Test–retest reliability of eye-tracking metrics for the measurement and classification of sign- trackers and goal-trackers

Marco Badioli[1], Claudio Danti[1], Luigi Degni[1], Gianluca Finotti[1,2], Valentina Bernardi[1], Lorenzo Mattioni[1], Francesca Starita[1], Sara Giovagnoli[1], Giuseppe di Pellegrino[1], Mariagrazia Benassi[1] & Sara Garofalo[1]

Affiliations:
[1]Department of Psychology, University of Bologna
[2]Department of Psychology, Birkbeck University of London

Corresponding author:
Sara Garofalo
Department of Psychology
University of Bologna
Via Rasi e Spinelli, 176, Cesena, Italy
sara.garofalo@unibo.it

Abstract

In animal research, reward-predictive cues shape behavior through Pavlovian conditioning, yet animals vary in the value they assign to these cues. Sign-trackers (ST) attribute both incentive and predictive values to the cues, orienting their attention to them, while goal-trackers (GT) assign solely predictive value, orienting their attention rapidly toward the forthcoming reward. Although most animal studies report sign-tracking and goal-tracking as stable, trait-like behavioral profiles, human research has produced inconsistent results, raising questions about the reliability and the stability of this behavior. To address these issues, we investigated the test-retest reliability and stability of the classification over a four-month period of the gaze index most frequently adopted in human sign-tracking and goal-tracking literature. Our findings revealed good stability for sign-tracking behavior, but limited consistency for goal-tracking behavior. These results raise the possibility that goal-tracking may be either genuinely rare in the population or poorly captured by the current index. Overall, while the gaze index holds promise for identifying sign-tracking behavior, methodological refinements or alternative approaches may be needed to more reliably detect these behaviors in future research.

Keywords: Sign-trackers, Goal-trackers, Reliability, Reinforcement learning, Eye-tracking, Pavlovian learning

1. Introduction

Environmental cues associated with rewards (like a restaurant logo signaling the availability of appealing food) provide fundamental information for predicting the outcomes of our choices and for preparing to exert adaptive behavior. However, such cues can also lead to maladaptive behavior, prompting externalizing disorders like addiction and impaired impulse control (Everitt & Robbins, 2016; Robinson & Berridge, 1993; Robinson & Berridge, 2025). Experimental evidence has revealed substantial inter-individual variability in how environmental cues are processed and exploited for guiding behavior (Badioli et al., 2024; Degni, 2024; Degni & Garofalo, 2025; Doya, 2008; Garofalo et al., 2019; Garofalo & Di Pellegrino, 2015; Hogarth & Duka, 2006). Specifically, a growing body of research has pointed out individual differences in the propensity to attribute incentive salience to reward-predictive cues, distinguishing two behavioral phenotypes: sign-trackers (ST) and goal-trackers (GT) (for recent reviews, see Anselme & Robinson, 2020; Colaizzi et al., 2020; Felix & Flagel, 2024; Heck et al., 2024; Sarter & Phillips, 2018).

In animal models, ST and GT are usually identified through a Pavlovian conditioning paradigm in which a neutral cue (the “sign”; e.g., a light) is repeatedly paired with an outcome delivery (the “goal”; e.g., a food pellet). While GT attribute only a predictive value to such cues (i.e., upon cue presentation, they focus on the incoming reward location, with a tonic dopaminergic response that peaks at reward delivery), ST assign both predictive and incentive value to the cues (i.e., upon cue presentation, they approach the cue itself, with a greater phasic dopaminergic response that peaks at cue presentation) (Brown & Jenkins, 1968; Colaizzi et al., 2020; Flagel et al., 2011; Robinson & Flagel, 2009). In other words, for sign-trackers, the cue itself becomes attractive and functions as a “motivational magnet” (Anselme et al., 2013; Anselme & Robinson, 2020).

The investigation of ST and GT differences in humans faces important challenges in mirroring the animal approach behavior toward the sign or the goal (Colaizzi et al., 2020; Heck et al., 2024). The first translational method was introduced by Garofalo and di Pellegrino (2015), who developed a computerized Pavlovian conditioning task coupled with an eye-tracking device to measure oculomotor responses. The eye-gaze was proposed to mimic animals’ approach behavior. In this task, the cue was presented alone in a predictable location, and only after five seconds was the outcome displayed in a different predictable location. Participants’ gaze was free to explore, and the dwell time was used to calculate an eye-gaze index indicating the proportion of time spent on the reward-predictive cue (the sign) relative to the reward (the goal). A median split on this index classified participants with higher values as ST and those with lower values as GT.

Although alternative measures, such as event-related potentials (Versace et al., 2016, 2019), physical Pavlovian conditioning tasks (Colaizzi et al., 2023; Cope et al., 2023; Joyner et al., 2018) and value-modulated attention capture (VMAC) tasks (Duckworth et al., 2022; Liu et al., 2021; Watson et al., 2024), have been explored for human ST and GT classification, they face both theoretical and practical limitations (see Colaizzi et al., 2020; Heck et al., 2024 for details), therefore the original paradigm implemented by Garofalo and di Pellegrino (2015) currently remains the most widely used (Cherkasova et al., 2024; Degni et al., 2024a; Degni et al., 2024c; Dinu et al., 2024; Schad et al., 2019; Schettino et al., 2024). Nevertheless, the psychometric properties of such measures are unexplored, posing the critical but often underestimated issue of assessing the reliability of measures used in cognitive tasks (Enkavi et al., 2019; Hedge et al., 2018; Pennington et al., 2025; Saeedpour et al., 2023; Zorowitz & Niv, 2023).

Relatedly, although studies have reported that GT animals may shift to ST behavior following extensive Pavlovian conditioning training (Keefer et al., 2020; Srey et al., 2015; Villaruel & Chaudhri, 2016) or under conditions of reward uncertainty (Robinson et al., 2015), and that exposure to an auditory cue may induce the reverse switch from ST to GT behavior (Meyer et al., 2014), the prevailing view in the animal literature tends to consider ST and GT as stable traits (Campus et al., 2016; Colaizzi et al., 2023; Dickson et al., 2015; Felix & Flagel, 2024; Flagel et al., 2008; Meyer et al., 2012; Robinson & Flagel, 2009). Crucially, ST has been consistently associated with higher impulsivity, novelty seeking, and risk propensity. Moreover, studies inducing maladaptive behaviors through the administration of specific substances (e.g., quinpirole, a D2/D3 agonist) have reported a link between ST and addiction-like and impulsive control-related behaviors (Felix & Flagel, 2024; Flagel et al., 2010; King et al., 2016; Lovic et al., 2011; Swintosky et al., 2021; Yager & Robinson, 2010, but see also Fraser & Holland, 2019; Saunders et al., 2014). Nevertheless, such associations appear less clear in the human population (Felix & Flagel, 2024). Although complex gene-environment interactions could introduce higher variability and intraindividual fluctuations that prevent such firm conclusions in humans (Colaizzi et al., 2020; Meyer et al., 2012), translational differences in the task, as well as low reliability and validity of the measures adopted to classify STs and GTs, also likely contribute to such conflicting results (Colaizzi et al., 2020).

Standardizing tasks and methods, along with examining the stability of the ST and GT behavioral phenotypes, appears thus essential to enhance the robustness of results in this literature and understand its clinical implications (Colaizzi et al., 2020; Heck et al., 2024). To address these gaps, the present study aims to provide critical insights into the psychometric properties of the most widely used gaze index to investigate ST and GT by evaluating its test-retest reliability over a four-month period.

Establishing test-retest reliability, along with validity and categorization stability measures, will provide information about the robustness of this measure (Koo & Li, 2016), as well as provide new insights as to whether ST and GT behavioral phenotypes can be intended as stable trait-like characteristics or as influenced by state-dependent fluctuations (Atkinson & Nevill, 1998; Fleeson & Jayawickreme, 2015; Shrout & Fleiss, 1979).

2. Methods

2.1 Participants

A total of 173 participants were recruited for the first experimental session (T1) from the Italian population. Approximately four months later (Mean = 126.07 days; SD = 16.80), participants returned for the second experimental session (T2). A total of 76 participants completed the T2 session. To minimize experimenter bias, data were analyzed only after the T2 session. Eight participants were further excluded due to eye-tracking registration issues that did not allow for recording sufficient data for at least one experimental session (see Eye-tracking data pre-processing and analysis). The final sample consisted of 68 participants (male = 36; mean age (SD) = 22.83 (±2.81) years; mean years of education (SD) = 16.24 (±1.50)).

The sample size was determined based on Mokkink et al. (2023) by simulating the Intraclass Correlation Coefficient (ICC) with a two-way random effects model for consistency and agreement. For a power = 0.8, the expected correlation between repeated measurements = 0.6, and a 95% confidence interval width = 0.4, a minimum of 50 participants was required. Recruitment was conducted across four separate experiments, each employing a comparable Pavlovian conditioning protocol (see Pavlovian conditioning task). The inclusion criteria for the participant recruitment were: 1) no diagnosis of neurological or psychiatric disorder; 2) not taking medications that could alter cognitive abilities; 3) having normal or corrected-to-normal vision.

A binomial test against the expected value of equal distribution was used to check whether the drop-out rate between T1 and T2 could be systematically associated with ST or GT group membership. Results indicated an equal number of participants per group for both the median split classification (ST = 47; GT = 47, p = 1) and the tertiary split classification (ST = 32, p = 0.83; GT = 31, p = 1), suggesting no specific group imbalance in the drop-out rate.

2.2 Procedure

Participants were instructed to refrain from eating for at least 3 hours before the experiment to enhance the incentive value of the food reward. Upon arrival at the laboratory, they were seated comfortably in a quiet room and provided with an informed consent form to review and sign. A PC monitor was positioned at the center of the participant's visual field at a viewing distance of 60 cm. The eye-tracking system was mounted on the participant's head and combined with forehead-chin support to ensure comfort and stability throughout the session.

Before the Pavlovian conditioning, each participant rated their subjective liking ("How much do you usually enjoy eating it?") and wanting ("How much would you like to eat it now?") of four savory (e.g., chips) and five sweet (e.g., chocolate) highly palatable foods via two separate Likert scales ranging from 0 (not at all) to 9 (very much). Participants then performed a computerized Pavlovian conditioning task (see Pavlovian conditioning task section), where the corresponding image of the food with the highest wanting rating was inserted into the experimental task as the rewarding outcome. The real food item selected was placed near the participant to enhance motivation for winning the reward during the task. After the experiment, participants were rewarded with the selected food rewards. For example, a participant who selected chips saw an image of the chips during each rewarded trial and, at the end of the task, received a small bag of chips. This procedure was designed to prevent a rapid sense of satiation and maintain high motivation throughout the experiment. All participants ultimately earned and received the same total amount of reward.

