Supplementary Materials

Selection of human subjects. Consensus diagnoses for MCI-AD and neurologically healthy subjects were made by a team of neurologists, nurses, and neuropsychologists, using the results of a neurological exam, neuropsychological assessment, and an informant interview. Because we were interested in MCI patients who had likely hippocampal dysfunction and were in the prodromal stage of AD, we included only MCI patients who had memory impairment on standard memory testing, and who were likely to have AD pathology, as judged by the consensus team from the pattern and progression of symptoms. Clinical Dementia Rating scores were <1.0 for all MCI subjects, suggesting minimal or no functional decline. The MCI-AD and control groups did not differ in age (73.6 ± 8.3 vs. 72.2 ± 7.6, respectively), percent female (43% and 52%, respectively), or education (16.3 ± 2.3 vs. 17.3 ± 2.2 years) (all $P > .05$).

Mouse Morris water maze. The water maze consisted of a pool (122-cm diameter) filled with water (21 ± 2 °C) made opaque with nontoxic white tempera paint powder. The pool was surrounded by distinct visual cues 1–3 meters from the edge of the pool. The escape platform was square (225 cm^2, ~2% of the surface area of the pool) and submerged 1.5 cm below the surface. Swimming behavior was monitored with a top-down video tracking system (Noldus, Ethovision XT8.5). The protocol consisted of three phases: visible-target training, hidden-target learning, and a probe trial.

Visible-target training. A black-and-white striped pole (15 cm tall) was placed on the submerged platform, and extra-maze cues were removed from the wall. Each day for 2 consecutive days, mice had two training sessions per day with a 3-h intersession interval. Each session consisted of two training trials with a 30-min intertrial interval. For each trial, mice were placed into the water facing the sidewalls at different locations. The maximum time per trial was 60 s. Mice that did not mount on the platform were guided to it gently and allowed to sit on it for 10 s and were then removed.

Hidden-target learning. The day after visible-target training, mice had two learning trials per day with a 3-h intertrial interval for 6 consecutive days. For each trial, the drop location varied semirandomly. Mice were allowed to stay on the submerged platform for 10 s and were then removed. The maximum time per trial was 90 s.

Probe. The probe trial was done 18–20 h after the last hidden-target learning trial. The platform was removed, and mice were allowed to swim for 90 s before they were removed. The drop location for probe trials was opposite to the target location during hidden-target learning.

Virtual Morris maze. All human subjects were tested with a virtual adaptation of the Morris water maze test, which runs as a desktop application with a 180 degree field of view on Windows. The software was developed with a 3D Toolkit from Microsoft Research. Subjects were placed in front of a 30” monitor with a simple driving simulator that included a gas pedal and a steering wheel (Logitech Driving Force GT Steering Wheel). We selected the driving simulator because the subjects quickly learned how to use it; in pilot tests, a joystick was confusing for many subjects. All three tasks—visible-target training, hidden-target learning, and probe—used the same virtual environment, a tan circular field that was 100 units in diameter (See Supplementary Video). The program logged all gas pedal and steering wheel movements and maze position 80 times per second. During visible-target training, the field was surrounded by a blue skybox with no clouds or other landmarks. During hidden-target learning and probe trials, the skybox showed a sun, water tower, mountains, houses, and trees, whose locations remained constant. The target comprised 4% of the field and was located halfway between the perimeter and center of the field. On all three tasks, the subject used the steering wheel and the gas pedal to drive within the circular field but could not drive off the field and could not drive in reverse. The subject’s position was represented by a small vehicle, and the viewing direction within the scene was updated continuously whenever the subject turned. Turning was possible

even when the gas pedal was not depressed, which allowed for sharper turns. The gas pedal allowed for variable speed with a maximum speed of 10 distance units per second.

Visible-target training. On each of four visible-target trials, the subject started from a unique location on the edge of the circular field, and a lavender box appeared at the target location. Subjects were instructed to drive to the box as quickly as possible. When the subject drove into the box, it turned into a treasure chest. After 5 s, the subject was virtually placed at the starting position for the next trial. Instructions on how to use the driving simulator were administered during a sample visible-target trial before the first trial. There was no time limit.

Hidden-target learning. On each of 10 hidden target trials, the subject started from a unique position on the edge of the circular field. The subject was told that a treasure was buried in the field and would appear when they drove over it, that they would have many trials to look for the treasure, and that the treasure would always be buried in the same place relative to the trees, mountains, and other cues that surrounded the field. If the subject did not find the treasure within 120 s, the treasure appeared, and the subject was instructed to drive directly to the treasure. After the treasure was reached, the subject remained at that location for 5 s, and then was virtually placed at the starting position for the next trial.

