. B., Stevenson, S. B., & Cormack, L. K. (2013). Reflexive and voluntary control of smooth eye movements. *Paper presented at the Human Vision and Electronic Imaging XVIII. Proceedings of the SPIE, Volume 8651, id. 86510Z* 22 pp. Retrieved from: https://ui.adsabs.harvard.edu/abs/2013SPIE.8651E..0ZM/abstract#:~:text=Reflexive%20and%20voluntary%20control%20of%20smooth%20eye%20movements,be%20encountered%20in%20vehicles%20like%20aircraft%20and%20automobiles.

Nieder, A. (2019). *A Brain for Numbers: The Biology of the Number Instinct*. New York, NY: MIT Press.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision, 5*(5), 376–404.

Pelli, D. G. (1991a). *The quantum efficiency of vision*. Cambridge, UK: Cambridge University Press.

Pelli, D. G. (1991b). *Noise in the visual system may be early*. New York, NY: The MIT Press.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10*(4), 437–442.

Pelli, D. G., & Farell, B. (1999). Why use noise? *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 16*, 647–653.

Pooresmaeili, A., Arrighi, R., Biagi, L., & Morrone, M. C. (2013). Blood oxygen level-dependent activation of the primary visual cortex predicts size adaptation illusion. *Journal of Neuroscience, 33*(40), 15999–16008.

Ross, J., & Burr, D. C. (2010). Vision senses number directly. *Journal of Vision, 10*(2), 10.1–8.

Tonelli, A., Pooresmaeili, A., & Arrighi, R. (2020). The Role of Temporal and Spatial Attention in Size Adaptation. *Frontiers in Neuroscience, 14*, 539.

Appendix A: Cross-correlogram as an estimate of the transfer function

To ease the notation, we define

$$conv\,(A\,(t)\,,B\,(t)) = \int A\,(\tau)\,B\,(t\ -\ \tau)\,d\tau$$

and

$$xcorr\,(A\,(t)\,,B\,(t)) = \int A^{*}\,(\tau)\,B\,(t+\tau)\,d\tau$$

where A(t) and B(t) are two generic functions and A*(t) is the complex conjugate of A.

Assuming a linear model for the observer, the output $\boldsymbol{O}$ can be written in function of the input $\boldsymbol{I}$ as a convolution

$$\boldsymbol{O} = conv\,(\boldsymbol{I},\boldsymbol{K}) \quad (1)$$

where $\boldsymbol{K}$ is the kernel of the transfer function. When computing the cross correlation between input and output we have

$$\boldsymbol{XC} = xcorr\,(\boldsymbol{O},I)$$

where we neglect a minus sign associated with the annulling paradigm. Rewriting the output as in Equation 1 we have

$$\boldsymbol{XC} = xcorr\,(conv\,(\boldsymbol{I},K)\,,I)$$

Using the property of convolutions and cross-correlation

$$xcorr\,(conv\,(\boldsymbol{A},\boldsymbol{B})\,,\boldsymbol{C}) = conv\,(xcorr\,(\boldsymbol{A},C)\,,\boldsymbol{B})$$

we have

$$\boldsymbol{XC} = conv\,(xcorr\,(\boldsymbol{I},I)\,,\boldsymbol{K})$$

and the term $xcorr(\boldsymbol{I},\boldsymbol{I})$ is the autocorrelation of the input. Changes in input are randomly distributed in time and instantaneous, that is, an aperiodic comb function:

$$\boldsymbol{I} = \sum\nolimits_{n} I_n * \ \delta\,(t-t_n)$$

where t_n are the random instants when changes occur. Their autocorrelation is then a delta function in $t = 0$.

We have:

$$\boldsymbol{XC} = conv\left(\mathrm{E} * \delta\left(0\right), \boldsymbol{K}\right)$$

Where $E = \sum_n I_n^2$. Then

$$\boldsymbol{K} = \frac{\boldsymbol{XC}}{\boldsymbol{E}}.$$

This enables us to estimate the filter kernel up to a scaling factor. In real situations, both the input and the output have a noise component, which results in noise in the cross correlation.

Appendix B: Root explained variance as a measure of efficiency

The response from the human observer H can be thought as the response from the ideal observer V with noise η, so

$$H = V + \eta$$

where V is obtained from Equation 1 using the kernel estimated from tracking data using the procedure in Appendix A.

The correlation between the human and ideal is then

$$r = \frac{\sigma_{VH}}{\sigma_V \sigma_H}$$

where σ_{V} and σ_{H} are the standard deviations of the ideal and human observer responses respectively, and σ_{VH} is the covariance between the two. σ_{VH} can be made explicit as

$$\sigma_{VH} = \sigma_{V,V+\eta} = \sigma_{VV} + \sigma_{V\eta} = \sigma_V^2$$

because the covariance $\sigma_{\mathrm{V}\eta}$ between the ideal observer and noise is zero. Note that the last equality holds only if the human and virtual observer are considered with the same temporal order. Then

$$r = \frac{\sigma_{VH}}{\sigma_V \sigma_H} = \frac{\sigma_V^2}{\sigma_V \sigma_H} = \frac{\sigma_V}{\sigma_H}$$

Taking the square of the correlation yields the explained variance, resulting in the definition of efficiency E given in (Pelli & Farell, 1999) as the ratio of the energies of the ideal and human observer:

$$E = \frac{\sigma_V^2}{\sigma_H^2}$$

The two ways of defining efficiency are therefore strictly linked, but correlation is better suited for our implementation, as it depends on the timing of changes. In addition, it preserves the relative sign of mouse movements and changes in the stimulus, because in our implementation responding with a rightward movement of the mouse is not equivalent to respond with a leftward movement of the same size to the same stimulus change.