Additionally, participants rated the subjective liking of the visual cues used as CSs both before and after the Pavlovian conditioning task, to assess the initial comparability of the CS and the selective increase in liking for the CS+, as compared to the CS-. Following the Pavlovian conditioning task, participants completed a brief 5-point eye-tracking calibration task (50 seconds) to ensure data accuracy (Hooge et al., 2019). Participants performed the same procedure and version of the task at T1 (test) and T2 (retest).

2.3 Pavlovian conditioning task

The experimental task (Figure 1A) was based on the Pavlovian conditioning task from Degni and colleagues (2024a), implemented using OpenSesame v3.2 (Mathôt et al., 2012). In this task, participants learned stimulus-outcome associations through repeated pairings, whereby initially neutral cues became conditioned stimuli (CS). Each trial began with the presentation of two empty squares, one on the top (CS location, the “sign”) and another on the bottom (outcome location, the “goal”). A central fixation cross was displayed for a variable intertrial interval (ITI) between 5000-6000 ms. During this period, participants were instructed to maintain their gaze on the fixation cross. A CS was then presented for 6000 ms in the top square, followed by the corresponding outcome during the last 1000ms of the CS presentation (i.e., the CS and the outcome were presented together during the last second). One CS (CS+) was paired with a reward in 80% of trials, while in the remaining 20%, a black "X" appeared, indicating no reward. Another CS (CS-) was associated with the black "X" in 100% of the trials; hence, CS- was never paired with a reward.

The CSs consisted of two distinct fractals, balanced for luminance, complexity, and color saturation (Finke et al., 2021), and counterbalanced across participants. Different fractals were used in T1 and T2 sessions to avoid bias from preexisting preferences or recall of previous associations. The rewarding outcome consisted of the corresponding image of an individually tailored food reward (see Procedure).

Task instructions were displayed on the monitor, and participants were required to read them aloud. The instructions were as follows: "Some stimuli will appear on the upper screen of the slot machine. Food will appear on the lower screen. Every time the slot machine is empty, look at the central cross. IMPORTANT: Pay attention to the association between the stimulus that appears at the top and the food you receive. Every time you see the food, you will win it. When you see the X, you will not win anything". The experimenter then provided a verbal summary to ensure the participant fully understood the task.

After two practice trials, the main task began. The task consisted of a minimum of two blocks of 20 trials each (10 per CS per block). After each block, participants were required to report all stimulus-outcome associations to confirm the learning of contingencies. The learning criterion was met when the participant correctly identified all associations for two consecutive blocks. Participants who correctly met this criterion successfully completed the experimental task; otherwise, the task terminated after four incorrect responses.

The Pavlovian conditioning task was used across four experiments during the T1 recruitment. The task was identical across all experiments, except for a few aspects detailed below, with corresponding adjustments implemented to ensure comparability. First, in three experiments, two CS+ stimuli (each paired with equally liked and wanted outcomes) and one CS– were used, whereas in the first experiment, only a single CS+ and a single CS– were included. Accordingly, the three experiments consisted of 30 trials per block (10 trials per CS per block), while the first experiment consisted of 20 trials per block (10 trials per CS per block). The scores were averaged across CS+ and trials for comparability across all experiments. Second, ocular parameters were online recorded with a sampling rate of 100 Hz in three experiments and 200 Hz in one; the latter was offline downsampled at 100 Hz to ensure data comparability. Of note, the eye-tracking device was the same across all experiments (see Eye tracking parameters).

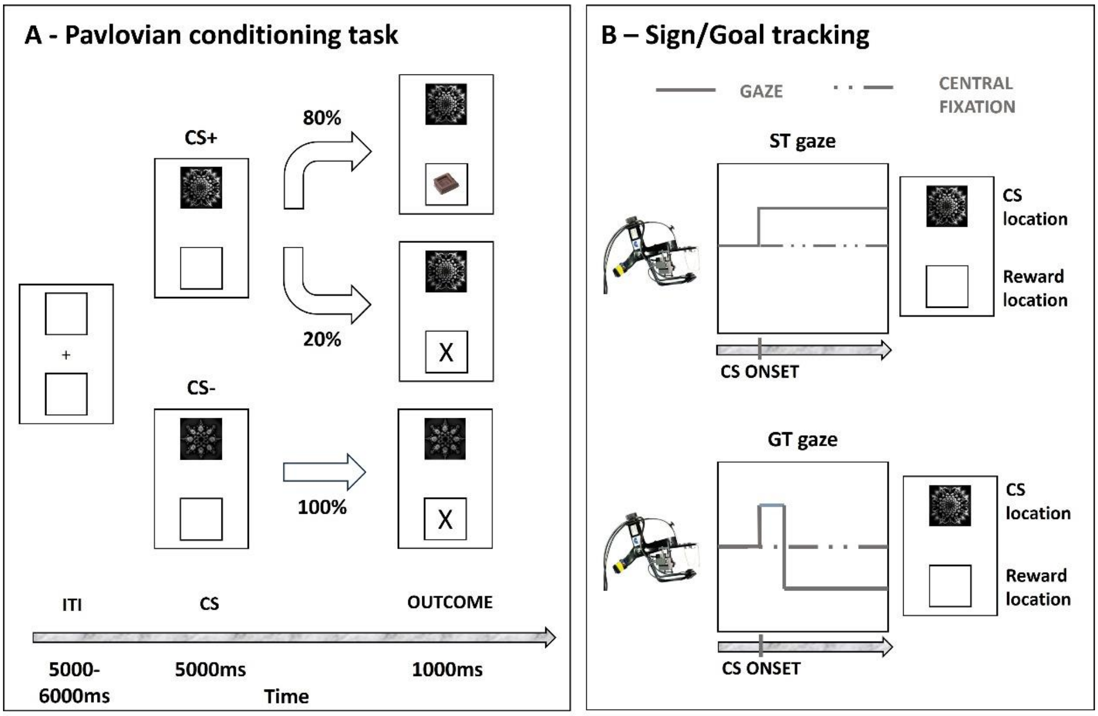

Figure 1: Pavlovian conditioning task and example of sign/goal tracking. A) Task design: each trial began with a 5000-6000ms inter-trial-interval (ITI), followed by the presentation of one of two CS (CS+ or CS-) for 5000ms. The outcome then appeared and remained on display for 1000ms. The CS+ was paired with an individually tailored food reward in 80% of trials and with no reward ("X") in the remaining 20%. The CS- was never paired with a reward (100% "X"). B) Example of sign- and goal-tracking eye-gaze behavior: following CS onset, ST direct their gaze toward the CS location. In contrast, GT initially direct their gaze toward the CS location and then quickly shift toward the reward delivery location.

2.4 Eye-tracking recording and preprocessing

Fixation time (horizontal and vertical) during the Pavlovian conditioning task was recorded using a binocular 2D eye-tracking system (Chronos Vision GmbH, Berlin, Germany) with a sampling rate of 100 Hz. The device utilized two remote cameras to track pupil profiles via online digital image processing (2D, 11-bit output range). Infrared LEDs emitting at 940 nm facilitated both online and offline visualization of eye movements. The device provided horizontal and vertical measurements ranging from −40° to +40°, with a resolution of <0.05° and a measurement error of less than 0.2°.

Eye-tracking data were offline processed using MATLAB v2024a (The MathWorks, Inc., 2024). Data were recorded from both eyes; however, only data from the eye with the clearer registration (i.e., the one with fewer frame losses) were analyzed during pre-processing. This approach was chosen because eye movements are typically synchronized across both eyes (Carter & Luke, 2020; Hooge et al., 2019).

The raw eye-tracking signal during the Pavlovian conditioning task was segmented into 10-second epochs for each trial, including 5000ms of ITI and 5000ms of CS presentation. From each epoch, the first second of CS presentation was excluded to eliminate the orienting response triggered by the visual stimulus' appearance (Gottlieb, 2012; Pietrock et al., 2019).

Out of the 8 participants excluded from the analysis (see Participants), 6 were removed for reporting less than 70% of the total available data, while 2 were excluded for having 60% or less of available trials.

2.5 Gaze index and sign-trackers/goal-trackers classification

The dwell time, defined as the amount of time during which consecutive fixations remain within the same area of interest (Garofalo & Di Pellegrino, 2015), was computed for the remaining 4000ms of CS presentation and used to calculate the gaze index (see below). The CS and outcome locations served as the two areas of interest (5 cm square), corresponding to the "sign" and the "goal" regions of interest, respectively. The gaze index was computed according to the following formula:

$$(1) \quad Gaze\ index = \frac{Dwell\ time\ on\ Sign - Dwell\ time\ on\ Goal}{Dwell\ time\ on\ Sign + Dwell\ time\ on\ Goal}$$

where the "sign" represents the area of interest around the CS location, and the "goal" represents the area of interest around the outcome location (see Figure 1). This index is bounded between -1 and +1, offering a symmetric and interpretable scale. A value of +1 indicates exclusive fixation on the sign (i.e., 100% of dwell time on the CS location), while a value of -1 reflects exclusive fixation on the goal (i.e., 100% of dwell time on the outcome location). A value of 0 denotes equal allocation of gaze time between the two areas of interest.

This index was calculated during the presentation of the CS+, excluding the first second of each trial, and restricted to the second block, thus including trials where learning was presumed to be already established (Cherkasova et al., 2024; Garofalo & Di Pellegrino, 2015). The classification is then usually based on a median split on such an index, with participants falling in the higher half being classified as sign-trackers while those in the lower half as goal-trackers (Cherkasova et al., 2024; Garofalo & Di Pellegrino, 2015). To confirm that gaze allocation reflects the incentive salience specifically attributed to the CS+, and not just a general attentional bias toward visual cues, the same index was also computed for the CS- and tested against the CS+. Since the CS- was visually identical to the CS+ but lacked any reward association, this comparison allowed us to isolate the motivational value of the CS+. As in previous studies, sign-trackers were expected to show significantly greater gaze attraction to the CS+ than to the CS-, supporting the idea that their attention is captured specifically by the motivational value of the cue rather than by a general attentional bias toward visual stimuli. Examples of ST and GT behavior during the Pavlovian conditioning task are presented in Figure 1B. No participants reported more than 80% of the analyzed epoch outside the CS or US locations (Dinu et al., 2024).

2.6 Statistical analyses

Statistical analyses were conducted using JASP 0.19.3.0 (Love et al., 2019) and RStudio v4.4.2 (R Core Team, 2024) with the following packages: *rstudioapi* (Ushey et al., 2024), *lmtest* (Zeileis & Hothorn, 2002), *openxlsx* (Schauberger & Walker, 2024), *tidyverse* (Wickham et al., 2019), *ggplot2* (Wickham, 2016), *irr* (Gamer et al., 2019), *patchwork* (Pedersen, 2019), and *boot* (Canty & Ripley, 2024; Davison & Hinkley, 1997). Normality of the distribution and heteroscedasticity were evaluated via visual inspection of data distribution and residuals, as well as with Shapiro-Wilk and Breusch-Pagan tests, respectively. Non-parametric statistics were employed for non-normally distributed or heteroscedastic data.