Probe. After a 40-min delay, the subject was placed in the same environment at the edge of the field. The subject was asked to look for the treasure, which was buried in the same place as before, but this time it would not appear. Even though the treasure did not appear, the subject was instructed to keep looking for the treasure, as if they wanted to drive over it as many times as possible. The probe trial lasted 90 s.

Measure Correction for Start Location. To correct measures for the start location on each trial, we did not count the initial path until the subject had traveled a distance equal to the direct distance between the start location and the target. Thus, the measures represented search error and were not biased by the distance of the starting position to the target, which varied by trial

(1). In mice, this adjustment was made for distance but not for time measures because of the substantial variability in swim speed within each trial (see Supplementary Figure 1).

Rank-summary measures. The canonical method for analyzing Morris maze data is repeated-measures ANOVA—despite reports that this method is inappropriate for several reasons. Often a nonlinear trend is observed, as controls learn very quickly, with this learning effect leveling off once the maze is fully understood (2). A "saw-blade" effect is often seen, as subjects are more likely to perform better on the last trial of a given day than on the first trial of the next day. Nonconstant variance is common, as there is typically a mean-variance relation in time to event data. Since the trial is aborted after a fixed amount of time, the data are often subject to right censoring (3). Finally, the correlation structure between observations on a given subject is more complicated than the compound symmetry assumed by repeated-measures ANOVA (4).

New methods have been proposed to address some of these issues. Linear mixed-effects models account for variable learning rates within a group (4) but do not account for nonlinear trends, nonconstant variance, or censoring. Such models can account for nonlinear trends (2, 5), but the form of the trends is difficult to predict *a priori* (2). The Cox proportional-hazards mixed-effects model addresses the nonconstant variance and the censoring (6) but does not model nonlinear learning rates. These methods are difficult to apply appropriately by nonstatisticians, and the data must be examined before the model is selected—a source of bias in statistical tests and inappropriate for clinical trials.

We devised a rank-summary score that avoids the problems of repeated-measures ANOVA by taking advantage of the balanced nature of the experiments to transform the data into a more manageable form. Because the subjects are always trained in an identical fashion for a given experiment, we first used a matched design that makes it unnecessary to model the complete mean structure: only outcomes of subjects at the same trial number are compared directly. This

is particularly useful when combining data from different cohorts or protocols. For example, we found some differences between cohorts of mice even with the same protocol (Supplementary Figure 2). Second, we replaced the rank scores with quantile scores. For example, the subject who finishes third among 20 subjects in a given trial would receive a score of 3/20. Quantiles were applied separately for each mouse cohort. Using quantile scores greatly reduces the influence of outliers and the high variance observed in the early trials and accounts for censoring. Subjects who did not complete the task in the allotted time were censored as "tied for last." Finally, rather than trying to model a correlation structure of the multiple scores on each subject, we average the quantile scores across trials to get a single summary score per subject. Analyzing summary measures rather than trial-level data greatly simplify the statistical methods. For example, the learning performance of two groups can be compared by *t* test. Thus, the rank-summary method requires very little model inspection, simplifies the analysis and interpretation of results, and enables valid combination of data across cohorts and comparison of results across species.

To assess the validity of the rank-summary method, we randomly assigned treatment categories to our dataset. This procedure was repeated 5,000 times, and the number of significant results (i.e., false positives) was recorded to get a bootstrap estimate of the true significance level under the null hypothesis. This was done for each of six measurements: latency, distance, and CSE in both mice and human subjects. The observed significance of the rank-summary was close to the nominal, 0.0466–0.0566. The observed significance of repeated-measures ANOVA was slightly higher, 0.0496 to 0.116. These results are consistent with reports that repeated-measures ANOVA can increase Type I error rates (2).

Rank-summary scores can be calculated in Excel by using the PERCENTRANK function to transform the raw scores into percentile ranks for each subject in each trial and then using the

AVERAGE function to get the average percentile rank. Rank-summary scores can be calculated with R-code (see https://github.com/pistacliffcho/rankSummaries.git).

