Examining individual learning patterns using generalised linear mixed models

Sean Commins, Antoine Coutrot, Michael Hornberger, Hugo J Spiers, Rafael de Andrade Moral

Published in
Behavior Research Methods
https://doi.org/10.3758/s13428-023-02232-z

Examining individual learning patterns using generalised linear mixed models

Sean Commins[1] · Antoine Coutrot[2] · Michael Hornberger[3] · Hugo J Spiers[4] · Rafael De Andrade Moral[5]

Accepted: 31 August 2023

Abstract
Everyone learns differently, but individual performance is often ignored in favour of a group-level analysis. Using data from four different experiments, we show that generalised linear mixed models (GLMMs) and extensions can be used to examine individual learning patterns. Producing ellipsoids and cluster analyses based on predicted random effects, individual learning patterns can be identified, clustered and used for comparisons across various experimental conditions or groups. This analysis can handle a range of datasets including discrete, continuous, censored and non-censored, as well as different experimental conditions, sample sizes and trial numbers. Using this approach, we show that learning a face-named paired associative task produced individuals that can learn quickly, with the performance of some remaining high, but with a drop-off in others, whereas other individuals show poor performance throughout the learning period. We see this more clearly in a virtual navigation spatial learning task (NavWell). Two prominent clusters of learning emerged, one showing individuals who produced a rapid learning and another showing a slow and gradual learning pattern. Using data from another spatial learning task (Sea Hero Quest), we show that individuals' performance generally reflects their age category, but not always. Overall, using this analytical approach may help practitioners in education and medicine to identify those individuals who might need extra help and attention. In addition, identifying learning patterns may enable further investigation of the underlying neural, biological, environmental and other factors associated with these individuals.

Keywords Learning · GLMMs · Spatial · Individual · Cluster analysis

Introduction

Learning is a dynamic process and fluctuates across time. Repeated examination of a task generally leads to improved performance; however, learning rates are individualistic, with some learning a task more quickly than others, while still others may never learn. Such variation across time and individuals is often well captured by variation around the mean, relying on repeated-measures analyses of variance (ANOVAs) or mixed-factorial ANOVAs when comparing group performance across time (see Barnhart et al., 2015; Bootsma et al., 2018; Raboyeau et al., 2010, across spatial, visuo-motor and lexical learning domains, and our own Farina et al., 2015). However, examining variation and how individuals perform a task across time is important. Given the replication crisis across many fields including psychology and the neurosciences, there is an increased emphasis on displaying all data points in a clear and explicit manner (see Allen et al., 2019, for data visualisation using raincloud plots). In addition to being more transparent, examination of individual performance and data may enable a better evaluation of outliers, a comparison of performance in various settings (e.g. in education, see Braithwaite et al., 2019), an examination of individual diagnosis or treatment plans (Chiang et al., 2020; Simon, 2001) and the ability to account for individual behavioural patterns (Seidler et al., 2015).

✉ Sean Commins
Sean.Commins@mu.ie

1 Department of Psychology, Maynooth University, Maynooth, Co Kildare, Ireland

2 Laboratoire d'InfoRmatique en Image et Systèmes d'information, CNRS, Université Claude Bernard, Lyon,, France

3 Norwich Medical School, University of East Anglia, Norwich NR4 7TJ, UK

4 Department of Experimental Psychology, Institute of Behavioural Neuroscience, Division of Psychology and Language Sciences, University College London, London WC1H 0AP, UK

5 Department of Mathematics and Statistics, Maynooth University, Maynooth, Co Kildare, Ireland

Generalised linear mixed models (GLMMs) are widely used across many fields (Bolker et al., 2008; Demétrio et al., 2014). The recognition that data may not be normally distributed and the addition of random effects to the linear predictor are important features of GLMMs. GLMMs have also been increasingly used in psychology and cognitive science (Baayen et al., 2002), and the method has even been suggested as the main analytical tool for quantitative data (Meteyard & Davies, 2020). GLMMs are very versatile and have been used to examine many psychological and cognitive constructs, such as associating grip strength with cognitive decline in older adults (Quesque et al., 2020; Chou et al., 2019), as well as in patients with varying forms of depression (Firth et al., 2018). In addition, they have been used to look at the relationship between cognitive processes and neural biomarkers, again in a variety of cohorts, from young adolescents (Paulus et al., 2019) to patients suffering from dementia (McDade et al., 2018). One of the great strengths of GLMMs is their use in the modelling and prediction of outcomes using longitudinal or repeated data. For example, Song et al. (2020) used baseline cardiovascular scores to predict cognitive decline and neural changes in the subsequent 21 years. Similarly, in a 2-year follow-up examination of adults with type 2 diabetes, Mattei et al. (2019) showed that those who adopted a Mediterranean diet demonstrated higher cognitive scores than those who did not.

GLMMs account for main and interaction differences (for fixed experimental effects, e.g. we may use mixed-model ANOVAs for Gaussian GLMMs). Importantly, GLMMs are used to estimate variance components associated to random effects. Such random effects may arise from individual differences, with participants deviating from the grand mean with respect to time, location or other unknown factors. Although the inclusion of such random effects may be typically used as a control feature, it is also an ideal way to examine individual variation in performance with respect to task items or across time, such as learning, or examining the effects of sleep loss on attention. For example, Cochrane et al. (2021) used a GLMM (in comparison to their own model) to specifically examine and predict individual vigilance scores from a number of sleep-related measures. Similarly, Kliegl et al. (2011) provided a nice illustration of how individual performance can be examined using GLMMs. In this study, participants had to complete a visual attention task and respond as quickly as possible to stimuli presented on a screen (there were three experimental conditions—response to changes in object, spatial and central fixation conditions). The authors assessed the significance of main (fixed) effects for all three conditions. Importantly, they also showed how individual responses vary within a condition and that individual responses may correlate across some conditions and not others (random effects). For example, the authors showed that responses of individuals in the spatial condition were very variable (with some performing very well compared to others), whereas in the object condition there was limited variability across individuals. Further, participants' responses in the spatial condition correlated strongly (negatively) with responses for the central fixation condition, but did not correlate at all with the object condition. In addition, the authors show how individual responses differ and how correlations might change, depending on the model used.

Although used in the above examples, this type of extra individual-level analysis is seldom performed, with many studies including only random effects in their models to control for dependence between experimental or observational units, i.e. to simply reflect design. As a result, interesting insights into individual learning (or indeed patterns of learning across conditions) may be lost. Further to this, and mirroring the criticisms of Meteyard and Davies (2020) regarding the multiplicity of approaches of GLMM reporting in general, there is little guidance as to how such individual-level analysis should be examined. Here, we offer a set of approaches and show how GLMMs can be used to examine individual learning patterns by exploring individual-level random effects combined with outlier detection and clustering methodologies. We illustrate how a general framework may be applied using three different datasets, chosen specifically due to the different nature of the response variables to be analysed. The first (face-name pairs task) has discrete proportions as a response, thereby constituting an example of data that can be analysed using a binomial GLMM. The second (virtual navigation task) measured time as a response, which is a strictly positive and continuous response, and in this case right-censored, which can help illustrate continuous and censored GLMMs. The third (Sea Hero Quest) presents a strictly positive and continuous score as the response, and illustrates the use of continuous GLMMs with flexible functions included in the linear predictor to model non-linear behaviour.

Methods

A number of datasets and procedures were used to examine how the GLMMs and analysis would deal with different types of data (continuous, non-continuous, censored, non-censored), different numbers of participants and trials, and different learning tasks.

Behavioural procedures

The face-name pairs task was used as an example of discrete non-censored data and was used to assess associative learning and memory in a previous experiment (see Caffrey

& Commins, 2022). The task consisted of eight face-name paired stimuli, presented twice in a block; there were four blocks in total (see Zeineh et al., 2003). Each face-name pair was presented in random order on screen for five seconds. Either the four blocks were presented sequentially on the same day (massed condition) or one block of face-name pairs was presented each day for four days (spaced condition). After each of the four blocks, a test trial was given. Each trial consisted of the eight faces presented once in random order, without their corresponding names. The number of correctly recalled names associated with each of the eight faces (out of eight) was used to measure learning performance across the four blocks.

The first dataset was conducted in a controlled laboratory setting with 57 participants in the massed condition and 61 in the spaced condition. Both groups were well matched in terms of gender (massed: M/F = 31/30, spaced: M/F = 29/28), age (massed = 22.9 [SD = 1], spaced = 23.4 [SD = 1.2]) and general IQ as measured by the National Adult Reading Test (NART; massed = 23.4 (SD = 1.6), spaced = 24.3 (SD = 1.7).

The second experiment was conducted online through the Qualtrics online survey platform and included 358 participants in total (179 per condition, massed and spaced). Because of the uncontrolled online conditions, the conditions were not well matched for gender (85 male/94 female for the massed condition and 70 male/109 female for the spaced condition) or age, with the massed group being older (mean = 29.4 [SD=13.1]) than the spaced group (mean = 24.1 [SD=8.8]).

We used both these datasets to explore how our analysis would deal with non-continuous data and differing numbers of participants.