Pavlovian learning assessment

These analyses aimed to test comparable learning between ST and GT measured as a selective increase in liking of the CS+ (vs CS-), at both T1 and T2. To achieve this, two separate Bayesian analyses of variance (ANOVA) were conducted at T1 and T2, with a 2 (CS: CS+, CS-) X 2 (Group: ST, GT) factorial design. The difference between pre- and post-liking of the CS was used as the dependent variable (i.e., values > 0 indicated a higher liking for the CS following Pavlovian conditioning). These analyses evaluated support for the selective increase in CS+ liking compared to CS-, as well as the null hypothesis that posited no differences between ST and GT. The Bayes Factor (BF_{10}) quantified the probability of the data under the alternative hypothesis relative to the null hypothesis (Degni et al., 2022; Kruschke, 2021).

Gaze index distribution and validity

The distribution of the gaze index was inspected and reported for descriptive purposes. Construct validity (specifically, divergent validity) was assessed to verify that gaze behavior toward the CS+ reflected incentive salience rather than a general attentional bias. To this end, we conducted two independent Welch two-sample t-tests, one for each session (T1 and T2), comparing ST and GT classified as previously described. For each participant, the difference between the CS+ and CS- gaze index served as the dependent variable. This approach tested whether the two groups differed in their selective attention to the reward-predictive cue (CS+) relative to a non-rewarded but perceptually identical stimulus (CS-).

Test-retest reliability

Intraclass Correlation Coefficient (ICC), Lin's Concordance Correlation Coefficient (CCC), and Bland-Altman analysis were used to assess the test-retest reliability of the gaze index. The ICC was computed by using a two-way random effects model for a single rater, multiple measurements, and both absolute agreement ($ICC_{agreement}$) and consistency ($ICC_{consistency}$) (Koo & Li, 2016; McGraw & Wong, 1996; Shrout & Fleiss, 1979). The value of ICC ranged from 0 to 1, with a higher value indicating that measurements account for greater true variance than error variance and are therefore more reliable. $ICC_{agreement}$ considered the reliability in absolute terms, penalizing the reliability if systematic error between participants was manifested (e.g., all participants showed higher gaze index at T2 than T1). Conversely, $ICC_{consistency}$ considers only the rank order of the measurements, so reliability remains stable even in the presence of systematic error. A higher value of $ICC_{consistency}$ than $ICC_{agreement}$ suggested the presence of a systematic error (Parsons et al., 2019). The CCC ranged from 0 to 1 and evaluated the precision (i.e., how data follow a linear relation) and the accuracy (i.e., how the correlation line fits with the identity correlation line), considering both casual and systematic error. A higher CCC value represented higher concordance between measurements. Since the data distributions did not fit a normal distribution, a non-parametric bootstrap with 10.000 iterations with adjusted bootstrap percentile (BCa) was computed for both ICC and CCC to estimate 95% confidence intervals (Mehta et al., 2018; Ukoumunne et al., 2003; Williamson et al., 2007). Comparable ICCs and CCC values suggest the absence of systematic error.

The Bland-Altman analysis (and plot) evaluated the agreement between measurements by fitting a simple linear regression model (systematic error line) to investigate the relationship between the difference and the average of the gaze indices at T1 and T2 with 95% confidence intervals around the regression line. When the 95% confidence interval around the regression line include zero (i.e., the difference between T1 and T2 equaled zero), there is absence of systematic bias, i.e., no consistent tendency for values to be higher at T1 than T2, or vice versa (Giavarina, 2015; Koo & Li, 2016; Martin Bland & Altman, 1986; Shrout & Fleiss, 1979; Weir, 2005). Moreover, to predict the 95% confidence intervals of the expected between-session variability, two predictive models were computed around the systematic error line with 95% confidence intervals indicating lower and upper Limits of Agreement (LoAs). Wider LoAs indicate a higher between-session variability and lower agreement. Since heteroscedasticity was observed between the difference and the average of the gaze indices, instead of the standard use of the mean difference and parallel LoAs, we computed the simple linear regression models and the relative predictive values to obtain more robust results (Bland & Altman, 1999; Dewitte et al., 2002; Ludbrook, 2010). Minimum, maximum, and mean values (with 95% confidence intervals) of the systematic error line and the LoAs were reported.

Stability of the sign-trackers/goal-trackers classification

To investigate the stability of the ST and GT classifications, we compared the number of participants assigned to a different group between T1 and T2, both when the median (i.e., participants classified as ST at T1 and as GT at T2, or vice versa) (Colaizzi et al., 2023; Dinu et al., 2024; Garofalo & Di Pellegrino, 2015) and the tertiary (i.e., participants classified as ST at T1 and as either GT or intermediate at T2, or vice versa) (Cherkasova et al., 2024; Dinu et al., 2024; Schad et al., 2019; Schettino et al., 2024) splits were used. When a median split was used, participants with scores above the median were classified as ST, while those below the median were classified as GT. With a tertiary split, participants in the first tertile were classified as ST, and those in the third tertile were classified as GT. Participants in the second tertile were classified as intermediate and excluded from the analyses.

3. Results

3.1 Pavlovian learning assessment

Participants showed selective increase of CS+ liking from pre- to post-Pavlovian conditioning at both T1 (BF_{10} = 6.95*10^4; err% = 4.66) and T2 (BF_{10} = 4.26*10^8; err% = 1.20). This enhancement was comparable in ST and GT at T1 (BF_{10} = 0.24; err% = 0.98) and T2 (BF_{10} = 0.42; err% = 1.4) (Figure 2; Table 1). These findings indicate that both groups learned to discriminate equally well between CS+ and CS− in the Pavlovian conditioning task.

All participants completed at least two blocks of the Pavlovian conditioning task at both time points (T1 and T2) and accurately reported the stimulus-outcome contingencies. At T1, 7 participants required one additional block, and 2 participants required two additional blocks to meet the learning criterion. At T2, 1 participant required one extra block to meet the learning criterion. To exclude possible biases due to learning imbalances, we conducted a supplementary analysis including only those participants who reached the Pavlovian learning criterion within 2 blocks at T1 and T2 (N = 58). Critically, we found the same pattern of results obtained from the original sample, showing a selective increase of CS+ liking from pre- to post-task at T1 (BF_{10} = 1.13*104; err% = 0.92) and T2 (BF_{10} = 6.91*107; err% = 8.25), comparable between ST and GT at T1 (BF_{10} = 0.23; err% = 3.22) and at T2 (BF_{10} = 0.41; err% = 8.25). These results suggest that requiring more blocks to complete the Pavlovian conditioning task didn't influence the explicit pre- to post-CS liking.

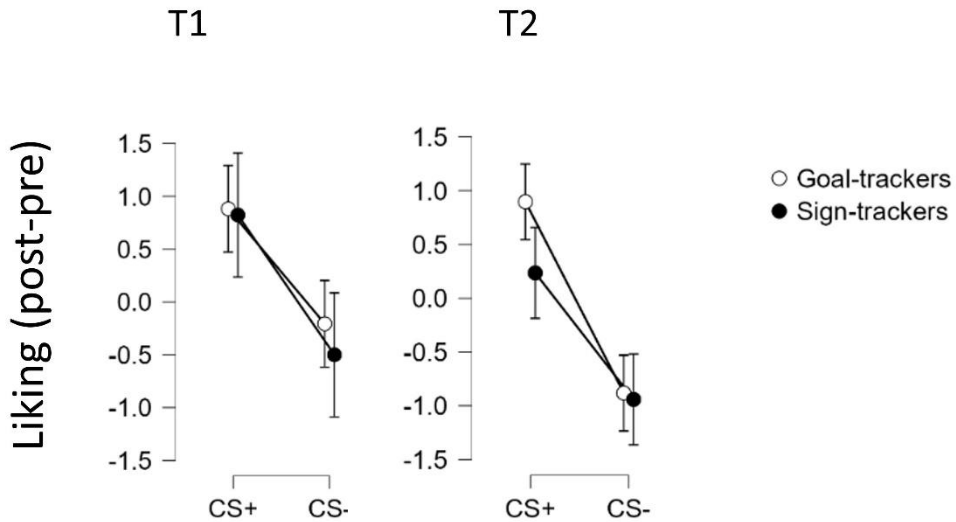

Figure 2: CS liking assessment. The figure displays the change in CS liking (post minus pre) following Pavlovian conditioning task at T1 (left) and T2 (right). Dots represent the mean CS+ and CS- liking ratings with 95% credible intervals.

Table 1: Descriptive statistics for CS liking assessment

			Mean	SD	95% credible interval (lower)	95% credible interval (upper)
T1	CS+	GT	0.88	1.07	0.51	1.26
		ST	0.82	1.27	0.38	1.27
	CS-	GT	-0.21	1.12	-0.60	0.19
		ST	-0.50	1.83	-1.14	0.14
T2	CS+	GT	0.90	1.01	0.55	1.25
		ST	0.24	1.23	-0.20	0.67
	CS-	GT	-0.88	1.25	-1.32	-0.45
		ST	-0.94	1.81	-1.57	-0.31

T1 = test; T2 = retest; CS = conditioned stimulus; ST = sign-trackers; GT = goal-trackers

3.2 Gaze index distribution and validity

The gaze index showed a highly skewed distribution (Figure 3), where many participants clustered near absolute sign-tracking behavior (gaze index = 1), most participants presented positive values indicative of the prevalence of sign-tracking, a few participants presented negative values indicative of the prevalence of goal-tracking, and no participant presented absolute goal-tracking behavior (gaze index = -1). Shapiro-Wilk test (T1: $W = 0.77$, $p < 0.001$; T2: $W = 0.73$, $p < 0.001$) also confirmed that gaze indices at T1 and T2 deviate from a normal distribution. For the median split, the gaze index cut-off scores were 0.79 for T1 and 0.86 for T2. For the tertiary split, the gaze index cut-off scores were 0.67 and 0.95 for T1, and 0.75 and 0.94 for T2.

Construct validity analysis showed that, as compared to goal-trackers, sign-trackers exhibited a positive difference between CS+ and CS- (i.e., higher gaze index for CS+ than CS-) at both T1 ($t_{(50.50)} = 4.88$, $p < .001$, Cohen's $d = 1.18$) and T2 ($t_{(41.26)} = 2.34$, $p = .023$, Cohen's $d = 0.57$). These results confirm that the gaze index reflects a difference in how incentive value modulates attention, rather than a general attentional bias towards visual cues.