References

1. Gallagher M, Burwell R, and Burchinal MR. Severity of spatial learning impairment in aging: development of a learning index for performance in the Morris water maze. *Behav Neurosci.* 1993;107(4):618.
2. Young M, Clark M, Goffus, A, and Hoane, M. Mixed effects modeling of Morris water maze data: Advantages and cautionary notes, *Learn Motiv*. 2009;40:160-177.
3. Stein J, Bergman W, Fang Y, et al. Behavioral and neurochemical alterations in mice lacking the RNA-binding protein Translin. *J Neurosci.* 2006;26(8):2184-96.
4. Wagner A, Brayer S, Hurwitz M, et al. Non-spatial pre-training the water maze as a clinical relevant model for evaluation learning and memory in experimental TBI. *Neurobio Learn Mem.* 2013;106:71-86.
5. Gillani R, Tsai S, Wallace D, et al. Cognitive recovery in the aged rate after stroke and anti-Nogo-A immunotherapy. *Behav Brain Res*. 2010;(208)415-24.
6. Jahn-Eimermacher A, Lasarik I, and Raber J. Statistical analysis of latency outcomes in behavioral experiments. *Behav Brain Res.* 2011;221(1):271-5.

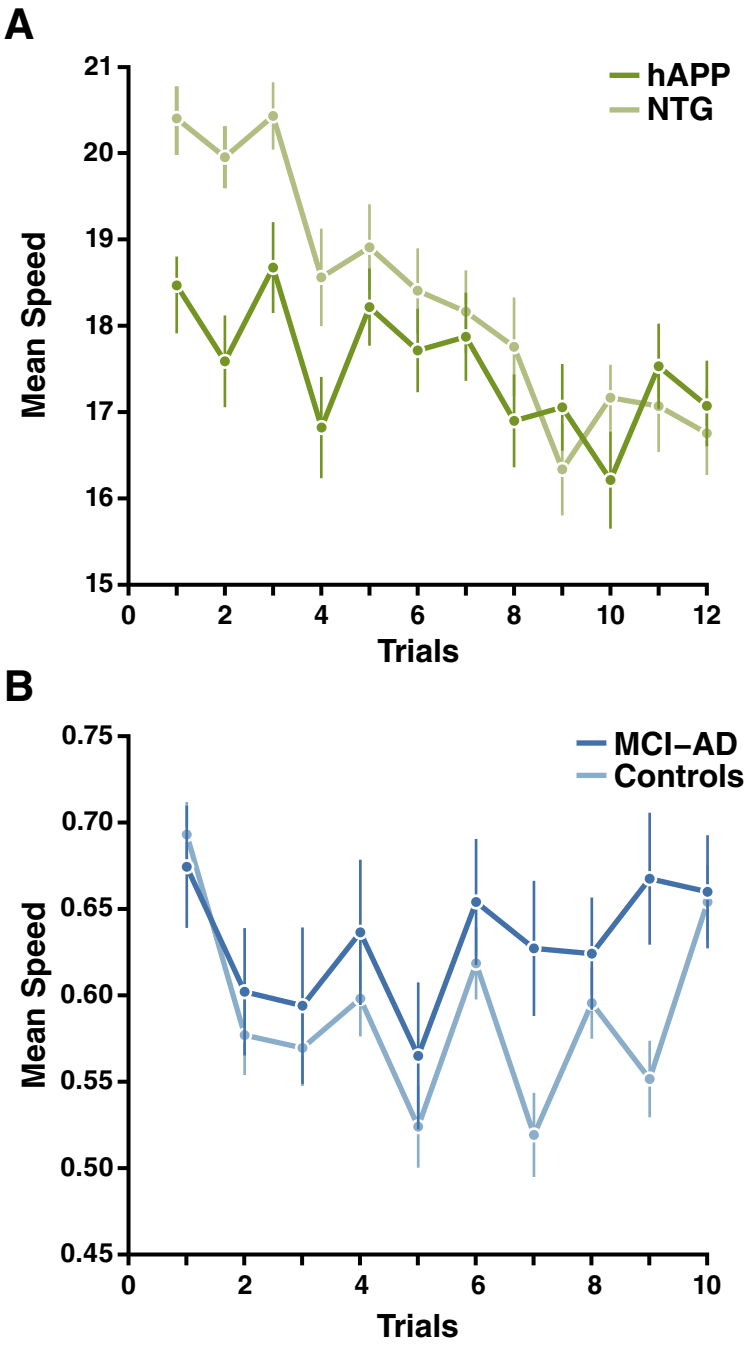
Supplementary Video. Human virtual Morris maze. One visible-target training trial and three hidden-target learning trials are presented. This task is administered with a 30” monitor and a simple driving simulator.

Supplementary Figure 1. Navigational speed in hidden-target learning trials across species. Mean speed by trial in hAPP and NTG mice **(A)** and MCI-AD patients and their controls **(B).** Rank summary speed scores were significantly different between hAPP and NTG mice ($P < 0.05$), but did not significantly differ in the human groups ($P = 0.22$). Units: A, cm per second; B, virtual distance units per second where the field is 100 units in diameter.