The *virtual navigation task* (NavWell, Commins et al., 2020) was used as an example of continuous data that was right-censored (i.e. with a maximum limit). This task is the human equivalent of the Morris water maze task that is generally used to assess spatial learning and memory (Morris, 1981). The task consists of participants virtually navigating around an enclosed circular arena (22 virtual metres or 15.75 seconds to traverse its diameter) in an attempt to locate an invisible target, located somewhere on the ground. The hidden target only becomes visible once the participant traverses it. On subsequent trials, participants must try to recall this specific location and make their way to the hidden target as quickly as possible. To aid their recall, two cues (large shapes) are located on the wall of the circular arena. Each participant is given 12 trials; participants must try to find the target within 60 seconds. If the participant cannot locate the target within this time, they are transported to the hidden target and instructed to look around the arena and try to recall the location on the subsequent trials. For those who successfully locate the target, they are also told to look around the arena and try to recall the specific location. Time taken to reach the target for each of the 12 trials is used as the dependent measure. Lower scores reflect better spatial learning and memory. For this task, 42 participants were included (M/F = 19/23; mean age 28.3 [SD = 14, range 19–62]).

The Sea Hero Quest (SHQ) mobile video game (Coutrot et al., 2018; Spiers et al., 2021) was used to examine continuous, non-censored data that required semi-independent learning (the task contains increasing levels of difficulty). The SHQ task was designed to measure human spatial navigation ability through gameplay. Currently, over four million people across 195 countries have downloaded and played the game. The game has a number of different features, but it primarily involves participants virtually navigating a boat through a series of waterways and rivers to find a target—the goal is to find a sea creature in a particular location and photograph it. Before setting off, participants are provided with a map that shows their current location and the target location to which they need to navigate. The task has a number of different levels; each level is increasingly difficult and contains more twists and turns (i.e. takes longer to complete), as well as having rivers that do not lead anywhere. There is no time limit to the task, but unlocking a more difficult level is dependent on completing the previous one. The length (in virtual metres) to reach the target is used to measure spatial learning. Gaming ability and difference in touchscreen handling/proficiency of participants has been taken into account and controlled by normalising SHQ performance for the first two levels (see Coutrot et al., 2018, for details). Although the task relies primarily on spatial learning and memory, it also depends on other cognitive processes including the translation of a 2D map into a 3D game, planning of routes, and the continuous monitoring of progress during the game.

The overall dataset contains 3317 participants; however, to illustrate our proposed methodology, we have just looked at a random subset of this ($n = 240$), with four age groups (18–20, 21–40, 41–60 and 61–80 years, $n = 60$ per cohort). Within each cohort we tried to match for gender and to have an even spread of age. The resulting data give us the following for the four groups, respectively: M/F: 30/30, mean age = 19.47, SD = 0.5; M/F: 31/29; mean age = 30.5, SD = 5.7; M/F: 31/29, mean age = 50.48, SD =5.4 and M/F: 33/27 mean age = 70.9, SD = 5.7.

Statistical procedures

Here we present the statistical methods used for each sample dataset. For a summary, see Table 1, which also includes the syntax used to fit the models in R (R Core Team, 2021).

Face-name pairs task

For this task, the response variable is the number of times the individual made a correct association between a face and a name. This response is discrete and bounded between 0 (all incorrect) and 8 (all correct). Typically, binomial models are used to analyse discrete proportion responses, also known as logistic regression (when the logit link is used). A normal distribution would be an inadequate assumption, since normal models assume unbounded and continuous responses.

Here we fitted a binomial GLMM using the logit link. Other options would be the probit or complementary log-log links, for example, but since there are only four trials, all link functions would perform similarly (one would choose the complementary log-log link, for instance, if the sigmoidal shape in the response were asymmetric). Since we are interested in studying the learning behaviour over time, and how it changes according to learning condition, we included a different linear effect of trial (as a continuous predictor, ranging from 1 to 4) per learning condition (massed vs. spaced) in the linear predictor, which yielded one intercept and one slope per learning condition (i.e. four fixed effects). The assumption of a linear effect in the linear predictor scale is common, but exploratory analyses are always useful to guide the analysis, and could suggest the inclusion of other terms in the linear predictor. We also included individual-level random intercepts and slopes, which are able to describe individual learning curves. These were assumed to be independent and to follow a normal distribution with mean zero and a variance to be estimated by the model (i.e. two variance components). Since we used a logit link function, the response is modelled in the log-odds scale. Consequently, the slope parameters represent here how one trial affects the log-odds of a correct response. Typically, we look at the exponentials of the slopes as the change in the odds of a correct response associated with one extra trial. For example, a slope of 0.8 yields $e^{0.8} = 2.2$, which means that the odds of making correct associations in the next trial are 2.2 times those in the current trial.

For the binomial distribution, the dispersion parameter is known (or fixed) and equal to 1. However, if the variability in the data is larger than expected by the binomial model, it is possible to estimate this dispersion parameter in a quasi-likelihood approach to accommodate the extra-variability. Other approaches include the use of mixtures of distributions, such as the beta-binomial, or the inclusion of an observational-level random effect within a GLMM framework.

Virtual navigation task

The response here is strictly positive (since it is the time taken for the individual to reach the target), and the right-censoring of the response is an important feature. This is because, if the individuals had not found the target within 60 seconds, the task would be interrupted. Therefore, for these cases, we have the information that it would have taken more than 60 seconds for that individual to find the target, but we do not know exactly how long. This information can be incorporated in the modelling framework by using a different term in the likelihood for censored observations, based on the survival function (for this case, defined as the probability of the observation being more than 60 seconds). Again, the normal distribution would not be suitable, since although the data are continuous, they are strictly positive (i.e. time cannot be negative).

There are many statistical distributions that can be used to model strictly positive data. Here, we chose to fit a right-censored gamma GLMM. The gamma distribution is very flexible, and accommodates different shapes of continuous, positive and right-skewed data. The gamma distribution can be parameterised to have a mean and a dispersion parameter proportional to the variance. For the mean parameter, we included a linear effect of trial (as a continuous predictor, ranging from 1 to 12), that is, an intercept and a slope, in the linear predictor as fixed effects, and random and independent intercepts and slopes per individual, to describe individual learning curves. Additionally, we included the linear effect of trial in the linear predictor for the dispersion parameter, which allowed us to model the changes in variability as the trials progressed. Although the canonical link function for the gamma GLM is the inverse, the mean was modelled with a log link for ease of interpretability. The dispersion was also modelled with a log link, because it is a strictly positive parameter, and the log link maps the real values to strictly positive values. In this case, the slope coefficients represent the change in time until reaching the target associated with one extra trial in the logarithmic scale. Typically, we look at the exponentials of the slopes as a measure of multiplicative change. For example, a slope of -0.5 yields $e^{-0.5} = 0.6$, which means that the time needed to reach the target in the next trial will be 60% of the current time, i.e. 40% faster.

Other potential approaches would include the use of different link functions for the mean parameter, such as the identity link, or the canonical inverse link previously mentioned, the inclusion of semi-parametric terms in the linear predictor such as splines (to capture non-linear behaviour), or the use of alternative distributions to the gamma, such as the inverse Gaussian, Weibull or log-normal, among others. There is a plethora of probability distributions that accommodate positive continuous data, with a continuously growing literature on the development of new models aimed at time-until-event and censored data.

Sea Hero Quest

For this study, the response is also strictly positive and continuous, as it is the length of time to reach a target. In contrast to the virtual navigation task, the data are not censored,

Table 1 Nature of response variable, modelling framework, structure of the linear predictors and methods used for individual-level exploration of the learning behaviour across trials, for each dataset described in the methodology section

Dataset	Nature of response variable	Model	Structure of linear predictors	Individual-level exploration
Face-Name Pairs task	Discrete proportions	Binomial GLMM	Linear predictor for the mean Fixed effects of trial (linear) and learning condition Random intercepts and slopes per individual (i.e. two random effects) Linear predictor for the dispersion Fixed dispersion of 1 R syntax (using lme4 package) `glmer(cbind(Score, total - Score) ~ Trial * Condition + (Trial || Condition:ID), family = binomial, data = dataset_name)`	Fitted curves Bivariate normal ellipsoids Clustering based on random effects
Virtual navigation task	Positive continuous, right-censored	Gamma GLMM	Linear predictor for the mean Fixed effect of trial (linear) Random intercepts and slopes per individual (i.e. two random effects) Linear predictor for the dispersion Fixed effect of trial (linear) R syntax (using gamlss package) `gamlss(Surv(time = Score, event = censoring_index, type = "right") ~ Trial + re(random = list( ID = pdDiag(~ Trial))), sigma.formula = ~ Trial, family = cens(GA), data = dataset_name)`	Fitted curves Bivariate normal ellipsoids Clustering based on random effects
Sea Hero Quest	Positive continuous	Gamma GLMM	Linear predictor for the mean Different fixed effects of trial (b-spline with three knots) per age category Random b-splines (three knots) per individual (i.e. four random effects) Linear predictor for the dispersion Different fixed effects of trial (b-spline with three knots) per age category R syntax (using glmmTMB package) `glmmTMB(Score ~ bs(scale(Trial, 3)) * Age + (bs(scale(Trial), 3) || ID), dispformula = ~ bs(scale(Trial, 3)) * Age, family = Gamma(link = log), data = dataset_name)`	Fitted curves Clustering based on random effects

which means that the trials were not limited in terms of a maximum length. The behaviour of the response is non-linear, and therefore a simple linear model with an intercept and slope would not be sufficient. This non-linearity would also not be properly described by a quadratic equation.