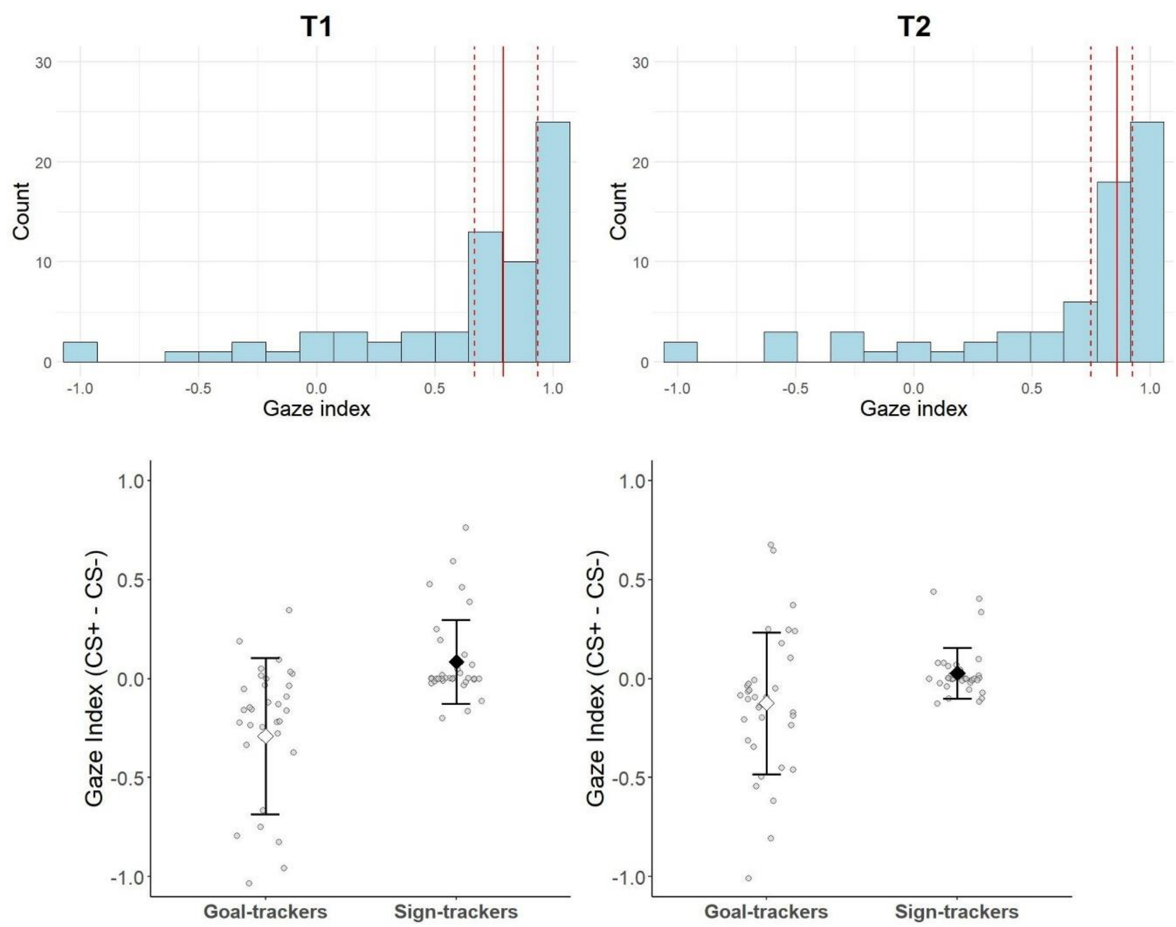

Figure 3. Gaze index distribution and construct validity. The figure displays participants' gaze index during the Pavlovian conditioning task at T1 (left) and T2 (right). The upper panels show the distribution of gaze index values (CS+ only); the x-axis represents gaze index scores, and the y-axis represents the number of participants. The solid vertical line marks the median, and the dotted lines indicate the tertiles. The lower panels present the validity analysis; here, the y-axis represents the difference in gaze index between rewarded and unrewarded trials (CS+ − CS−) in sign-trackers and goal-trackers. White and black diamonds indicate the mean scores of goal-trackers and sign-trackers, respectively, with 95% confidence intervals, while the dots represent individual scores.

3.3 Test-retest reliability

Test-retest reliability analysis indicates that approximately half of the variance in gaze index scores reflected stable individual differences, resulting in suboptimal reliability ($ICC_{agreement}$ = 0.54, 95% confidence intervals [0.30;0.72]; $ICC_{consistency}$ = 0.54, 95% confidence intervals = [0.28;0.71]; CCC = 0.54, 95% confidence intervals = [0.32;0.72]).

Visual inspection of the scatterplot (Figure 4) revealed a good dispersion of data around the identity line, suggesting the absence of a significant systematic behavioral change or measurement error (Berchtold, 2016). Additionally, the presence of comparable $ICC_{agreement}$, $ICC_{consistency}$, and CCC values indicates that adjusting for systematic error or comparing data variability with the identity line does not alter these findings. This key element ensures that measurement error remains random rather than systematic between T1 and T2, thereby not preferentially distorting any particular subgroup or trajectory (Berchtold, 2016; Rousson et al., 2002; Weir, 2005).

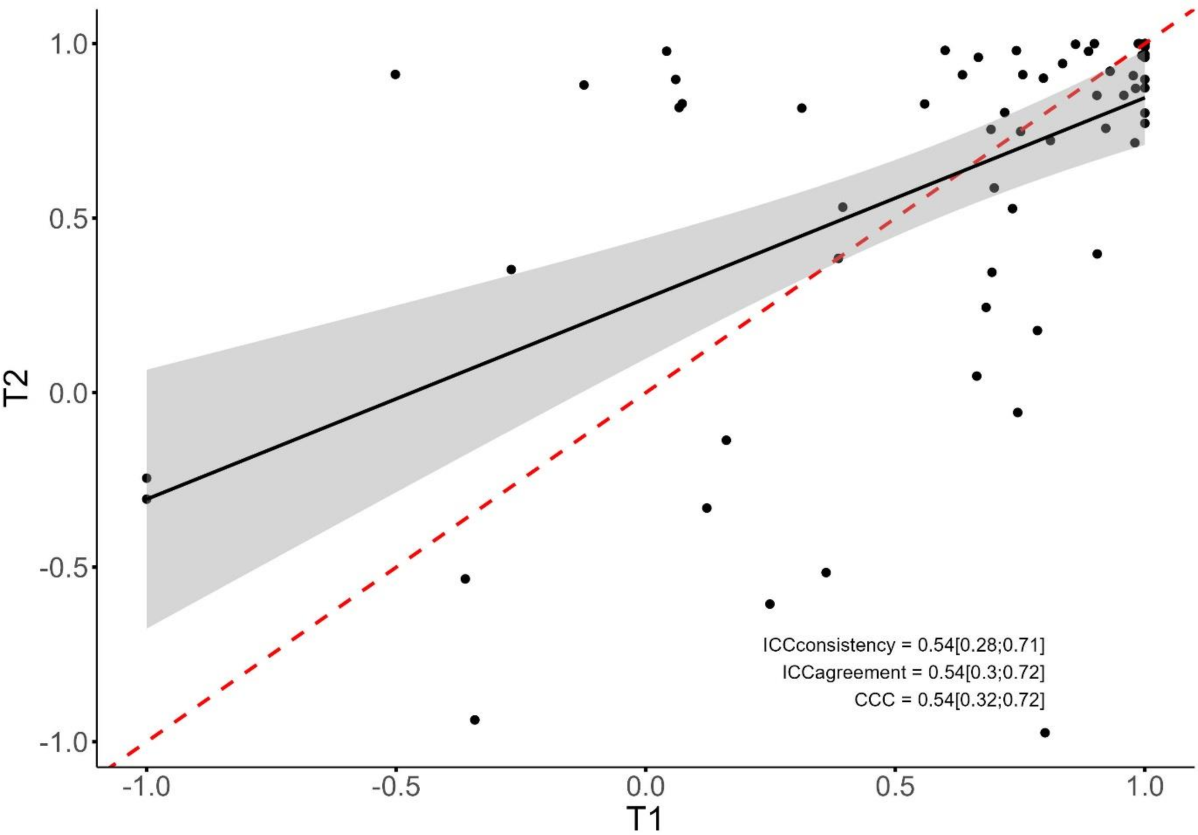

Figure 4: Scatterplot of gaze index test (T1) and retest (T2). This scatterplot presents the relationship between the test (T1, x-axis) and retest (T2, y-axis) gaze indices. Black dots represent individual scores. The solid black line represents the linear regression line fitted to the data, and the grey-shaded region represents the 95% confidence intervals. The red dashed line denotes the identity correlation line.

For the Bland-Altman analysis, data distributions of the difference and the mean between T1 and T2 were non-normally distributed (Difference: $W = 0.92$, $p < 0.001$; Mean: $W = 0.80$, $p < 0.001$) and presented heteroscedasticity ($BP = 17.59$, $p < 0.001$), hence more robust analyses were performed (see Statistical analysis). The Bland-Altman plot (Figure 5) reveals a mean difference between T1 and T2 around zero, with the zero-line included within the 95% confidence intervals of the regression line (Table 2, $\beta = -0.02$, $p = 0.57$), suggesting an overall agreement between the scores measured in the two sessions and the absence of systematic error.

Overall, these results indicate an absence of systematic behavioral change or measurement error but suboptimal overall reliability, with ICC (both consistency and agreement) and CCC approximating 0.5. These estimates imply that roughly 50% of the observed variance can be attributed to true differences among participants, while the remaining variability is due to random noise (Koo & Li, 2016). Such levels of reliability are considered suboptimal as they imply a high degree of random noise, which can distort effect estimates and limit the replicability of findings (Koo & Li, 2016; Loken & Gelman, 2017). This substantial random noise can obscure true effects, leading to difficulties in distinguishing genuine individual differences from the variability introduced by measurement imprecision. As a result, any interpretation of group differences or correlations involving the gaze index must account for the possibility that this inherent error attenuates observed effects. Notably, however, the Bland-Altman plot (Figure 5; Table 2) clearly shows much narrower LoAs for higher gaze index scores, corresponding to sign-tracking behavior, and considerably wider LoAs for lower scores, indicative of goal-tracking behavior. Narrower LoAs reflect reduced variability and greater measurement precision across sessions, thereby suggesting that sign-tracking behavior is captured more reliably by the gaze index than goal-tracking behavior. Furthermore, the "v-shaped" pattern of data observed in Figure 5 supports the interpretation that sign-tracking behavior is more stable than goal-tracking behavior between T1 and T2, and that the gaze index measure may be more reliable in capturing sign-tracking than goal-tracking behavior (Giavarina, 2015; Ludbrook, 2010).