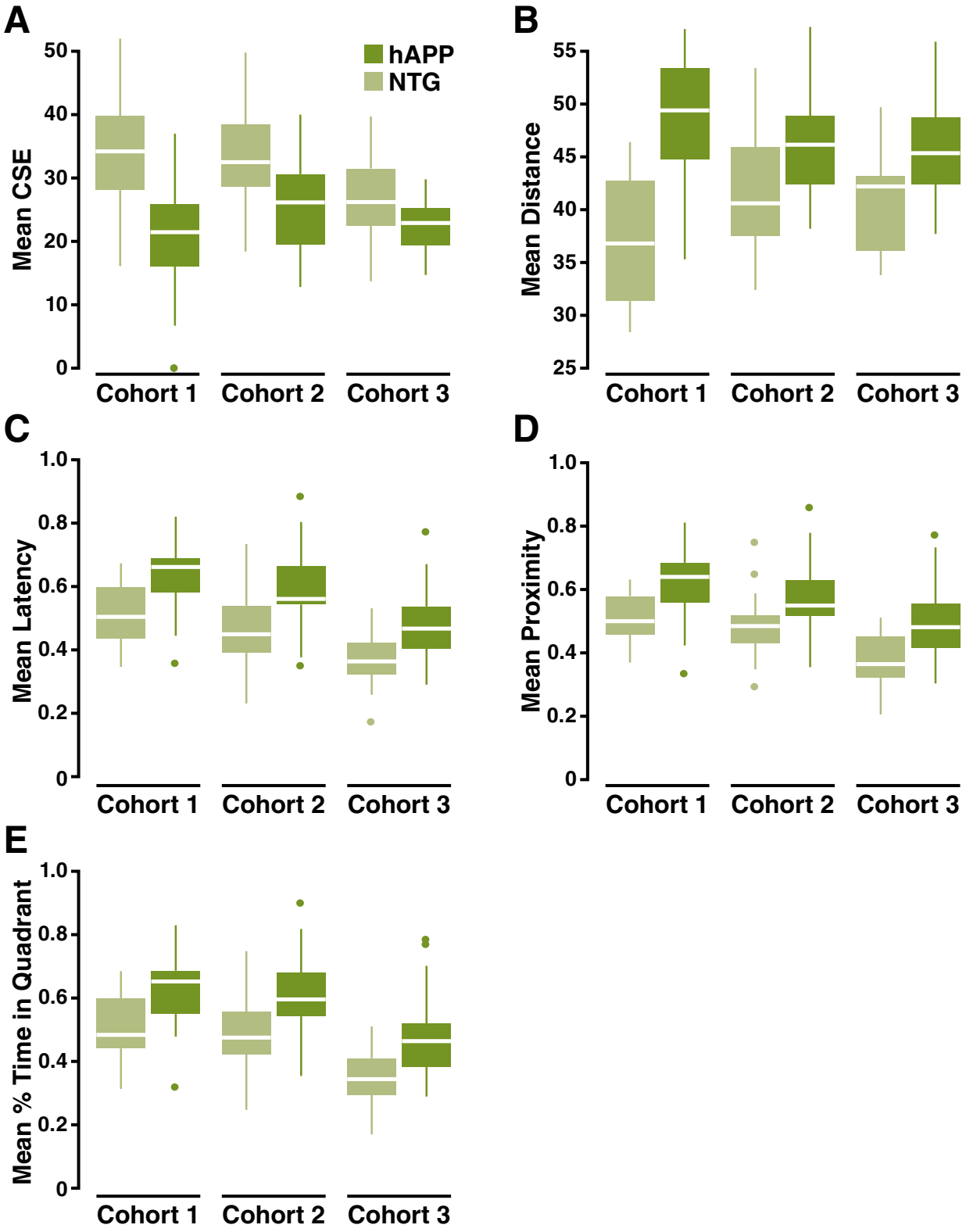
Supplementary Figure 2. Rank summary scores on hidden-target learning and probe for hAPP mice and their controls in 3 different cohorts. Hidden-target rank summary scores for distance, latency and CSE, and mean proximity and percent time in the target quadrant, are compared across three independent cohorts of hAPP and NTG mice.

Supplementary Figure 3. The Cox PH model does not increase the probability to detect impairment on hidden-target learning over the rank summary score method. Hidden-target learning performance analyzed by the rank summary score (solid line) or Cox PH model (dashed line) separately for Distance, Latency, and CSE in hAPP mice **(A)** and MCI-AD patients **(B)** is presented in power curves by sample size with Type I error rate set at .05. Abbreviation: CSE, cumulative search error.

Possin, et al, Supplementary Figure 1



Possin, et al, Supplementary Figure 2



Possin, et al, Supplementary Figure 3