Here, we fitted a gamma GLMM, but instead of a linear effect of trial, we included semi-parametric smooth functions based on b-splines. B-splines are linear combinations of values in the *x*-axis that offer a high degree of flexibility when modelling non-linear behaviours and at the same time are smooth. The b-splines used here had three knots across trials (again, taken as a continuous predictor, ranging from 1 to 5). This totalled four estimated parameters to represent a curve. We estimated a different curve per age group, totalling 16 parameters, or fixed effects, in the linear predictor for the mean of the distribution. We also included individual-level random effects, representing a different smooth curve per individual. Therefore, each individual is associated to four random effects, instead of two as described in the previous examples. These effects do not have a clear meaning such as an intercept and slope, but still hold the information used to generate the non-linear smooth curves that describe each individual's learning behaviour. We also included the same fixed effects used for the mean parameter in the linear predictor for the dispersion, i.e. different smooth functions across trials per age group (also totalling 16 parameters). Both mean and dispersion were modelled with a log link for the same reasons outlined in the virtual navigation task example.

Producing ellipsoids based on the bivariate normal distribution

For the models that estimate individual-level curves based on random intercepts and slopes, it is possible to plot the random intercepts versus the random slopes, and produce ellipsoids based on the bivariate normal distribution. By looking at a 95% ellipsoid, for example, one might observe individuals who fall beyond the area delimited by the ellipsoid and look further into their learning behaviour, since it may be considered "extreme", or "outlying", when compared to others. Moreover, when comparing e.g. treatment levels, different ellipsoids may be produced for different groups, and simple statistics may be computed from these ellipsoids, such as area and eccentricity indices, which may aid comparison of treatment levels.

The power of this type of analysis lies on how interpretable the random effects are. For the intercepts versus slopes case, it is a great tool to discriminate individuals according to their learning behaviour. The plot can be divided into four quadrants (see Fig. 1). Points close to the origin represent individuals who have intercepts and slopes very similar to the overall mean. If the response variable is directly proportional to learning (e.g. for the face-name pairs task experiment), in a plot with random intercepts as the *x*-axis and random slopes as the *y*-axis, points further in the first and fourth quadrants represent the faster learners in the pool (larger slopes), while points in the second and third quadrants represent the slower learners (smaller slopes). Points further in the third and fourth quadrants represent learners who started at a lower level (smaller intercepts), while points further in the first and second quadrants represent learners who started at a higher level (larger intercepts). This relationship is reversed when the response is inversely proportional to learning, e.g. for the virtual navigation task experiment, where the response is the time taken until reaching the target (shorter times represent a higher level of learning).

For higher dimensions (e.g. four random effects per individual) it becomes harder to visualise, but multi-dimensional ellipsoids based on the multivariate normal distribution can still be computed. However, in the example shown here (SHQ), the coefficients do not have easily interpretable meaning, and therefore ellipsoids are not explored.

Clustering individuals based on their random effects

The individual-level random effects may then be used to carry out clustering. This may help to uncover groupings in the data based on a summarised individual-level profile, if this profile is well represented by the random effects. In this paper we use hierarchical clustering based on Euclidean distances between individuals and Ward's method, but any other clustering method may be used. Ward's method focuses on minimising the variability within clusters whilst maximising the variability between clusters.

Assessing model goodness of fit

It is important to carry out analyses of the model residuals to check goodness of fit. If the model does not fit the data well, then inferences made based on the estimated parameters may be misleading. For instance, type I errors may be inflated when the model does not accommodate extra-variability in the data (i.e. failure to account for overdispersion); in this case, the model estimates less uncertainty than there actually is (see Demétrio et al., 2014, for more details). If, however, the variance estimated by the model is greater than the variability in the data (i.e. underdispersion), type II errors will be inflated. For the face-name pairs task and SHQ study, to assess whether the model fitted the data well, we used half-normal plots with a simulated envelope for the Pearson residuals (Moral et al., 2017). These plots are such that if the data are a plausible realisation of the fitted model, most residuals will lie within the simulated envelope. For the virtual navigation task, since the model was fitted using the gamlss package (Rigby & Stasinopoulos, 2005) in R (R Core Team, 2021),

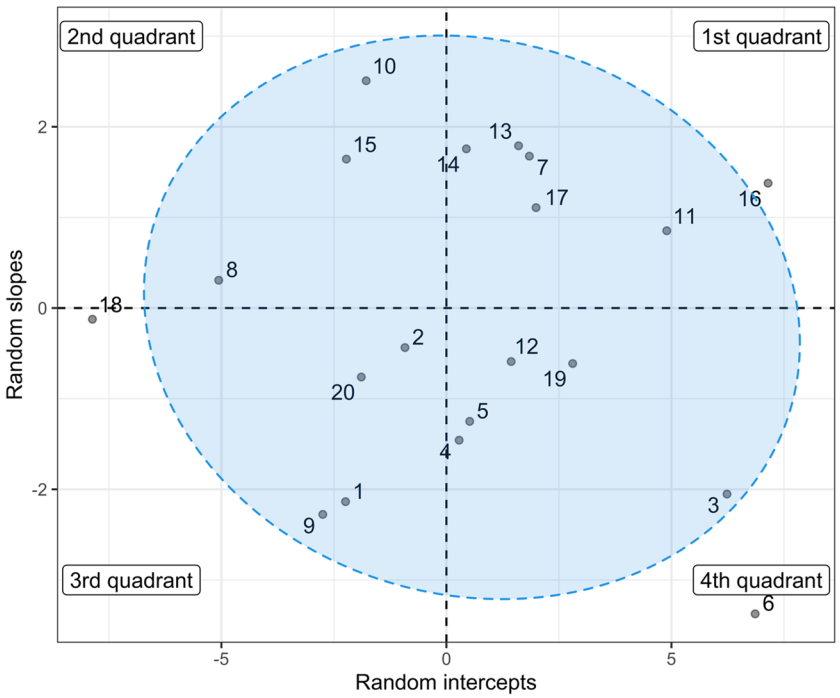

Fig. 1 Bivariate plot for random samples simulated from a normal distribution with mean zero and variance 4 to simulate random intercepts and variance 2 to simulate random slopes, for a fictitious sample of 20 individuals. The plot includes a 95% ellipsoid based on the bivariate normal distribution. For this particular plot, individuals 6, 16 and 18 present outlying behaviour. Individuals 6 and 16 present larger intercepts, which could represent advantageous initial learning states if the response is directly proportional to learning (e.g. face-name pairs task experiment) or disadvantageous states if the response is inversely proportional to learning (e.g. virtual navigation task experiment). Individual 16, however, presents a larger slope than the pool of individuals, while individual 6 presents a smaller slope. If the response is directly proportional to learning, then individual 16 would be a faster learner, while individual 6 would be a slower learner. Individual 18, on the other hand, presents a smaller intercept compared to the sample, but a slope close to zero, which represents an average learning rate

we used worm-plots for the randomised quantile residuals. These are similar to the half-normal plots described above, but they do not require the construction of a simulated envelope, since the randomised quantile residuals should follow a uniform (0,1) distribution if the model fits the data well.

Simulation studies

To better understand the reliability and robustness of the GLMM estimates and predicted random effects, as well as to compare with the standard GLM (without random effects), we carried out simulation studies based on 18 different scenarios. We simulated from three main models (binomial GLMM with random intercepts and slopes, gamma GLMM with random intercepts and slopes, and gamma GLMM with only random intercepts), three sample sizes (20, 40 and 80 participants) and two numbers of repeated measures/trials (4 and 8). For each scenario, we simulated 1000 datasets, and fitted the corresponding GLMMs and standard GLMs to each simulated dataset. The true models included an intercept and a linear effect of trial in the linear predictor. The true parameter values were set as $\beta_0 = -1.22$ and $\beta_1 = 1.11$ for the binomial GLMM, and $\beta_0 = 2.60$ and $\beta_1 = -0.08$ for the gamma GLMMs, and the variances for the random intercepts and slopes were set as 0.46 and 0.22, respectively (inspired by the estimates obtained from the models fitted to the online face-name pairs task and virtual navigation task).

We compared the individual-level estimates (random effects for the GLMMs and regression coefficients for the standard GLMs) with the true individual random effects by calculating the sum of squared differences across all individuals, then averaging over the 1000 simulated datasets for each scenario, thereby obtaining mean squared errors. For the GLMM fixed effects and variance components estimates, we calculated the mean relative bias, which is obtained by averaging the relative bias (estimate minus true parameter value divided by the true parameter value) across all 1000 simulated datasets for each scenario.

Software

All analyses were carried out in R (R Core Team, 2021). Binomial GLMMs were fitted using package lme4 (Bates et al., 2015), gamma GLMMs were fitted using package

glmmTMB (Brooks et al., 2017) and the gamma GLMMs for censored data were fitted using packages gamlss and gamlss.cens (Stasinopoulos et al., 2018). Model goodness of fit was assessed using package hnp (Moral et al., 2017). All data and code are made available at https://github.com/rafamoral/individual_learning_GLMM.

Results

Face-name pairs task

For the **online** face-name experiment, we began by assessing the significance of the variance component associated with the slopes. There was evidence that its inclusion significantly improved our model goodness of fit (LR = 154.67, df = 1, $p < 0.0001$), which means that different individuals have significantly different learning speeds (represented in our model by the slopes of the linear predictor, or the linear effect of trial). Looking at the fixed effects, there was a significant interaction between trial and group condition (LR = 12.12, df = 1, $p = 0.0005$) (Fig. 2a). The slope for the spaced group (0.86) is smaller than that for the massed group (1.11), i.e. on average the spaced condition is associated to slower learning (see Table 2). These slopes are interpreted in the log-odds scale, which means that in the massed group, adding a new trial increases the log-odds of correctly matching a face and a name by 1.11, which translates to being $e^{1.11} \approx 3$ times more likely to correctly match when compared to the previous trial. For the spaced group, individuals are, on average, $e^{0.86} = 2.36$ times more likely to obtain correct matching from one trial to the next one.