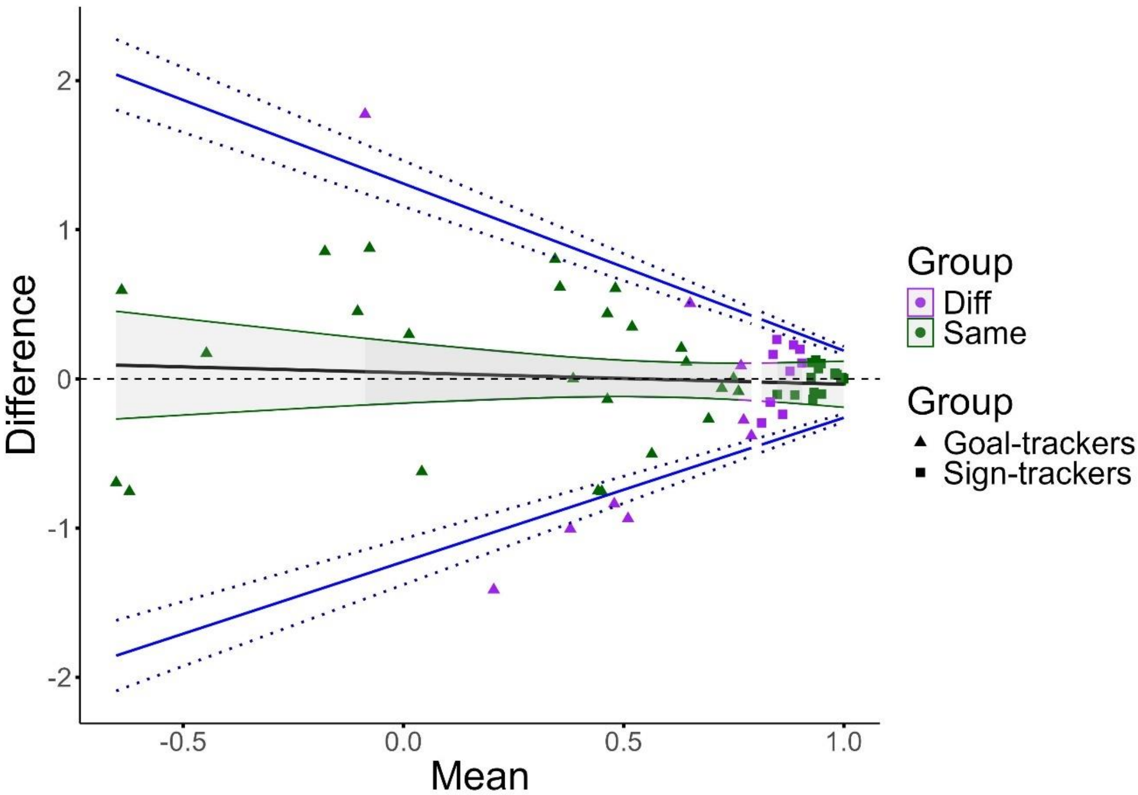

Figure 5: Bland-Altman plot. This Bland-Altman plot illustrates the agreement between the T1 and T2 scores. The y-axis displays the difference between T1 and T2. The x-axis shows the mean

between T1 and T2. The shape of dots denotes individual measurements for ST (square) and GT (triangle) based on the median split of the mean gaze index. Colored dots indicate the stability (green) and instability (purple) of the group assignment between T1 and T2 sessions. The dashed black line depicts the zero-systematic error line (i.e., no difference between T1 and T2). The solid black line represents the linear regression line fitted to the data representing the bias measurement, and the grey-shaded region represents the 95% confidence intervals around this regression line. The solid blue lines represent the limits of agreement, and the dotted line denotes the 95% confidence intervals around the limits of agreement.

Table 2 – Systematic error and limits of agreement

	Mean (95% confidence intervals)	Min (95% confidence intervals)	Max (95% confidence intervals)
Systematic error	-0.01 [-0.04; 0.09]	-0.04 [-0.19; 0.12]	0.09 [-0.27; 0.45]
Upper LoA	0.61 [0.19; 2.04]	0.19 [0.16; 0.22]	2.04 [1.80; 2.28]
Lower LoA	-0.62 [-0.26; -1.85]	-1.86 [-1.62; -2.09]	-0.26[-0.23; -0.29]

LoA = Limits Of Agreement

3.4 Stability of the sign-trackers/goal-trackers classification

The stability of group classification between T1 and T2, based on both the median and tertiary splits, is reported in Table 3. The results indicate a higher stability of the classification between T1 and T2 when the median split classification is used, as compared to the tertiary split classification. In addition, the tertiary split classification reveals an asymmetry in stability between ST and GT: participants initially classified as GT appear more prone to change over time than those initially classified as ST. This may suggest that, in line with the previous analysis, individuals at the lower end of the gaze index distribution exhibit more variability in gaze behavior, or that classification thresholds near the lower tertile cut-off are more sensitive to minor fluctuations.

Table 3 – Stability of the sign-trackers/goal-trackers classification

	Median split		Tertiary split	
Group	**Same**	**Different**	**Same**	**Different**
Overall	50 (73.5%)	18 (26.5%)	27 (58.7%)	19 (41.3%)
Sign-trackers	25 (73.5%)	9 (26.5%)	15 (65.2%)	8 (34.8%)
Goal-trackers	25 (73.5%)	9 (26.5%)	12 (52.2%)	11 (47.8%)

Note: the table reports the number and percentage (in brackets) of participants assigned to the same or different group between T1 and T2.

4. Discussion

4.1 Gaze index validity and reliability

This study aimed to evaluate the psychometric properties of a widely used measure for classifying ST and GT behavior in humans based on eye gaze (Garofalo & Di Pellegrino, 2015). In line with previous results (Garofalo & Di Pellegrino, 2015; Schad et al., 2019; Schettino et al., 2024), preliminary analyses confirmed comparable levels of Pavlovian learning between the two groups and controlled for construct validity by ensuring that the gaze index did not merely reflect an attentional bias. Test-retest reliability analyses on the continuous gaze index indicated an absence of systematic behavioral change or measurement error, but overall suboptimal reliability, which was particularly related to low scores. More precisely, the variability of the score across the two sessions (T1 and T2) increased as the gaze index moved from highly positive values (sign-tracking prevalence) toward highly negative values (goal-tracking prevalence), suggesting that this gaze index may be more reliable in capturing sign-tracking than goal-tracking behavior (Giavarina, 2015; Ludbrook, 2010). This result aligns with the pattern emerging from the classification stability analysis, where an imbalance was reported when using the tertiary split. Specifically, individuals initially identified as goal-trackers showed a greater tendency to shift classification over time than those classified as sign-trackers. This asymmetry may either suggest that goal-tracking behavior is more susceptible to temporal fluctuations or that it is less consistently captured by the gaze index, possibly due to higher estimated measurement error (Atkinson & Nevill, 1998; Koo & Li, 2016). Taken together, these findings suggest that the gaze index is a more robust and temporally stable indicator for sign-tracking than goal-tracking.

Importantly, such discrepancies in measurement reliability could be mitigated if the true effects under investigation are strong and clearly defined, as the impact of low measurement reliability on their detection would be substantially reduced, even though in specific situations it may lead to a spurious overestimation of the effect (Loken & Gelman, 2017). A strong effect produces a powerful indicator that stands out against background noise, so even if the measure contains a significant amount of random error, the true effect can still clearly emerge. Conversely, when effects are weak, meaningful differences may be obscured, thereby limiting both the replicability and the interpretability of the findings. This issue is especially relevant when studying subtle behavioral or neuropsychological differences, where real effects can be quite small and more easily hidden by measurement error (Hedge et al., 2018; Loken & Gelman, 2017). To the best of our knowledge, only one previous study measured the stability of a different ST and GT classification based on a reaction time index extracted from the Value-Modulated Attentional Capture (VMAC) paradigm (Duckworth et al., 2022). In this task, a stimulus target was presented in a specific location, and participants were required to quickly press the button corresponding to the target location, while a distractor stimulus was presented, signalling the amount of reward at stake. Sign-trackers were defined using a tertiary split on reaction time, assuming that slower responses would reflect higher attraction to the distractor stimulus and thus sign-tracking behavior. The authors found higher stability for the ST (50%) than for the GT (30%) classification. However, these results were limited by the small sample size, which included only 6 ST, 4 intermediate, and 10 GT.

4.2 Gaze index distribution

The data distribution of the gaze index conveys important considerations. In line with other studies (Cherkasova et al., 2024; Colaizzi et al., 2023; Schettino et al., 2024), the gaze index was predominantly high and clustered near 1, with only a few participants showing negative values, thus denoting a high propensity to manifest sign-tracking behavior.

The use of median splits on skewed or narrowly distributed data can lead to artificial groupings that exaggerate or obscure true individual differences, limit cross-study comparability, and may misrepresent the continuous nature of the underlying construct. These are particularly problematic in human ST and GT research because the gaze index distribution rarely approximates a symmetric (e.g., normal or binomial) distribution with enough spread to distinguish sign-tracking from goal-tracking. Ideally, ST and GT groups should be expected to occupy distinct regions of the scale, with highly positive values for sign-tracking and highly negative values for goal-tracking, but this separation only holds when the underlying distribution provides sufficient variability around zero (Cohen, 1983; MacCallum et al., 2002). Of note, studies that reported more symmetric distributions (Cherkasova et al., 2024; Dinu et al., 2024) did not report significant differences compared to the rest of the literature. In particular, Cherkasova and colleagues (2024) directly investigated this issue by adding a second experiment in which a consummatory response was required to obtain the reward. Despite producing a more symmetric distribution in the eye gaze index, neither experiment provided support for a link between sign-tracking and risk-taking propensity, as hypothesized. In general terms, the absence of large-scale studies unravelling the true distribution of ST and GT in the human population renders all choices inherently arbitrary, increasing the risk of misclassification.

Nevertheless, considering the gaze index as a continuous variable does not fully resolve the problem, since skewed distributions introduce their own limitations. When the majority of participants cluster within a narrow range of values, the effective variability of the measure becomes restricted, reducing statistical power to detect associations with external variables and increasing the influence of a small number of extreme scores. Such outliers may disproportionately affect estimates of reliability and inflate error variance, ultimately compromising the stability of the construct (Enkavi et al., 2019; Hedge et al., 2018; Pennington et al., 2025; Zorowitz & Niv, 2023). Moreover, skewness can attenuate correlations with other measures, since classical parametric tests assume a roughly symmetric distribution of errors and may underestimate true effect sizes when distributions deviate substantially from normality (Enkavi et al., 2019; Hedge et al., 2018; Pennington et al., 2025; Zorowitz & Niv, 2023). Finally, a skewed continuous index complicates the interpretability of intermediate values: if most individuals fall just above or below zero, the distinction between putative “intermediate trackers” and noise becomes blurred, undermining the ability to make meaningful psychological inferences.

These limitations highlight the need for measurement models that explicitly account for distributional shape and error structure, and underscore the importance of mapping the true population distribution of sign- and goal-tracking tendencies before drawing strong theoretical conclusions.

Until more precise data on the population distribution is available, deviations from the statistical ideal do not preclude meaningful analysis or interpretation. For instance, Colaizzi and colleagues (2023), encountering a comparable distribution, designated the low-score group as “non-ST” rather than “GT,” thereby explicitly recognizing the absence of a well-defined goal-tracking phenotype. While this solution does not eliminate the underlying measurement challenges, it illustrates a pragmatic strategy for safeguarding interpretative validity without imposing categorical distinctions unsupported by the empirical distribution.