One of the perks of a mixed model is that we are able to obtain fitted curves at an individual level without having

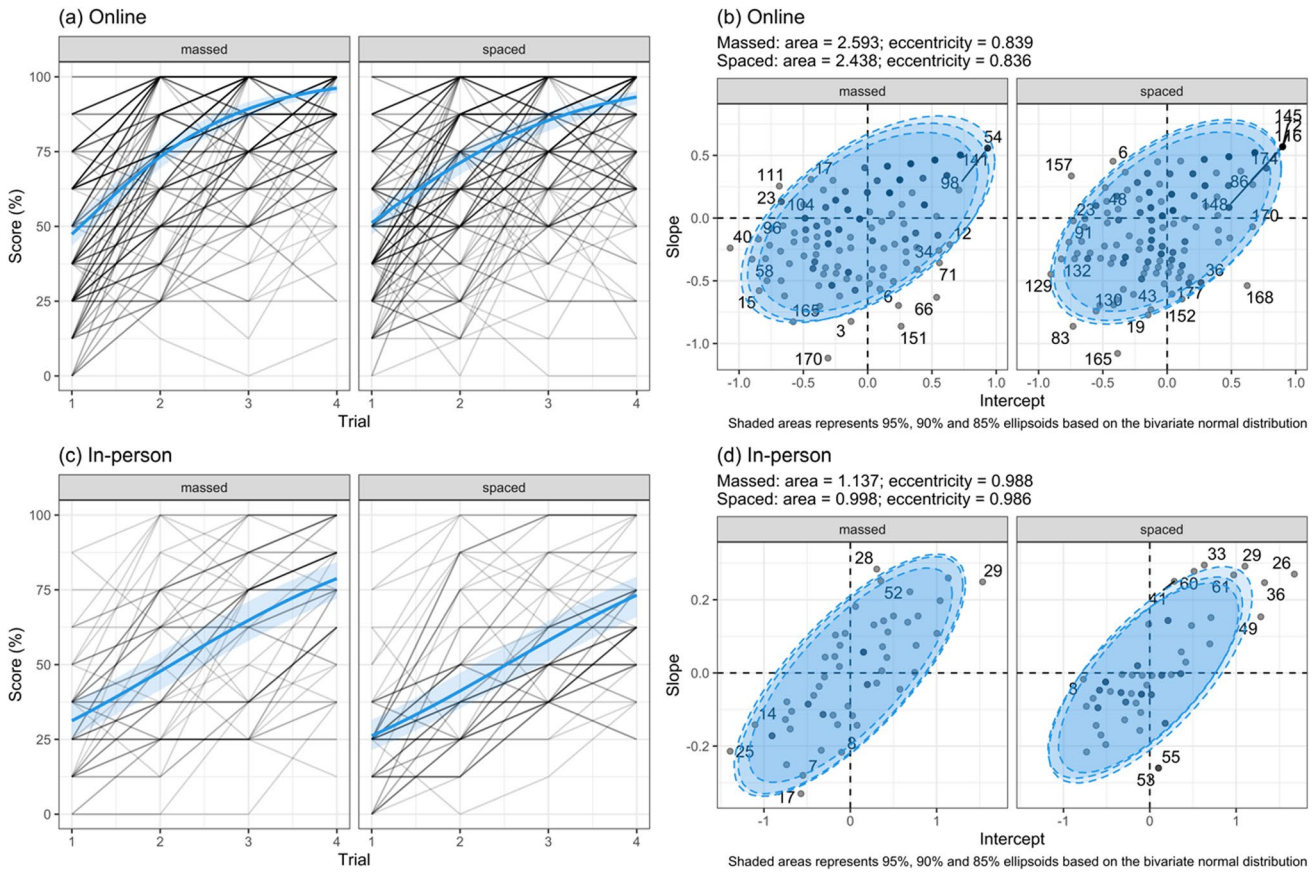

Fig. 2 **a** Observed (black lines) and predicted (blue curves) scores for individuals under massed vs. spaced learning conditions for the online experiment. The shaded areas represent the 95% confidence intervals for the true mean scores, based on the binomial GLMM (see Table 1). The curves from the massed and spaced conditions are statistically different (test for different slopes: LR = 12.12, df = 1, $p = 0.0005$). **b** Predicted random intercepts and slopes for each individual in the study who was at either a massed or spaced learning framework, as well as 85%, 90% and 95% ellipsoids based on the bivariate normal distribution, for the online experiment. Points falling outside of the 85% ellipsoids have their individual numbers indicated as potential outliers. The area of the ellipsis is calculated as $A = \pi ab$, where a and b are the major and minor axes, respectively. Eccentricity is calculated as $E = (1 - b^2/a^2)^{1/2}$. **c** Same as (a), but now for the in-person experiment; **d** same as (b), but now for the in-person experiment

to estimate a large number of parameters. In this particular case, estimating a linear predictor for all 358 individuals would require $358 \times 2 = 716$ parameters. Instead, we estimate an intercept and slope for each condition plus two variance components, related to random intercepts and slopes per individual, yielding four estimated parameters instead of 716, while maintaining the predictive power. This is because individual random effects are obtained through their conditional distribution given the data (McCulloch & Neuhaus, 2011), instead of being individually maximised in the model likelihood. In fact, they are not even present in the likelihood function, since they are integrated out as latent, unobserved variables.

Since we now have individual-level intercepts and slopes, which are assumed to arise from normal distributions with mean zero and variances estimated by the mixed model, it becomes straightforward to rank individuals based on their random effects. This may be done looking only at the univariate distribution of random intercepts or slopes, or at their bivariate distribution. Taking the random effects as a bivariate sample from a bivariate normal distribution with a zero vector of means and a variance covariance matrix estimated by the mixed model, we produce ellipsoids at different probability levels (see Fig. 2b). Although the overall eccentricity and area values for the two conditions are comparable, they do provide a good indication of spread. In this case, the massed condition has slightly more spread (area = 2.59) than the spaced (area = 2.44). Importantly, using this technique we can also examine individual learning patterns. We observe, for example, that individuals 54, 98 and 141 are some of the fastest learners in the massed learning condition, while individuals 116 and 174 are good learners in the spaced condition. Individuals 165 and 170 (massed) and individuals 129 and 83 (spaced) are relatively weak at learning this task. Those individuals in the fourth quadrant (23, 111 [massed] and 157, 6 [spaced]) rapidly learn the task but plateau quickly. These individuals have high slope but low intercept values. In contrast, those in the second quadrant have high intercept but low slope values, with some of these individuals starting well but getting fewer face-name pairs correct with additional trials (e.g. individual 152 in the spaced condition).

For the **in-person** face-name experiment, we also assessed the significance of the variance component associated with the slopes. There was evidence that its inclusion significantly improved our model goodness of fit (LR = 7.18, df = 1, $p = 0.0074$), which means that different individuals have significantly different learning speeds. Looking at the fixed effects, the interaction between trial and group condition was not significant (LR = 0.06, df = 1, $p = 0.8089$). The main effect of group condition was also not significant (LR = 2.14, df = 1, $p = 0.1438$), but the main effect of trial was (LR = 164.13, df = 1, $p < 0.0001$) (Fig. 2c). The slope for both groups was estimated to be approximately 0.70 (see Table 2), and is also interpreted in the log-odds scale. This means that adding a new trial increases the log-odds of correctly matching a face and a name by 0.70, which translates to being $e^{0.70} \approx 2$ times more likely to correctly match when compared to the previous trial. Although examination of the in-person experiment showed no significant difference between the conditions, we see more spread in the massed (area = 1.137) than in the spaced condition (area = 0.998, Fig. 2d). Similar exploration of individual learning patterns shows that there are more individuals in the > 85% area of the top right quadrant for the spaced compared to the massed condition (e.g. 26, 29, 33).