4.3 Alternative approaches

To mitigate these limitations, alternative computational and experimental approaches may be helpful.

Among the possible approaches to index computation, some studies have adopted unsupervised machine-learning methods for data-driven classification (Cope et al., 2023; Versace et al., 2016). This methodology takes advantage of the full data distribution to identify naturally occurring clusters without relying on arbitrary splits. Unsupervised machine learning methods, such as latent profile and cluster analysis, enable researchers to detect latent patterns that reflect genuine population differences, thereby reducing classification errors and leading to more precise subgroup definitions (Spurk et al., 2020). For instance, Cope and colleagues (2023) employed a physical Pavlovian conditioning task in which participants interacted with both a lever (representing the sign) and a magazine (representing the goal). Then, a latent profile analysis was conducted on standardized measures of magazine gaze, lever gaze, and lever presses, thereby grouping participants into ST, intermediate, and GT categories. Although their design, reporting reliable fit indices with as few as 30 participants, demonstrated a translational approach from rodent to human behavior, the latent profile analysis resulted in the exclusion of 68% of participants (the intermediate group) and yielded unbalanced groups, with approximately 20% ST and 12% GT. In contrast, simulation studies suggest that a minimum of 500 participants is necessary to achieve precise and reliable grouping (Nylund et al., 2007; Spurk et al., 2020). Considering the large sample sizes required and the high rate of participant exclusion, the feasibility of such a method becomes challenging. Versace and colleagues (2016) classified lean and obese participants as ST or GT based on their late positive potentials (LPP) measured via electroencephalography, while presenting pictures with varying emotional values. In this case, higher LPP amplitudes for food-related cues were assumed to indicate sign-tracking behavior. However, the resulting classification was unbalanced (32% ST, 68% GT) and, as noted by Colaizzi and colleagues (2020), this paradigm did not allow participants to be grouped based on evidence of learned associations, highlighting a critical divergence from animal paradigms. It is worth noting that although highly unbalanced groups may lead to statistical comparability issues due to differing sample sizes, the observed distribution of ST and GT may reflect their actual prevalence in the general population. Nevertheless, the behavioral and neuropsychological bases of ST and GT in humans remain debated (Flagel et al., 2008; Robinson & Flagel, 2009), and relying exclusively on these methods carries the risk of misclassifying participants (Dy & Brodley, 2004; Ye et al., 2024). Without a strong theoretical framework specifying which behavioral or neuropsychological variables truly capture incentive salience or distinct learning systems, clustering algorithms risk being misled by spurious correlations or redundant variables. Overweighting such features, especially when behavioral measures are highly correlated due to shared variance, can lead to statistically distinct but conceptually meaningless clusters. As a result, classifications may reflect random fluctuations or chance associations rather than genuine differences in underlying psychological processes (Spurk et al., 2020).

Novel experimental paradigms may also help overcome the previously discussed limitations, particularly in detecting goal-tracking behavior. One limitation of the current experimental paradigm is that it allows for the computation of the gaze index within a time window (the last 3 seconds of CS presented alone) that presents visual competition between a complex fractal image (CS) and a simple blank square (US location). Although the observation of a higher gaze index for CS+ than for CS- only in ST speaks in favor of an absence of an attentional bias, this imbalance may bias the gaze toward the more visually salient sign location in both ST and GT. Although some evidence in this sense already exists (Cherkasova et al., 2024; Cope et al., 2023; Garofalo & Di Pellegrino, 2015), future studies could directly test whether increasing the relevance of the US location or inserting a more direct measure of US collection could compensate for this issue.

4.4 Stable traits vs state-dependent behaviors

A key question in interpreting ST and GT classifications concerns whether these groups reflect stable, trait-like characteristics or more transient, state-dependent patterns. In our study, we observed substantial variability in gaze index scores between T1 and T2. Although this instability may partly stem from measurement imprecision (Hedge et al., 2018), it could also indicate that sign-tracking and goal-tracking tendencies are not fixed traits but instead fluctuate in response to situational or contextual factors. If this is the case, ST and GT behaviors may reflect state-dependent processes that change with momentary motivational, attentional, or affective states (Volkow et al., 2016). This interpretation is also consistent with what is observed in animal literature. Indeed, although these behavioral phenotypes are often considered stable, some animals shift between sign-tracking and goal-tracking behavior across experimental sessions and under specific conditions. For example, when the CS is presented as a diffuse auditory cue, animals originally classified as ST manifested a switch to the GT group (Meyer et al., 2014), whereas under reward uncertainty, those beginning as GT may shift toward ST (Robinson et al., 2015). Importantly, individuals with extreme gaze index scores (i.e., approximating +1, absolute sign-tracking, or near -1, absolute goal-tracking) demonstrated greater consistency across sessions, whereas the highest degree of fluctuation was observed among participants with intermediate scores. This pattern suggests that while the gaze index may be unreliable in classifying individuals with ambiguous or mixed behavioral tendencies, it may still capture relatively stable, trait-like differences in those who display clear sign-tracking or goal-tracking behavior. In other words, although the data do not provide conclusive evidence, they suggest the possibility that extreme ST and GT scores might reflect more stable individual characteristics (Flagel et al., 2011; Robinson & Flagel, 2009). In other words, while the greater consistency among individuals with extreme scores (particularly ST) may suggest that the behavior may indeed reflect stable individual traits, the marked fluctuations observed among those with intermediate scores could indicate that motivational, attentional, or affective states also play a significant role.

That said, stronger empirical support is needed before drawing firm conclusions. Future research should aim to establish clearer associations between ST and GT profiles and other reliable trait markers, such as impulsivity or reward sensitivity, to validate their trait-like nature (Felix & Flagel, 2024; Robinson & Flagel, 2009). Furthermore, a more comprehensive understanding of the distribution of ST and GT tendencies in the general population is necessary to interpret individual differences meaningfully and to refine classification thresholds accordingly.

5. Conclusions

In conclusion, this study provides crucial information on the test-retest reliability of sign-tracking and goal-tracking behavior, as well as the stability of classifications for the ST and GT groups.

The findings suggest that the commonly used gaze index is valid and effective in reliably capturing sign-tracking behavior; however, its sensitivity is limited when it comes to consistently detecting goal-tracking behavior. Taken together, these results suggest two potential scenarios. The first scenario concerns the characteristics of the population: goal-tracking behavior may be very rare in the population, meaning that larger samples or a targeted sampling strategy could be required to detect it. Also, whether sign-tracking and goal-tracking should be considered stable, trait-like characteristics or more transient and state-dependent behaviors is yet to be clarified. The second scenario concerns measurement limitations: the currently used gaze index may be an inherently precise and reliable measure for capturing sign-tracking behavior, but it might not effectively reflect goal-tracking behavior. In this case, different computational or experimental approaches may be necessary to capture goal-tracking behavior more accurately. Addressing these challenges through methodological refinements and broader population-level data will be essential for improving the interpretability and replicability of future research in this field.

Declarations

Funding

The authors of this work are supported by the following grants:

- European Union's NextGenerationEU and the Italian Ministry of Research (MUR) National Recovery and Resilience Plan (NRRP-PNRR) Progetti di Rilevante Interesse Nazionale (PRIN) 2022 grant, project titled “Individual variations in resisting temptation: the balance between motivational and inhibitory control” (ES PNRR - M4C2 - I1.3 - Prot. P2022KAZ45_001 – CUP J53D23017240001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

- European Union's NextGenerationEU and the Italian Ministry of Research (MUR) National Recovery and Resilience Plan (NRRP-PNRR) Progetti di Rilevante Interesse Nazionale (PRIN) 2022 grant, project titled “Decision-making through the lens of the motor system: a neurocomputational perspective” (Prot. 2022W5TTK8 - CUP J53D23019450006).

- European Union's NextGenerationEU and the Italian Ministry of Research (MUR) National Recovery and Resilience Plan (NRRP-PNRR), project titled “MNESYS” (N. PE0000006; DN. 1553 11.10.2022)

- Italian Ministry of Research (MUR) Progetti di Rilevante Interesse Nazionale (PRIN) 2020, project titled “Adaptive and multidimensional assessment of executive functions” (Prot. 20209WKCLL)

Conflict of interest

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study was approved by the Bioethics Committee of the University of Bologna.

Consent to participate

Informed consent was obtained from all individual participants included in the study.

Consent to publish

Not applicable

Data, materials and code availability

The datasets analyzed during the current study are shared according to FAIR principles in the OSF repository, https://osf.io/fvhqk/.

Author contribution

Author contribution statement follows the CRediT standard:

Marco Badioli: conceptualization, data curation, formal analysis, investigation, methodology, software, writing - original draft, visualization

Claudio Danti: formal analysis, investigation, software, writing – review & editing

Luigi Degni: formal analysis, investigation, software, writing – review & editing

Gianluca Finotti: conceptualization, formal analysis, software, writing – review & editing

Valentina Bernardi: investigation, writing – review & editing

Lorenzo Mattioni: writing, review & editing

Francesca Starita: writing – review & editing, funding acquisition

Sara Giovagnoli: methodology; supervision

Giuseppe di Pellegrino: conceptualization, supervision, funding acquisition

Mariagrazia Benassi: supervision

Garofalo Sara: conceptualization, funding acquisition, project administration, resources, methodology supervision, writing – review & editing, visualization

Acknowledgment

We thank Lucia Cortese, Benedetta Luciani, Andrea Savelli, Maico Valicelli and Alessia Del Conte for their collaboration in data acquisition.

Open Practice Statements

The data and materials for all experiments are available at https://osf.io/fvhqk/. The experiment was not preregistered.