Table 2 Coefficient estimates and standard errors for the binomial GLMM fitted to the face-name pairs task data, online and in person. SE = standard error

Parameter	Online estimate (SE)	In-person estimate (SE)
Intercept (massed)	−1.22 (0.10)	−1.49 (0.17)
Slope (massed)	1.11 (0.06)	0.70 (0.06)
Intercept (spaced)	−0.81 (0.10)	−1.72 (0.17)
Slope (spaced)	0.86 (0.05)	0.68 (0.06)
$\sigma^2_{\text{Intercept}}$	0.46	0.56
σ^2_{Slope}	0.22	0.05

Virtual navigation task

Examination of the virtual navigation task (NavWell) demonstrated that the estimate of the variance component associated with the slope was very small (0.000006), and the likelihood ratio test statistic was very close to zero (LR < 0.0001, df = 1, $p = 1$). Therefore, the slopes seem to be very similar across all individuals. However, the effect of trial was highly significant (Fig. 3a), both in the linear predictor for the mean (LR = 142.43, df = 1, $p < 0.0001$) and in that for the variance (LR = 43.83, df = 1, $p < 0.0001$). This means that the learning increases from one trial to the other, with the time taken to reach the target decreasing by $1 - e^{-0.084} = 8\%$, on average, for the subsequent trial. Also, the variance in the data decreases from trial to trial, on average, by $1 - e^{-0.077} = 7.4\%$, until almost all participants have learned how to reach the target. Since the random slopes are based on a normal distribution with very low variance, they are all very close to zero, i.e. there is little to no individual deviation from the overall mean slope (see Table 3). Nevertheless, we can still rank individuals based on their predicted random intercepts. In this experiment, time decreases with learning; as a result, individuals in the second and third quadrants (e.g. 29 and 17) show good learning across the trials, whereas

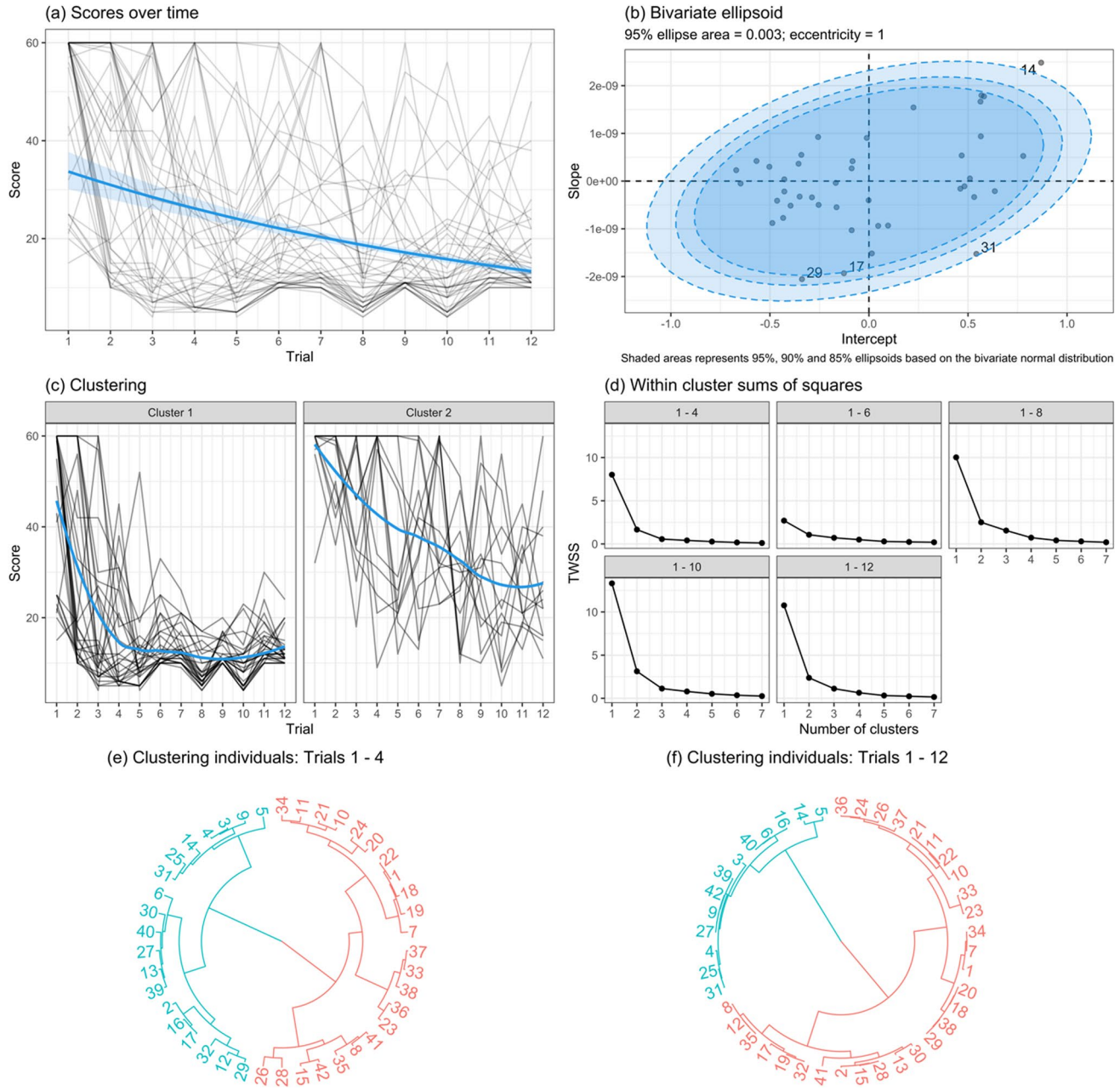

Fig. 3 **a** Observed (black lines) and predicted (blue curves) scores for individuals who took the NavWell virtual navigation task. The shaded area represents the 95% confidence intervals for the true mean scores, based on the right-censored gamma GLMM (see Table 1). **b** Predicted random intercepts and slopes for each individual in the study, as well as 85%, 90% and 95% ellipsoids based on the bivariate normal distribution. **c** The two-cluster solution obtained when carrying out hierarchical clustering analysis based on the Euclidean distance and Ward's method, using the predicted random intercepts and slopes per individual obtained from the right-censored gamma GLMM fitted to the full dataset. Blue curves are estimated from LOESS regression and are meant as a visual aid only. **d** Total within-cluster sums of squares (TWSS) for one- to seven-cluster solutions obtained when carrying out the cluster analysis to subsets of the data (trials 1–4, 1–6, 1–8, 1–10 and 1–12). Dendrogram and colour scheme representing the two-cluster solution obtained after fitting the model using trials (**e**) 1–4 and (**f**) 1–12, i.e. the full dataset

individuals in the first and fourth quadrants (e.g. 14) are slower learners (Fig. 3b).

Even though the random slopes were very close to zero, we again carried out a hierarchical clustering analysis based on both the individual-level intercepts and slopes (however, the predicted slopes will play almost no role in the clustering unless we scale both random intercepts and slopes to have unit variance). Over the full 12 trials, two clusters

Table 3 Coefficient estimates and standard errors for the censored gamma GLMM fitted to the virtual navigation task data. SE = standard error

Parameter	Estimate (SE)
Intercept (mean)	3.60 (0.06)
Slope (mean)	−0.08 (0.01)
Intercept (dispersion)	−0.16 (0.07)
Slope (dispersion)	−0.08 (0.01)
$\sigma^2_{\text{Intercept}}$	0.46
σ^2_{Slope}	< 0.01

emerged—those individuals who could be classified as fast learners and those who showed a slower learning pattern (Fig. 3c and f). The majority of individuals (29/42, red numbers in Fig. 3f) showed a pattern that learned rapidly over the first four trials before plateauing across the remaining ones (cluster 1, Fig. 3c), whereas the others (blue individuals in Fig. 3f) show a more gradual learning pattern, slowly decreasing across trials (cluster 2, Fig. 3c). The slow versus fast learning in this setting is highly dependent on the initial performance of the individuals. The distinction between early and late trials can be confirmed by looking at the total within-cluster sums of squares using data from fewer trials (Fig. 3d). This showed that running more trials was important for discriminating between the two groups. If the experiment had been stopped at e.g. trial 6, it would be difficult to separate the individuals into two well-defined clusters, as we are able to do when looking at the full dataset.

An advantage of our analysis is that we can track individual learning patterns across trials. For example, in trials 1–4 there is a relatively even split in the numbers of individuals showing rapid learning (red, $n = 23$) and gradual learning (blue, $n = 19$) (Fig. 3e). This proportion changes with the addition of extra trials (Fig. 3f). Some individuals change their pattern across trials (e.g. individuals 32, 12 and 29, clustered as slow learners [blue] for trials 1–4 but as fast learners [red] across trials 1–12). Others continue with the same pattern throughout (e.g. individuals 5, 14 and 39 remain gradual [blue] learners).

Sea Hero Quest

Here we observe non-linear behaviour of the task scores across trials, as the trials are semi-independent of each other. This is why we opted to model this behaviour using b-splines. We do have few trials (only five), and could include the trial predictor as a categorical factor in our model, but we opted to use random b-splines per individual to showcase the use of these random effects. Interpretation of the b-spline estimated coefficients is now not as straightforward as in the previous examples. In fact, they introduce great flexibility at the cost of direct interpretability of the estimated coefficients. Nevertheless, we can still look at the fitted curves and make inferences the same way as before. Since this is a slightly more complex model, we experienced convergence issues when fitting it to the data when including a random b-spline with three degrees of freedom per individual. The main convergence issues occurred when computing standard errors for the dispersion parameter estimates (see Table 4). This is likely due to having five points in the x-axis and fitting a model with four degrees of freedom, which essentially reproduces the marginal means at each point. However, the estimates obtained were reasonable and reproduced the behaviour in the data well at marginal and individual levels. We performed the likelihood ratio tests based on models fitted with random intercepts and slopes, which converged, allowing for the computation of log-likelihoods. The interaction between the non-linear effects of trial and age were highly significant for the linear predictors for both the mean ($\text{LR} = 112.08$, $\text{df} = 9$, $p < 0.0001$) and dispersion ($\text{LR} = 123.6$, $\text{df} = 9$, $p < 0.0001$), suggesting that different age groups are associated with different non-linear learning behaviour (see Fig. 4a).

The output of the cluster analysis with four clusters clearly shows that age group dominates the clustering.

Table 4 Coefficient estimates and standard errors for the gamma GLMM fitted to the Sea Hero Quest data. SE = standard error; S1 = spline knot 1; S2 = spline knot 2; S3 = spline knot 3; N/A = not available due to convergence issues

Parameter	Mean estimate (SE)	Dispersion estimate (SE)
Intercept (age 18–20)	3.52 (0.02)	4.78 (0.27)
S1 (age 18– 20)	1.86 (0.04)	8.57 (1.10)
S2 (age 18– 20)	0.36 (0.08)	−22.49 (0.74)
S3 (age 18– 20)	1.35 (0.03)	20.21 (N/A)
Intercept (age 21– 40)	3.48 (0.02)	16.07 (N/A)
S1 (age 21– 40)	1.83 (0.03)	−14.06 (N/A)
S2 (age 21– 40)	0.62 (0.07)	−26.58 (0.70)
S3 (age 21– 40)	1.35 (0.03)	6.41 (0.81)
Intercept (age 41– 60)	3.60 (0.03)	3.25 (0.25)
S1 (age 41– 60)	1.73 (0.10)	2.22 (1.11)
S2 (age 41– 60)	0.20 (0.16)	−3.63 (0.97)
S3 (age 41– 60)	1.54 (0.05)	−0.26 (0.52)
Intercept (age 61– 80)	3.59 (0.02)	3.19 (0.29)
S1 (age 61– 80)	1.59 (0.14)	−0.04 (1.04)
S2 (age 61– 80)	1.07 (0.20)	−4.12 (0.79)
S3 (age 61– 80)	1.75 (0.06)	−1.10 (0.42)
$\sigma^2_{\text{Intercept}}$	0.02	-
σ^2_{S1}	0.04	-
σ^2_{S2}	0.31	-
σ^2_{S3}	0.06	-

This is noteworthy because age group plays no role within the clustering algorithm itself; only the individual-level random effects do, and still we see a clear separation by age group, with individuals 18–40 in cluster 1, 21–60 in cluster 2, 41–80 in cluster 3, and 61–80 in cluster 4 (see Fig. 4b). An interesting feature of the cluster analysis (Fig. 4c) is that there are many individuals who have a learning pattern not typical for their age group. For example, there are some individuals in the 21–40 age category (green) who show learning patterns more similar to the 18–20 age group (red). In addition, the 41–60 age category (blue) is split fairly evenly, with a group of individuals showing learning patterns more similar to a younger age group (red and green) and a second group showing patterns more similar to the 61–80 age group (purple). Using this level of analysis may allow for early identification of individuals with spatial learning issues. This split in the 41–60-year-old cohort suggests that the data are best divided into three rather than four clusters (Fig. 4d).

Finally, we analysed each of the four experiments using linear models (see supplementary Table) to assess whether GLMMs would reveal anything extra at the overall (non-individual) level. In general, the results were comparable across the two methods.

Simulation studies

The results from the simulation studies suggest that (1) the GLMMs display better performance than the standard GLMs when estimating individual differences, based on both random intercepts and slopes (Fig. 5a and b); (2) the performance improves slightly when the numbers of participants and trials increase (Fig. 5a and b); (3) the fixed effects are estimated well and reliably for the GLMM, with slight performance improvement with an increase in the number of participants, and better improvement observed with an increase in the number of trials (Fig. 5c); and (4) the estimates of the variance components in the GLMM

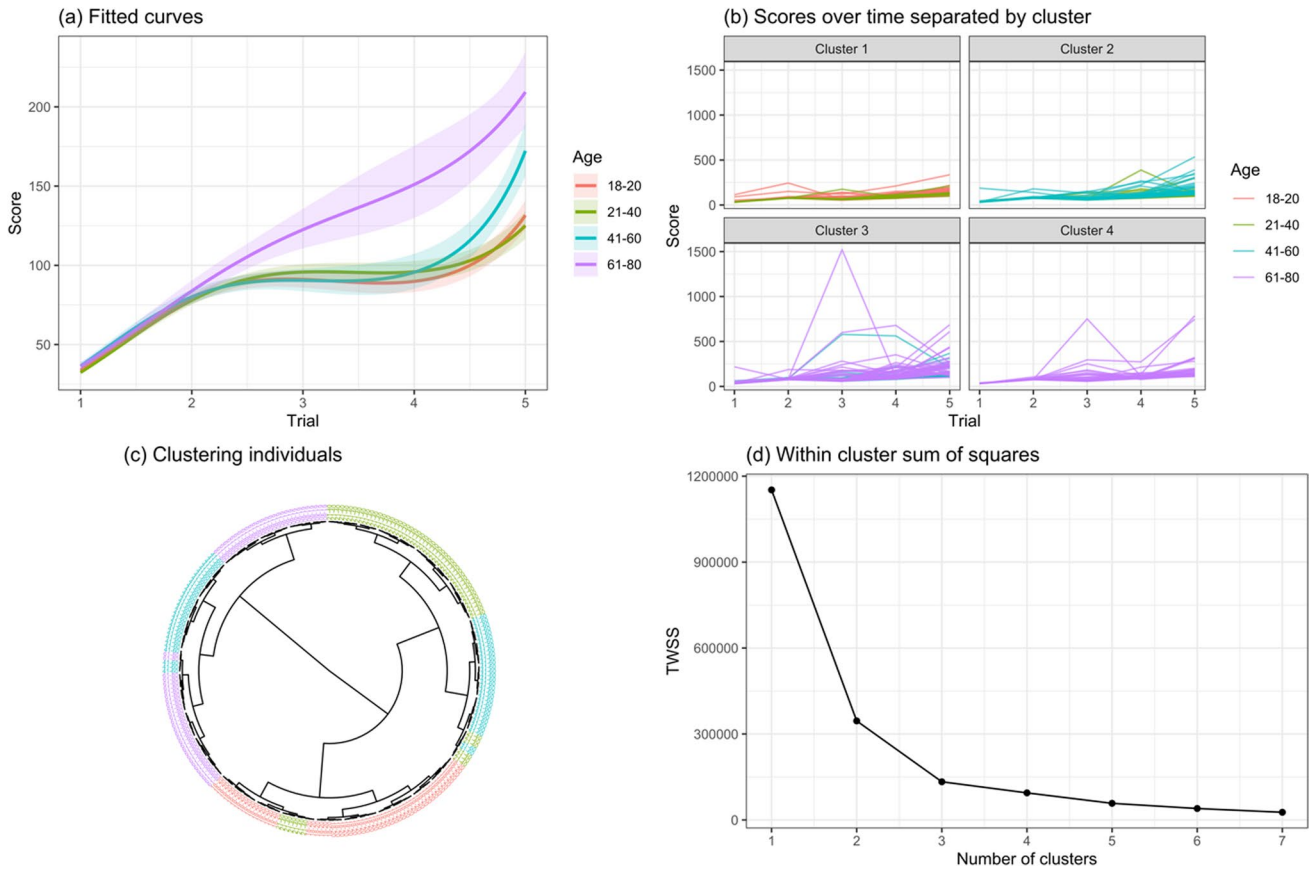

Fig. 4 **a** Predicted scores (solid curves) for individuals of different age groups from the Sea Hero Quest experiment, and associated 95% confidence intervals for the true mean scores (shaded areas) based on the gamma GLMM (see Table 1). **b** Observed scores for each individual split by cluster, obtained from the four-cluster solution of the hierarchical cluster analysis carried out using the random effects from the model, using the Euclidean distance matrix and Ward's method. **c** Dendrogram representing the full hierarchical clustering analysis, with colour scheme reflecting age group for each individual. **d** Total within-cluster sums of squares (TWSS) for one- to seven-cluster solutions obtained when carrying out the cluster analysis of the full dataset

are reliable and improve when sample size and number of trials increase, with poorer performance observed when fitting the gamma GLMM assuming random intercepts and slopes when the true model only has random intercepts (albeit still improved with larger sample sizes and number of trials; Fig. 5d).

Discussion

Here we show that the use of GLMMs can be broadened, so that individual variation in a dataset is not just controlled for but used to provide a deeper analysis on individuals' performance. While we have applied our approach to discrete, continuous, censored and non-censored data to demonstrate the possibility of using it across multiple datasets, the analysis may be suitable for datasets beyond these. Furthermore, while we have focused on learning among individuals and whether patterns of learning could be identified, a similar type of analysis may equally be applied to other psychological constructs and across other fields (e.g. examination of individual growth patterns in plants, effectiveness of a drug or therapeutic programme in individual patients). In terms of our face-name associative learning task, we able to report significant effects for condition (massed vs spaced) for the online experiment but not for the in-person experiment. Moreover, we were able to examine individual learning patterns, and show where an individual may lie along the slope/intercept and whether they exhibit outlying behaviour. Further, individuals' performance can be compared across conditions as well as within conditions. For example, with the in-person experiment, individuals in the spaced condition showed more extreme patterns, with many displaying very good learning (high intercept and slope values) compared to those in the massed condition.