References

Anselme, P., & Robinson, M. J. F. (2020). From sign-tracking to attentional bias: Implications for gambling and substance use disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *99*, 109861. https://doi.org/10.1016/j.pnpbp.2020.109861

Anselme, P., Robinson, M. J. F., & Berridge, K. C. (2013). Reward uncertainty enhances incentive salience attribution as sign-tracking. *Behavioural Brain Research*, *238*, 53–61. https://doi.org/10.1016/j.bbr.2012.10.006

Atkinson, G., & Nevill, A. M. (1998). Statistical Methods For Assessing Measurement Error (Reliability) in Variables Relevant to Sports Medicine. *Sports Medicine*, *26*(4), 217–238. https://doi.org/10.2165/00007256-199826040-00002

Badioli, M., Degni, L. A. E., Dalbagno, D., Danti, C., Starita, F., di Pellegrino, G., Benassi, M., & Garofalo, S. (2024). Unraveling the influence of Pavlovian cues on decision-making: A pre-registered meta-analysis on Pavlovian-to-instrumental transfer. *Neuroscience & Biobehavioral Reviews*, *164*, 105829. https://doi.org/10.1016/j.neubiorev.2024.105829

Berchtold, A. (2016). Test–retest: Agreement or reliability? *Methodological Innovations*, *9*, 2059799116672875. https://doi.org/10.1177/2059799116672875

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Brown, P. L., & Jenkins, H. M. (1968). Auto-Shaping of the Pigeon’s Key-Peck. *Journal of the Experimental Analysis of Behavior*, *11*(1), 1–8. https://doi.org/10.1901/jeab.1968.11-1

Campus, P., Accoto, A., Maiolati, M., Latagliata, C., & Orsini, C. (2016). Role of prefrontal 5-HT in the strain-dependent variation in sign-tracking behavior of C57BL/6 and DBA/2 mice. *Psychopharmacology*, *233*(7), 1157–1169. https://doi.org/10.1007/s00213-015-4192-7

Canty, A., & Ripley, B. (2024). *boot: Bootstrap R (S-Plus) Functions* (Versione 1.3-31) [Software]. https://CRAN.R-project.org/package=boot

Carter, B. T., & Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology*, *155*, 49–62. https://doi.org/10.1016/j.ijpsycho.2020.05.010

Cherkasova, M. V., Clark, L., Barton, J. J. S., Stoessl, A. J., & Winstanley, C. A. (2024). Risk-promoting effects of reward-paired cues in human sign- and goal-trackers. *Behavioural Brain Research*, *461*, 114865. https://doi.org/10.1016/j.bbr.2024.114865

Cohen, J. (1983). The Cost of Dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253. https://doi.org/10.1177/014662168300700301

Colaizzi, J. M., Flagel, S. B., Gearhardt, A. N., Borowitz, M. A., Kuplicki, R., Zotev, V., Clark, G., Coronado, J., Abbott, T., & Paulus, M. P. (2023). The propensity to sign-track is associated with externalizing behavior and distinct patterns of reward-related brain activation in youth. *Scientific Reports*, *13*(1), 4402. https://doi.org/10.1038/s41598-023-30906-3

Colaizzi, J. M., Flagel, S. B., Joyner, M. A., Gearhardt, A. N., Stewart, J. L., & Paulus, M. P. (2020). Mapping sign-tracking and goal-tracking onto human behaviors. *Neuroscience & Biobehavioral Reviews*, *111*, 84–94. https://doi.org/10.1016/j.neubiorev.2020.01.018

Cope, L. M., Gheidi, A., Martz, M. E., Duval, E. R., Khalil, H., Allerton, T., & Morrow, J. D. (2023). A mechanical task for measuring sign- and goal-tracking in humans: A proof-of-concept study. *Behavioural Brain Research*, *436*, 114112. https://doi.org/10.1016/j.bbr.2022.114112

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (Reprinted with corrections 2003). Cambridge University Press.

Degni, L. A. E., Dalbagno, D., Starita, F., Benassi, M., di Pellegrino, G., & Garofalo, S. (2022). General Pavlovian-to-instrumental transfer in humans: Evidence from Bayesian inference. *Frontiers in Behavioral Neuroscience*, *16*. https://doi.org/10.3389/fnbeh.2022.945503

Degni, L. A. E., Danti, C., Finotti, G., Garofalo, S., & di Pellegrino, G. (2024). *Pavlovian bias instigates suboptimal choices*. PsyArXiv. https://doi.org/10.31234/osf.io/qdru7

Degni, L. A. E., & Garofalo, S. (2025). Toward a Translational Model of Sex-Associated Pavlovian Phenotypes. *Addiction Biology*, *30*(6). https://doi.org/10.1111/adb.70054

Degni, L. A. E., Garofalo, S., Finotti, G., Starita, F., Robbins, T. W., & Di Pellegrino, G. (2024). Sex differences in motivational biases over instrumental actions. *Npj Science of Learning*, *9*(1), 62. https://doi.org/10.1038/s41539-024-00246-6

Degni, L. A. E., Mattioni, L., Danti, C., Bernardi, V., Finotti, G., Badioli, M., Starita, F., & Garofalo, S. (2024). *The cost of Pavlovian bias: Maladaptive decision-making in human sign-trackers and goal-trackers*. PsyArXiv. https://doi.org/10.31234/osf.io/eqnjd

Dewitte, K., Fierens, C., Stöckl, D., & Thienpont, L. M. (2002). Application of the Bland–Altman Plot for Interpretation of Method-Comparison Studies: A Critical Investigation of Its Practice. *Clinical Chemistry*, *48*(5), 799–801. https://doi.org/10.1093/clinchem/48.5.799

Dickson, P. E., McNaughton, K. A., Hou, L., Anderson, L. C., Long, K. H., & Chesler, E. J. (2015). Sex and strain influence attribution of incentive salience to reward cues in mice. *Behavioural Brain Research*, *292*, 305–315. https://doi.org/10.1016/j.bbr.2015.05.039

Dinu, L.-M., Georgescu, A.-L., Singh, S. N., Byrom, N. C., Overton, P. G., Singer, B. F., & Dommett, E. J. (2024). Sign-tracking and goal-tracking in humans: Utilising eye-tracking in clinical and non-clinical populations. *Behavioural Brain Research*, *461*, 114846. https://doi.org/10.1016/j.bbr.2024.114846

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, *11*(4), 410–416. https://doi.org/10.1038/nn2077

Duckworth, J. J., Wright, H., Christiansen, P., Rose, A. K., & Fallon, N. (2022). Sign-tracking modulates reward-related neural activation to reward cues, but not reward feedback. *European Journal of Neuroscience*, *56*(7), 5000–5013. https://doi.org/10.1111/ejn.15787

Dy, J., & Brodley, C. (2004). *Feature Selection for Unsupervised Learning*. *5*, 845–889.

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477. https://doi.org/10.1073/pnas.1818430116

Everitt, B. J., & Robbins, T. W. (2016). Drug Addiction: Updating Actions to Habits to Compulsions Ten Years On. *Annual Review of Psychology*, *67*(1), 23–50. https://doi.org/10.1146/annurev-psych-122414-033457

Felix, P. C., & Flagel, S. B. (2024). Leveraging individual differences in cue–reward learning to investigate the psychological and neural basis of shared psychiatric symptomatology: The sign-tracker/goal-tracker model. *Behavioral Neuroscience*, *138*(4), 260–271. https://doi.org/10.1037/bne0000590

Finke, J. B., Roesmann, K., Stalder, T., & Klucken, T. (2021). Pupil dilation as an index of Pavlovian conditioning. A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *130*, 351–368. https://doi.org/10.1016/j.neubiorev.2021.09.005

Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., Willuhn, I., Akers, C. A., Clinton, S. M., Phillips, P. E. M., & Akil, H. (2011). A selective role for dopamine in stimulus–reward learning. *Nature*, *469*(7328), 53–57. https://doi.org/10.1038/nature09588

Flagel, S. B., Robinson, T. E., Clark, J. J., Clinton, S. M., Watson, S. J., Seeman, P., Phillips, P. E. M., & Akil, H. (2010). An Animal Model of Genetic Vulnerability to Behavioral Disinhibition and Responsiveness to Reward-Related Cues: Implications for Addiction. *Neuropsychopharmacology*, *35*(2), 388–400. https://doi.org/10.1038/npp.2009.142

Flagel, S. B., Watson, S. J., Akil, H., & Robinson, T. E. (2008). Individual differences in the attribution of incentive salience to a reward-related cue: Influence on cocaine sensitization. *Behavioural Brain Research*, *186*(1), 48–56. https://doi.org/10.1016/j.bbr.2007.07.022

Fleeson, W., & Jayawickreme, E. (2015). Whole Trait Theory. *Journal of Research in Personality*, *56*, 82–92. https://doi.org/10.1016/j.jrp.2014.10.009

Fraser, K. M., & Holland, P. C. (2019). Occasion setting. *Behavioral Neuroscience*, *133*(2), 145–175. https://doi.org/10.1037/bne0000306

Gamer, G., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement* (Versione 0.84.1) [Software]. https://CRAN.R-project.org/package=irr

Garofalo, S., Battaglia, S., & di Pellegrino, G. (2019). Individual differences in working memory capacity and cue-guided behavior in humans. *Scientific Reports*, *9*(1), 7327. https://doi.org/10.1038/s41598-019-43860-w

Garofalo, S., & Di Pellegrino, G. (2015). Individual differences in the influence of task-irrelevant Pavlovian cues on human behavior. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00163

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, *25*(2), 141–151. https://doi.org/10.11613/BM.2015.015

Gottlieb, J. (2012). Attention, Learning, and the Value of Information. *Neuron*, *76*(2), 281–295. https://doi.org/10.1016/j.neuron.2012.09.034

Heck, M., Durieux, N., Anselme, P., & Quertemont, E. (2024). Implementations of sign- and goal-tracking behavior in humans: A scoping review. *Cognitive, Affective, & Behavioral Neuroscience*. https://doi.org/10.3758/s13415-024-01230-8

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hogarth, L., & Duka, T. (2006). Human nicotine conditioning requires explicit contingency knowledge: Is addictive behaviour cognitively mediated? *Psychopharmacology*, *184*(3), 553–566. https://doi.org/10.1007/s00213-005-0150-0

Hooge, I. T. C., Holleman, G. A., Haukes, N. C., & Hessels, R. S. (2019). Gaze tracking accuracy in humans: One eye is sometimes better than two. *Behavior Research Methods*, *51*(6), 2712–2721. https://doi.org/10.3758/s13428-018-1135-3

Joyner, M. A., Gearhardt, A. N., & Flagel, S. B. (2018). A Translational Model to Assess Sign-Tracking and Goal-Tracking Behavior in Children. *Neuropsychopharmacology*, *43*(1), 228–229. https://doi.org/10.1038/npp.2017.196

Keefer, S. E., Bacharach, S. Z., Kochli, D. E., Chabot, J. M., & Calu, D. J. (2020). Effects of Limited and Extended Pavlovian Training on Devaluation Sensitivity of Sign- and Goal-Tracking Rats. *Frontiers in Behavioral Neuroscience*, *14*. https://doi.org/10.3389/fnbeh.2020.00003

King, C. P., Palmer, A. A., Woods, L. C. S., Hawk, L. W., Richards, J. B., & Meyer, P. J. (2016). Premature responding is associated with approach to a food cue in male and female heterogeneous stock rats. *Psychopharmacology*, *233*(13), 2593–2605. https://doi.org/10.1007/s00213-016-4306-x

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kruschke, J. K. (2021). Bayesian Analysis Reporting Guidelines. *Nature Human Behaviour*, *5*(10), 1282–1291. https://doi.org/10.1038/s41562-021-01177-7

Liu, C., Yücel, M., Suo, C., Le Pelley, M. E., Tiego, J., Rotaru, K., Fontenelle, L. F., & Albertella, L. (2021). Reward-Related Attentional Capture Moderates the Association between Fear-Driven Motives and Heavy Drinking. *European Addiction Research*, *27*(5), 351–361. https://doi.org/10.1159/000513470