Although cluster analysis did not reveal any significant patterns in the face-name task, two distinct patterns of learning emerged with the virtual navigation NavWell task. One pattern showed an initial rapid period of learning, plateauing at strong performance that was sustained for the rest of the trials. The second cluster showed a slower and more gradual learning pattern that saw individuals learn across all

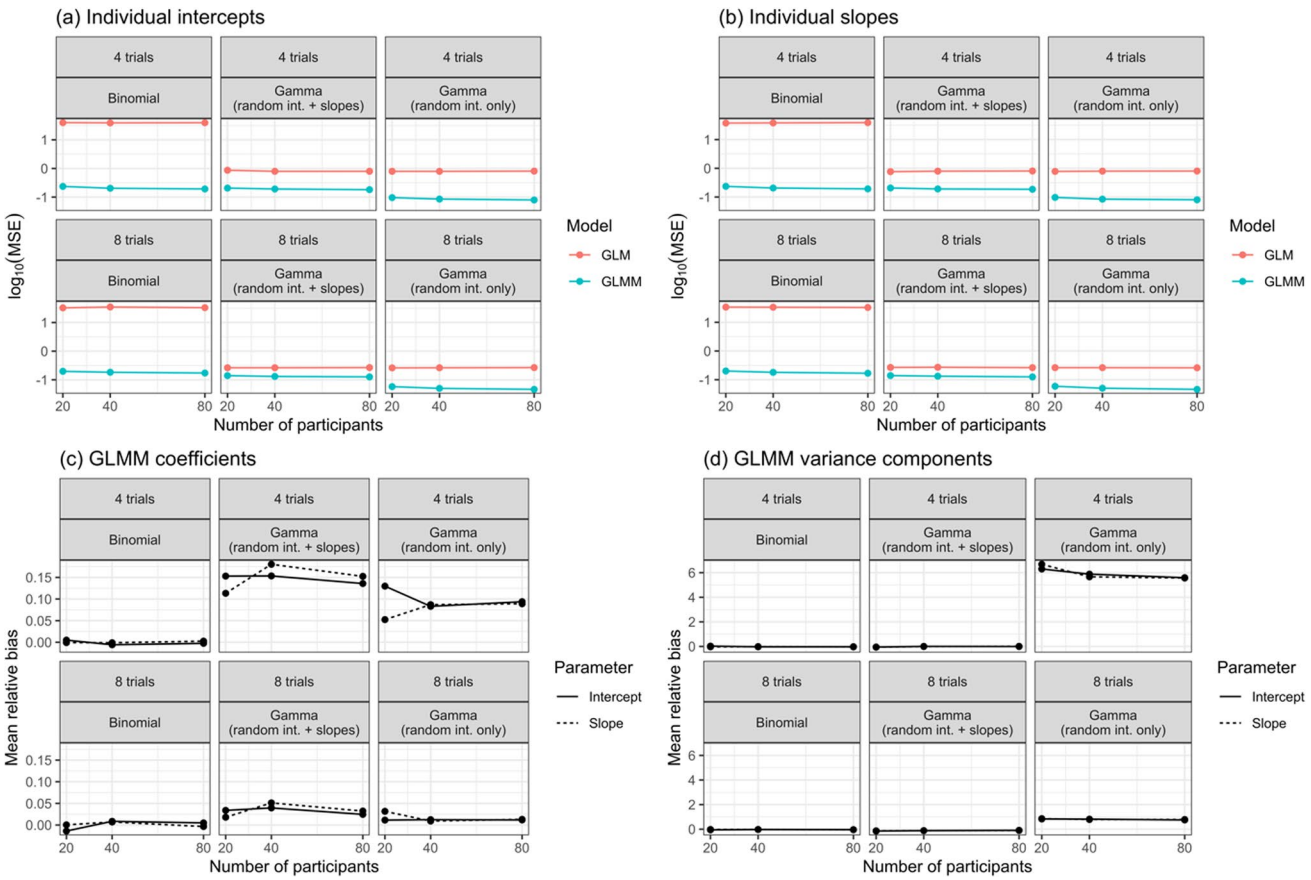

Fig. 5 Logarithm (base 10) of the mean squared error (MSE) of the individual-level random intercepts (**a**) and slopes (**b**) predicted by the generalised linear mixed models (GLMMs) and the standard generalised linear models (GLMs, which do not include random effects) across 1000 simulated datasets for 18 simulation scenarios. Mean relative bias for the fixed effects (**c**) and variance components (**d**) calculated across 1000 simulated datasets for 18 simulation scenarios for the GLMMs

the trials. This finding may support recent research showing that humans vary substantially in their spatial learning (Newcombe, 2018). Using a virtual college campus, Weisberg and Newcombe (2016) were able to classify individuals as integrators, non-integrators and imprecise navigators depending in whether individuals were able to form a cognitive map from different environments. Our results would support the idea of individual variation among individuals, showing patterns of both fast and slower learners. Although the water maze task is relatively simple, it is one of the most popular tools used to study learning and memory in both human and non-human animal research (Thornberry et al., 2021; Vorhees & Williams, 2006), as well as being an important assay for different diseases and neurological conditions (Goodrich-Hunsaker et al., 2010). Despite its popularity, group results rather than individual patterns of learning are typically reported (e.g. Woolley et al., 2015, and our own work, Farina et al., 2015). As a result, important information underlying the mechanism of spatial learning may be missed. The two patterns of water maze learning may be supported by different neural patterns. For example, it is known that the hippocampus supports fast learning, while other structures including the striatum and neocortex support slower learning (Kumaran et al., 2016). However, learning is dynamic, and this is reflected in individuals changing from one pattern to another across trials; again, this may be indicative of dynamic neural interactions between different learning systems (see e.g. Kosaki et al., 2015). Further investigation is warranted to examine the underlying biological and/or environmental factors that support the different patterns.

One of the interesting findings to emerge from the SHQ data is that learning patterns show good correspondence to the age group. As might be expected, the learning patterns are similar between the 18–20-year-olds and 21–40-year-olds, who show good learning across the five levels of difficulty, especially when compared to the older groups. This is not unexpected, as spatial cognition declines with age (Techentin et al., 2014), and the Irish sample fits readily to what has been observed internationally with the larger SHQ data (see Coutrot et al., 2018). However, here we can examine individual performance in more detail and check to see how such individuals compare to their age-matched cohort. We show that some individuals demonstrate patterns that more closely match a younger or older cohort, rather than their own. For example, the 41–60 age group is particularly interesting, as some individuals show patterns that are more similar to a younger cohort (21–40), while many others show patterns more similar to the older group (61–80). Indeed, our cluster analysis shows that three clusters may better account for the data than four. This age group may therefore represent an important transition period in terms of spatial learning. Further research is needed to examine whether this transition is similar for other cognitive processes. In addition, as spatial learning deficits may be an early indicator of dementia and Alzheimer's disease (Coughlan et al., 2018), the ability to identify individuals in a younger age cohort who perform as if much older may help to provide an early warning and may be useful as a cognitive digital marker for future disease. However, caution is warranted, as the age groups chosen for our analysis were wide, spanning 20 years (e.g. 41–60) for the most part (except for the 18–20 cohort); therefore, it is possible that those in their early 40s show patterns similar to the 21–40 age group. Similarly, those in their late 50s may show patterns similar to the 61–80 age group; further analyses are needed to examine this. It is also important to note that the SHQ data present only four distinct trials, and therefore using splines to model the non-linear behaviour may lead to overfitting. However, the objective here was to describe these trends as well as possible through individual random effects, so that the subsequent cluster analyses could reveal learning patterns with respect to the participants' ages. If, in contrast, the objective had been to predict responses for different trials, then a less flexible linear predictor could have been more adequate.

It is important to mention that model choice plays a significant inferential role. If the GLMM is mis-specified, this can lead to erroneous conclusions. Model misspecification includes, but is not limited to, choice of distribution that is incompatible with the response variable (e.g. use of a normal distribution to model discrete data), choice of inappropriate link function (e.g. use of a symmetrical link for a binomial model, such as the logit or probit, when the data exhibit asymmetrical behaviour), failure to account for overdispersion (when the variability in the data is larger than accounted for by the model; there are many model extensions capable of dealing with this phenomenon), omission of important predictors (e.g. not accounting for experimental design or failure to include important interactions) and failure to model dispersion appropriately (e.g. when dispersion changes according to treatment). Similarly, the choice of distance and clustering method can significantly alter the results of the clustering analysis. Therefore, it is important to use appropriate distances and compare the results from different clustering methods. Future simulation studies will be helpful for understanding how different types of model misspecification (linear predictor specification, choice of link function, etc.) and different clustering techniques affect the outcomes when analysing individual differences.

One particular strength of GLMMs is that they can address missing data, especially for longitudinal studies. For participants who did not complete all tasks in a study

(after dropout), GLMMs can still provide predictions, which will be based on the overall mean intercepts and slopes when there is not enough information to predict individual random effects. This allows one to make better use of the information available in the data, rather than, as an alternative, omit data from participants when the available data points are insufficient to estimate a proposed model. Similarly, for time-until-event responses, when a participant did not complete the task in the allotted time (e.g. NavWell data), the data are right-censored, which means that if the participant had been given enough time, they would have eventually completed the task. Therefore, we know that the time taken to complete is greater than the allotted time; however, we do not know exactly what that time is. One alternative is to omit these observations and carry out a conditional analysis. However, this analysis would not make use of all the available information in the data. Therefore, it is best to incorporate into the likelihood the fact that we know the time is greater than a certain threshold, which yields inferential results unconditional to the fact that the participant did not complete the task.

In conclusion, we have demonstrated the flexibility of using GLMMs when using repeated measures. Examination of individual differences is important to identify outliers or simply those who do or do not perform well. Such an approach may be useful for educators or in a clinical setting to help identify individuals who might need further assistance or attention. Furthermore, identifying clusters of individuals and learning patterns may allow further investigation of the underlying biological, environmental and other factors that may help explain why such patterns occur.

All data and R scripts are made available through http://www.github.com/rafamoral/individual_learning_GLMM.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-023-02232-z.

Authors' contributions RdAM: Involved in concept, analysed data, wrote and edited manuscript. SC: Involved in concept, analysed data, wrote and edited manuscript. HS, AC, MH: provided data, wrote and edited manuscript.

Data availability Not applicable

Code availability All code used is available through GitHub.

Declarations

Conflicts of interest/Competing interests There are no competing interests associated with this paper.

Ethics approval Experiments received approval from Maynooth University, and the procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Consent for publication All authors have approved submission.

References

Allen, M., Poggiali, D., Whitaker, K., et al. (2019). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research, 4*, 63.

Baayen, R. H., Tweedie, F. J., & Schreuder, R. (2002). The subjects as a simple random effect fallacy: subject variability and morphological family effects in the mental lexicon. *Brain Lang, 81*(1–3), 55–65. https://doi.org/10.1006/brln.2001.2506

Barnhart, C. D., Yang, D., & Lein, P. J. (2015). Using the Morris water maze to assess spatial learning and memory in weanling mice. *PLoS One, 10*(4), e0124521.

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. M., & White, J.-S.S. (2008). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution, 24*(3), 127–135.

Bootsma, J. M., Hortobágyi, T., Rothwell, J. C., & Caljouw, S. R. (2018). The Role of task difficulty in learning a visuomotor skill. *Medicine and Science in Sports and Exercise, 50*(9), 1842–1849.

Braithwaite, D. W., Leib, E. R., Siegler, R. S., & McMullen, J. (2019). Individual differences in fraction arithmetic learning. *Cognitive Psychology, 112*, 81–98.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal, 9*(2), 378–400.

Caffrey, M., & Commins, S. (2022). *Examination of Distributed Learning on Recent and Remote Memory Using Inperson and Online Experimental Paradigms.* https://doi.org/10.31234/osf.io/xk9rw

Chiang, S., Haut, S. R., Ferastraoaru, V., Rao, V. R., Baud, M. O., Theodore, W. H., Moss, R., & Goldenholz, D. M. (2020). Individualizing the definition of seizure clusters based on temporal clustering analysis. *Epilepsy Research, 163*, 106330.

Chou, M. Y., Nishita, Y., Nakagawa, T., Tange, C., Tomida, M., Shimokata, H., Otsuka, R., Chen, L. K., & Arai, H. (2019). Role of gait speed and grip strength in predicting 10-year cognitive decline among community-dwelling older people. *BMC Geriatrics, 19*(1), 186.

Cochrane, C., Ba, D., Klerman, E. B., & St Hilaire, M. A. (2021). An ensemble mixed effects model of sleep loss and performance. *Journal of Theoretical Biology, 509*, 110497.

Commins, S., Duffin, J., Chaves, K., Leahy, D., Corcoran, K., Caffrey, M., Keenan, L., Finan, D., & Thornberry, C. (2020). NavWell: A simplified virtual-reality platform for spatial navigation and memory experiments. *Behavior Research Methods, 52*(3), 1189–1207.

Coughlan, G., Laczó, J., Hort, J., Minihane, A. M., & Hornberger, M. (2018). Spatial navigation deficits - overlooked cognitive marker for preclinical Alzheimer disease? *Nature Reviews Neurology, 14*(8), 496–506.

Coutrot, A., Silva, R., Manley, E., de Cothi, W., Sami, S., Bohbot, V. D., Wiener, J. M., Hölscher, C., Dalton, R. C., Hornberger, M., & Spiers, H. J. (2018). Global Determinants of navigation ability. *Current Biology, 28*(17), 2861-2866.e4.

Demétrio, C. G. B., Hinde, J., & Moral, R. A. (2014). Models for overdispersed data in entomology. In C. P. Ferreira & W. A. C. Godoy (Eds.), *Ecological modelling applied to entomology.* Springer.

Farina, F. R., Burke, T., Coyle, D., Jeter, K., McGee, M., O'Connell, J., Taheny, D., & Commins, S. (2015). Learning efficiency: The influence of cue salience during spatial navigation. *Behavioural Processes, 116*, 17–27.

Firth, J., Firth, J. A., Stubbs, B., Vancampfort, D., Schuch, F. B., Hallgren, M., Veronese, N., Yung, A. R., & Sarris, J. (2018). Association Between Muscular Strength and Cognition in People With Major Depression or Bipolar Disorder and Healthy Controls. *JAMA Psychiatry., 75*(7), 740–746.

Goodrich-Hunsaker, N. J., Livingstone, S. A., Skelton, R. W., & Hopkins, R. O. (2010). Spatial deficits in a virtual water maze in amnesic participants with hippocampal damage. *Hippocampus, 20*(4), 481–91.

Kliegl, R., Wie, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology, 1*, 238.

Kosaki, Y., Poulter, S. L., Austen, J. M., & McGregor, A. (2015). Dorsolateral striatal lesions impair navigation based on landmark-goal vectors but facilitate spatial learning based on a "cognitive map". *Learning and Memory, 22*, 179–191.

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences., 20*(7), 512–534.

Mattei, J., Bigornia, S. J., Sotos-Prieto, M., & Scot, t T., Gao, X., Tucker, K.L. (2019). The Mediterranean Diet and 2-Year Change in Cognitive Function by Status of Type 2 Diabetes and Glycemic Control. *Diabetes Care, 42*(8), 1372–1379.

McCulloch, C. E., & Neuhaus, J. M. (2011). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics, 67*(1), 270–279.

McDade, E., Wang, G., Gordon, B. A., Hassenstab, J., Benzinger, T. L. S., et al. (2018). Longitudinal cognitive and biomarker changes in dominantly inherited Alzheimer disease. *Neurology., 91*(14), e1295–e1306.

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Moral, R. A., Hinde, J., & Demetrio, C. G. B. (2017). 'Half-Normal Plots and Overdispersed Models in R: The hnp Package. *Journal of Statistical Software, 81*, 1–23.

Morris, R. G. (1981). Spatial localization does not require the presence of local cues. *Learning and Motivation, 12*, 239–260.

Newcombe, N. S. (2018). Individual variation in human navigation. *Current Biology, 28*, R952–R1008.

Paulus, M. P., Squeglia, L. M., Bagot, K., Jacobus, J., Kuplicki, R., Breslin, F. J., Bodurka, J., Morris, A. S., Thompson, W. K., Bartsch, H., & Tapert, S. F. (2019). Screen media activity and brain structure in youth: Evidence for diverse structural correlation networks from the ABCD study. *Neuroimage, 185*, 140–153.

Quesque, F., Coutrot, A., Cox, S., de Souza, L. C., Baez, S., Cardona, J. F., Mulet-Perreault, H., Flanagan, E., Neely-Prado, A., Clarens, M. F., Cassimiro, L., Musa, G., Kemp, J., Botzung, A., Philippi, N., Cosseddu, M., Trujillo-Llano, C., Grisales- Cardenas, J. S., Fittipaldi, S., … Bertoux, M. (2022). Does culture shape our understanding of others' thoughts and emotions? An investigation across 12 countries. *Neuropsychology., 36*(7), 664–682. https://doi.org/10.1037/neu0000817

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. URL https://www.R-project.org/

Raboyeau, G., Marcotte, K., Adrover-Roig, D., & Ansaldo, A. I. (2010). Brain activation and lexical learning: the impact of learning phase and word type. *Neuroimage, 49*(3), 2850–61.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics, 54*(3), 507–554.

Seidler, R. D., et al. (2015). Individual predictors of sensorimotor adaptability. *Frontiers in Systems Neuroscience, 9*, 100.

Simon, G. (2001). Choosing a first-line antidepressant: Equal on average does not mean equal for everyone. *JAMA, 286*(23), 3003–4.

Song, R., Xu, H., Dintica, C. S., Pan, K. Y., Qi, X., Buchman, A. S., Bennett, D. A., & Xu, W. (2020). Associations between cardiovascular risk, structural brain changes, and cognitive decline. *Journal of the American College of Cardiology, 75*(20), 2525–2534.

Spiers, H. J., Coutrot, A., & Hornberger, M. (2023). Explaining World-Wide Variation in Navigation Ability from Millions of People: Citizen Science Project Sea Hero Quest. *Top Cogn Sci., 15*(1), 120–138. https://doi.org/10.1111/tops.12590

Stasinopoulos, M. D., Rigby, R. A., & Bastiani, F. D. (2018). GAMLSS: A distributional regression approach. *Statistical Modelling., 18*(3–4), 248–273. https://doi.org/10.1177/1471082X18759144

Techentin, C., Voyer, D., & Voyer, S. D. (2014). Spatial Abilities and Aging: A Meta-Analysis. *Experimental Aging Research, 40*, 395–425.

Thornberry, C., Cimadevilla, J. M., & Commins, S. (2021). Virtual Morris water maze: opportunities and challenges. *Rev Neurosci, 32*(8), 887–903. https://doi.org/10.1515/revneuro-2020-0149

Vorhees, C. V., & Williams, M. T. (2006). Morris water maze: Procedures for assessing spatial and related forms of learning and memory. *Nature Protocols, 1*(2), 848–858.

Weisberg, S. M., & Newcombe, N. S. (2016). How do (some) people make a cognitive map? Routes, places, and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 768–785.

Woolley, D. G., Mantini, D., Coxon, J. P., D'Hooge, R., Swinnen, S. P., & Wenderoth, N. (2015). Virtual water maze learning in human increases functional connectivity between posterior hippocampus and dorsal caudate. *Human Brain Mapping, 36*(4), 1265–77.

Zeineh, M. M., Engel, S. A., Thompson, P. M., & Bookheimer, S. Y. (2003). Dynamics of the hippocampus during encoding and retrieval of face-name pairs. *Science, 299*(5606), 577–580.