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q. F., Smíra, M., Epskamp, S., Matzke, D., Wild, A., Knight, P., Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2019). **JASP**: Graphical Statistical Software for Common Statistical Designs. *Journal of Statistical Software*, *88*(2). https://doi.org/10.18637/jss.v088.i02

Lovic, V., Saunders, B. T., Yager, L. M., & Robinson, T. E. (2011). Rats prone to attribute incentive salience to reward cues are also prone to impulsive action. *Behavioural Brain Research*, *223*(2), 255–261. https://doi.org/10.1016/j.bbr.2011.04.006

Ludbrook, J. (2010). Confidence in Altman–Bland plots: A critical review of the method of differences. *Clinical and Experimental Pharmacology and Physiology*, *37*(2), 143–149. https://doi.org/10.1111/j.1440-1681.2009.05288.x

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. https://doi.org/10.1037/1082-989X.7.1.19

Martin Bland, J., & Altman, DouglasG. (1986). STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT. *The Lancet*, *327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Mehta, S., Bastero-Caballero, R. F., Sun, Y., Zhu, R., Murphy, D. K., Hardas, B., & Koch, G. (2018). Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies. *Statistics in Medicine*, *37*(18), 2734–2752. https://doi.org/10.1002/sim.7679

Meyer, P. J., Cogan, E. S., & Robinson, T. E. (2014). The Form of a Conditioned Stimulus Can Influence the Degree to Which It Acquires Incentive Motivational Properties. *PLoS ONE*, *9*(6), e98163. https://doi.org/10.1371/journal.pone.0098163

Meyer, P. J., Lovic, V., Saunders, B. T., Yager, L. M., Flagel, S. B., Morrow, J. D., & Robinson, T. E. (2012). Quantifying Individual Variation in the Propensity to Attribute Incentive Salience to Reward Cues. *PLoS ONE*, *7*(6), e38987. https://doi.org/10.1371/journal.pone.0038987

Mokkink, L. B., Eekhout, I., Boers, M., Van Der Vleuten, C. P., & De Vet, H. C. (2023). Studies on Reliability and Measurement Error of Measurements in Medicine – From Design to Statistics Explained for Medical Researchers. *Patient Related Outcome Measures*, *Volume 14*, 193–212. https://doi.org/10.2147/PROM.S398886

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695

Pedersen, T. L. (2019). *patchwork: The Composer of Plots* (p. 1.3.2) [Dataset]. https://doi.org/10.32614/CRAN.package.patchwork

Pennington, C. R., Birch-Hurst, K., Ploszajski, M., Clark, K., Hedge, C., & Shaw, D. J. (2025). Are we capturing individual differences? Evaluating the test–retest reliability of experimental tasks used to measure social cognitive abilities. *Behavior Research Methods*, *57*(2), 82. https://doi.org/10.3758/s13428-025-02606-5

Pietrock, C., Ebrahimi, C., Katthagen, T. M., Koch, S. P., Heinz, A., Rothkirch, M., & Schlagenhauf, F. (2019). Pupil dilation as an implicit measure of appetitive Pavlovian learning. *Psychophysiology*, *56*(12), e13463. https://doi.org/10.1111/psyp.13463

R Core Team. (2024). *_R: A Language and Environment for Statistical Computing_.* (Versione 4.4.2) [Software]. R Foundation for Statistical Computing.

Robinson, M. J. F., Anselme, P., Suchomel, K., & Berridge, K. C. (2015). Amphetamine-induced sensitization and reward uncertainty similarly enhance incentive salience for conditioned cues. *Behavioral Neuroscience*, *129*(4), 502–511. https://doi.org/10.1037/bne0000064

Robinson, T., & Berridge, K. (1993). The neural basis of drug craving: An incentive-sensitization theory of addiction. *Brain Research Reviews*, *18*(3), 247–291. https://doi.org/10.1016/0165-0173(93)90013-P

Robinson, T. E., & Berridge, K. C. (2025). The Incentive-Sensitization Theory of Addiction 30 Years On. *Annual Review of Psychology*, *76*(Volume 76, 2025), 29–58. https://doi.org/10.1146/annurev-psych-011624-024031

Robinson, T. E., & Flagel, S. B. (2009). Dissociating the Predictive and Incentive Motivational Properties of Reward-Related Cues Through the Study of Individual Differences. *Biological Psychiatry*, *65*(10), 869–873. https://doi.org/10.1016/j.biopsych.2008.09.006

Rousson, V., Gasser, T., & Seifert, B. (2002). Assessing intrarater, interrater and test–retest reliability of continuous measurements. *Statistics in Medicine*, *21*(22), 3431–3446. https://doi.org/10.1002/sim.1253

Saeedpour, S., Hossein, M. M., Deroy, O., & Bahrami, B. (2023). Interindividual differences in Pavlovian influence on learning are consistent. *Royal Society Open Science*, *10*(9), 230447. https://doi.org/10.1098/rsos.230447

Sarter, M., & Phillips, K. B. (2018). The Neuroscience of Cognitive-Motivational Styles: Sign- and Goal-Trackers as Animal Models. *Behavioral neuroscience*, *132*(1), 1–12. https://doi.org/10.1037/bne0000226

Saunders, B. T., O'Donnell, E. G., Aurbach, E. L., & Robinson, T. E. (2014). A Cocaine Context Renews Drug Seeking Preferentially in a Subset of Individuals. *Neuropsychopharmacology*, *39*(12), 2816–2823. https://doi.org/10.1038/npp.2014.131

Schad, D. J., Rapp, M. A., Garbusow, M., Nebe, S., Sebold, M., Obst, E., Sommer, C., Deserno, L., Rabovsky, M., Friedel, E., Romanczuk-Seiferth, N., Wittchen, H.-U., Zimmermann, U. S., Walter, H., Sterzer, P., Smolka, M. N., Schlagenhauf, F., Heinz, A., Dayan, P., & Huys, Q. J. M. (2019). Dissociating neural learning signals in human sign- and goal-trackers. *Nature Human Behaviour*, *4*(2), 201–214. https://doi.org/10.1038/s41562-019-0765-5

Schauberger, P., & Walker, A. (2024). *openxlsx: Read, Write and Edit xlsx Files}* (Versione R package version 4.2.7.1) [Software]. https://CRAN.R-project.org/package=openxlsx

Schettino, M., Mauti, M., Parrillo, C., Ceccarelli, I., Giove, F., Napolitano, A., Ottaviani, C., Martelli, M., & Orsini, C. (2024). Resting-state brain activation patterns and network topology distinguish human sign and goal trackers. *Translational Psychiatry*, *14*(1), 446. https://doi.org/10.1038/s41398-024-03162-w

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and “how to” guide of its application within vocational behavior research. *Journal of Vocational Behavior*, *120*, 103445. https://doi.org/10.1016/j.jvb.2020.103445

Srey, C. S., Maddux, J.-M. N., & Chaudhri, N. (2015). The attribution of incentive salience to Pavlovian alcohol cues: A shift from goal-tracking to sign-tracking. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00054

Swintosky, M., Brennan, J. T., Koziel, C., Paulus, J. P., & Morrison, S. E. (2021). Sign tracking predicts suboptimal behavior in a rodent gambling task. *Psychopharmacology*, *238*(9), 2645–2660. https://doi.org/10.1007/s00213-021-05887-8

The MathWorks, Inc. (2024). *MATLAB: A High-Level Language for Technical Computing* (Versione R2024a) [Software].

Ukoumunne, O. C., Davison, A. C., Gulliford, M. C., & Chinn, S. (2003). Non-parametric bootstrap confidence intervals for the intraclass correlation coefficient. *Statistics in Medicine*, *22*(24), 3805–3821. https://doi.org/10.1002/sim.1643

Ushey, K., Allaire, J., Wickham, H., & Ritchie, G. (2024). *_rstudioapi: Safely Access the RStudio API* (Versione R package version 0.17.1) [Software]. https://CRAN.R-project.org/package=rstudioapi

Versace, F., Frank, D. W., Stevens, E. M., Deweese, M. M., Guindani, M., & Schembre, S. M. (2019). The reality of “food porn”: Larger brain responses to food-related cues than to erotic images predict cue-induced eating. *Psychophysiology*, *56*(4), e13309. https://doi.org/10.1111/psyp.13309

Versace, F., Kypriotakis, G., Basen-Engquist, K., & Schembre, S. M. (2016). Heterogeneity in brain reactivity to pleasant and food cues: Evidence of sign-tracking in humans. *Social Cognitive and Affective Neuroscience*, *11*(4), 604–611. https://doi.org/10.1093/scan/nsv143

Villaruel, F. R., & Chaudhri, N. (2016). Individual Differences in the Attribution of Incentive Salience to a Pavlovian Alcohol Cue. *Frontiers in Behavioral Neuroscience*, *10*. https://doi.org/10.3389/fnbeh.2016.00238

Volkow, N. D., Koob, G. F., & McLellan, A. T. (2016). Neurobiologic Advances from the Brain Disease Model of Addiction. *New England Journal of Medicine*, *374*(4), 363–371. https://doi.org/10.1056/NEJMra1511480

Watson, P., Prior, K., Ridley, N., Monds, L., Manning, V., Wiers, R. W., & Le Pelley, M. E. (2024). Sign-tracking to non-drug reward is related to severity of alcohol-use problems in a sample of individuals seeking treatment. *Addictive Behaviors*, *154*, 108010. https://doi.org/10.1016/j.addbeh.2024.108010

Weir, J. P. (2005). Quantifying Test-Retest Reliability Using the Intraclass Correlation Coefficient and the SEM. *The Journal of Strength and Conditioning Research*, *19*(1), 231. https://doi.org/10.1519/15184.1

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (2nd ed. 2016). Springer International Publishing : Imprint: Springer. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Williamson, J. M., Crawford, S. B., & Lin, H. (2007). Resampling Dependent Concordance Correlation Coefficients. *Journal of Biopharmaceutical Statistics*, *17*(4), 685–696. https://doi.org/10.1080/10543400701329471

Yager, L. M., & Robinson, T. E. (2010). Cue-induced reinstatement of food seeking in rats that differ in their propensity to attribute incentive salience to food cues. *Behavioural Brain Research*, *214*(1), 30–34. https://doi.org/10.1016/j.bbr.2010.04.021

Ye, W., Zheng, G., Cao, X., Ma, Y., & Zhang, A. (2024). *Spurious Correlations in Machine Learning: A Survey* (Versione 2). arXiv. https://doi.org/10.48550/ARXIV.2402.12715

Zeileis, A., & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. *R News*, *2*(3), 7--10.

Zorowitz, S., & Niv, Y. (2023). Improving the Reliability of Cognitive Task Measures: A Narrative Review. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *8*(8), 789–797. https://doi.org/10.1016/j.bpsc.2023.02.004